*Revised Manuscript
Click here to view linked References

Development of Quantitative Structure Activity Relationship (QSAR) Model for Disinfection Byproduct (DBP) Research: A Review of Methods and Resources

Baiyang Chen[1*], Tian Zhang[1], Tom Bond[2], Yiqun Gan[1]

[1]Harbin Institute of Technology Shenzhen Graduate School, Shenzhen Key Laboratory of Water Resource Utilization and Environmental Pollution Control, Shenzhen 518055, China.

[2]Department of Civil and Environmental Engineering, Imperial College, London, United Kingdom SW7 2AZ.

[*]Corresponding author telephone: 86-134-8072-7605; Fax: 86-755-2603-2557; Email address: poplar_chen@hotmail.com.

Abstract

Quantitative structure-activity relationship (QSAR) models are tools for linking chemical activities with molecular structures and compositions. Due to the concern about the proliferating number of disinfection byproducts (DBPs) in water and the associated financial and technical burden, researchers have recently begun to develop QSAR models to investigate the toxicity, formation, property, and removal of DBPs. However, there are no standard procedures or best practices regarding how to develop QSAR models, which potentially limit their wide acceptance. In order to facilitate more frequent use of QSAR models in future DBP research, this article reviews the processes required for QSAR model development, summarizes recent trends in QSAR-DBP studies, and shares some important resources for QSAR development (e.g., free databases and QSAR programs). The paper follows the four steps of QSAR model development, i.e., data collection, descriptor filtration, algorithm selection, and model validation; and finishes by highlighting several research needs. Because QSAR models may have an important role in progressing our understanding of DBP issues, it is hoped that this paper will encourage their future use for this application.

Key Words: disinfection byproduct; QSAR/QSPR/QSTR/LFER; descriptor filtration; algorithm selection; model validation.

Abbreviation list

AD	Applicability domain
ANN	Artificial neural network
CA	Correlation analysis
DBP	Disinfection byproduct
DBPFP	DBP formation potential
DOC	Dissolved organic carbon
EU	European Union
E_{HOMO}	Energy of highest occupied molecular orbital
E_{LUMO}	Energy of lowest unoccupied molecular orbital
FS	Forward selection
GA	Genetic algorithm
HA	Halogenated acetaldehyde
HAA	Haloacetic acid
HAcAm	Haloacetamide
HAN	Haloacetonitrile
HK	Haloketone
HM	Halomethane
HNM	Halonitromethane
kNN	k-nearest neighbors
LFER	The linear free energy relationship
LogP, K_{OW}	Octanol-water partition coefficient, in decadic logarithm form
LMO	Leave many out, equivalent to leave N out (LNO)
LOO	Leave one out
MLP	Machine learning process
MLR	Multiple linear regression
MW	Molecular weight
MX	3-chloro-4-(dichloromethyl)-5-hydroxyl-2(5H)-furanone
NIST	National Institute of Standards and Technology
n-PLS	Nonlinear PLS
OECD	Organization for Economic Cooperation and Development
PCA	Principal component analysis
PCR	Principal component regression
pK_a	Acid dissociation constant, in negative decadic logarithm form
PLS	Partial least square
QSAR	Quantitative structure-activity relationship
QSPR	Quantitative structure-property relationship
QSTR	Quantitative structure-toxicity relationship

RM	Replacement method
SDR	Standardized residual
SLR	Single linear regression
SVM	Support vector machine
THM	Trihalomethane
TOX	Total organic halogen
USA	United States of America
WOK	Web of knowledge

Symbol list

r_H	hydrolysis rate of the reactant, d_C/d_t
k_H	hydrolysis rate constant
C	Concentration
k_a, k_n, k_b	the rate constants at acidic, neutral, and basic conditions, respectively
$k_{H,T}$	hydrolysis rate constant at specific temperature
E_A	Arrhenius activation energy
T	absolute temperature
R	gas constant (1.987 cal/degree-mole)
Y	activity of a compound
σ	electronic effect constant
σ*	aliphatic inductive effect
σ_m	meta-substituted constant
σ_p	para-substituted constant
Es	steric effect constant
π	hydrophobic effect constant and sometimes described as LogP
δ	steric effect coefficient
ρ	electronic effect coefficient
k, k'	coefficients of hydrophobic effect
X_n	descriptor value
n	number of observations
b	descriptor coefficient
m, m'	slopes obtained by predicted data versus experimental data with and without a y-intercept, respectively
R_i^2 , R_0^2	determination coefficients obtained by predicted data versus experimental data with and without a y-intercept, respectively
α, β, γ	Power factors in nonlinear regression equation

Introduction

1.1 Disinfection byproducts (DBPs)

In water and wastewater engineering systems, DBPs are formed by reactions between disinfectant(s) and organic precursors, both in the presence and absence of bromide and iodide [1, 2]. Chloroform and other trihalomethane (THM) species were the first DBPs identified in the 1970s [3]. Since then, epidemiological studies have consistently shown an association between long-term exposure to chlorinated drinking water and an enhanced risk of developing bladder cancer [4]. Other suspected public health impacts over shorter time scales, including disruption of *in utero* development and adverse birth outcomes, have also been reported [5, 6]. Consequently, four chloro- and/or bromo- THMs are currently regulated in the United States of America (USA) and European Union (EU), while five haloacetic acids (HAAs) are also regulated in the USA.

Although chlorine is the most common water disinfectant used globally, alternative disinfectants such as ozone, chlorine dioxide and chloramines are increasingly employed, and each has its own characteristic reactivity with organic compounds precursors in drinking water [7, 8]. Furthermore, organic precursors originate from a variety of sources, including decaying vegetation, soil, intracellular and extracellular algal organic matter, agricultural runoff (e.g. pesticides and herbicides), and municipal wastewater (including nitrogenous compounds, pharmaceuticals and endocrine disrupting chemicals) [9, 10]. With these diverse types of disinfectants and precursors, it is apparent that many types of DBPs can be formed. In fact, approximately 700 DBP species have been reported [2, 7], but the sum of them still accounts for less than half of the total organic halogen (TOX) according to the surveys undertaken in the USA [2, 8, 11]. Together with advances in analytical methods [12], the number of detectable DBPs is expected to proliferate

continuously. For example, there have been recent concerns about unregulated nitrogenous [2, 13] and iodinated DBPs [14-16] due to their relatively high genotoxicity and cytotoxicity [17-21], which has expanded the pool of DBPs. Unfortunately, the diversity and quantity of DBPs poses daunting challenges to researchers, regulators, and water suppliers to fully understand their analysis, occurrence, toxicity, property, formation, degradation, and removal, not to mention overwhelmingly financial burdens. Therefore, it is beneficial to find effective ways to streamline and prioritize future DBP research, which potentially includes an increasing deployment of quantitative structure-activity relationship (QSAR) models.

1.2 Quantitative structure-activity relationships (QSARs)

QSARs are a suite of tools used to link chemical activities with molecular structure and composition. QSAR techniques are well suited to deal with large quantities of data, which allows researchers to discern trends in existing data with environmental relevance and then extrapolate trends to untested compounds. For example, an early study developed a QSAR model for DBP hydrolysis from existing data and then applied the model to predict the half-life of triiodoacetic acid in water [22], which is shorter than 4 hrs. Consequently, in order to avoid decomposition, samples were extracted and methylated within 4 h before analysis [23].

The use of QSAR techniques is based upon two underlying principles: 1) structurally similar compounds behave comparably under similar environmental conditions, and 2) behavioral differences among compounds are linked to structural and compositional variations. The predictor variables are usually described as “descriptors” (or features, attributes, independent variables, structural/compositional components, etc.), and the resultant response variables (e.g., reactivity, toxicity or bioactivity) are termed as “activities” (or endpoints, dependent variables,

etc.). Alternative names for QSAR models are recognized in other research fields. In particular, the term quantitative structure property relationship (QSPR) is sometimes used to emphasize compound properties [24, 25], quantitative structure toxicity relationship (QSTR) is sometimes used in toxicity studies [26], while linear free-energy relationship (LFER) is accepted in some degradation studies [27]. Regardless of the name, these studies always correlate descriptor(s) with activities, and therefore follow common QSAR principles.

1.3 History and status of QSARs in environmental and DBP studies

The concept of using QSARs to link structure and activity was introduced over 100 years ago and subsequently widely used in medical and biological research [28, 29]. In 1975, Veith and Konasewich introduced QSAR models into environmental research [30]. Soon after, a book edited by Lyman [31] in 1982 evaluated the environmental behaviors of halogenated compounds using QSAR models, which included some DBPs even though these models were not designed specifically for DBPs. The first integration of DBPs and QSAR research occurred in 1991 when Tuppurainen and colleagues developed a QSAR model especially for DBP toxicity [32].

To better understand the status of QSAR models in DBP and, more broadly, environmental research, we undertook a literature search using two academic databases, Web of Knowledge (WOK) and Scopus. The keywords used for topic search were QSAR OR QSPR OR QSTR OR LFER, and the results were limited to the Environmental Sciences Ecology scope. This search returned over 1400 references; however, when the additional term “DBP” or “disinfection byproduct” or “disinfection by-product’ was included, fewer than 12 references were identified. Consequently, we conducted an additional search in Google Scholar with the keywords “QSAR OR QSPR OR QSTR OR LFER AND DBP”, followed by manual exclusion of unrelated

literature. Through this effort, ten studies on toxicity, five studies on formation, eight studies on transformation, and eight studies on properties were identified (Table 1), all showing reasonable goodness of fit ($R^2 > 0.7$) between modeled data and experimental data. Figure 1 shows temporal trends in the frequency of references identified in both environmental and DBP studies until the middle of 2014. The increasing number of studies indicates the growing acceptance of QSAR methods in environmental science, with over 100 references in recent years. By comparison, the number of QSAR-DBP studies is limited (i.e., <4/year), implying that DBP research is only a small portion of environmental research and QSAR techniques remain underutilized by DBP researchers. The reasons of such underutilization are confounding but may include: 1) the chemical structures of natural organic matter (NOM) as DBP precursors are not well understood, which makes it difficult to predict DBP formation; 2) only a small portion (<100) of the DBPs identified (~700) have been tested for toxicity so far, while most QSARs are based on training sets of chemicals that do not include any/many DBPs.

1.4 Criterion in selecting QSAR-DBP studies

The limited number of QSAR-DBP studies may be partly attributed to the strict screening criteria applied, which excluded 1) literature appearing twice in different formats; for example, once in a journal and again in a dissertation or thesis [24, 33]. In such cases, the journal article was selected because it is peer-reviewed and more accessible; 2) those using qualitative analysis [34] rather than QSAR analysis; 3) studies applying previously-developed QSAR models rather than developing novel models [35-37], because the focus of this article is to help understand the development process; and 4) articles applying non-structural and non-compositional bulk parameters, such as dissolved organic carbon (DOC), pH, and temperature, in models [38, 39] because these do not comply strictly with QSAR principles. Although there are various empirical

models correlating DBP formation with DOC and ultraviolet absorbance (UV_{254nm}) [40], they usually regard organic matter as a single, unified entity and therefore lacked mechanistic insights into the structural and compositional characteristics of DBP precursors. In addition, most empirical QSARs that have been developed for DBP formation only work within the single water supply system they were developed on, but do not extrapolate well to other waters.

1.5 Purpose of this article

Notwithstanding the strict search criteria used, it is evident that there is a paucity of research in deploying QSAR methods to investigate DBP behaviors. The purpose of this study is to fill the gap by reviewing and critiquing existing QSAR-DBP studies, identifying trends and limitations, and therefore facilitating better understanding of the QSAR model development process. This study does not restate or justify modeling results presented in literature, but rather summarizes typical approaches used for QSAR model development. This article also does not provide mathematical details of algorithms and descriptor calculation methods. However, to encourage acceptance and increased use of QSAR analysis, a list of resources that may facilitate QSAR-DBP studies is provided. Finally, we present several knowledge gaps that may merit additional QSAR-DBP study in the future.

2. Discussion

2.1 Data Collection

2.1.1 Introduction to activity data

Figure 2 presents a flow chart of QSAR model development. The first step is data collection, which involves collating activity and descriptor information while ensuring the data are as thorough and reliable as possible. Often, two databases are established in this process, one for

activity and another for descriptors. No matter which activities or descriptors are selected, there are three ways to obtain data: 1) experimentation 2) literature, i.e. academic journals, reports, websites, dissertations and theses; and 3) computer programs or databases, either online or off-line. A database contains experimental activity data and/or descriptor values collected from academic publications, whereas a computer program usually yields values from existing, embedded models.

The preferred data source varies among researchers. Sometimes, experimental results are regarded as the most authentic option, and the ideal data for QSAR model development are considered to be those derived from the same laboratory and by the same analyst(s) [41]. By contrast, literature data are often difficult to obtain and less reliable when they are collated from multiple resources. The main problem of using various data sources for generating a training set is that false negatives or false positives (essentially inaccurate data) may lead to a poor model and inconsistencies. Conversely, literature-based models may sometimes be considered more robust and useful than models derived from a single laboratory because they incorporate data from multiple sources and diverse types of compounds. Similarly, data from computer programs have their own limitations; if a program produces data by an embedded model, the data are unsuitable for the development of a new QSAR model because it risks circular reasoning. However, if the program is a collection of experimental and literature results, it is essentially a literature resource and its use should be encouraged.

2.1.2 Existing activity data collection approaches

Regardless of data source, obtaining reliable and standardized data is critical to develop robust models. Table 1 summarizes existing QSAR-DBP literature identified in this study. The majority

of studies (26 out of 38) collected data from literature and databases. Twelve cases utilized their own experimental results, which mainly fell in the area of DBP toxicity and transformation. The total number of information sources (38) exceeds the number of studies (31), because some literature used data from multiple sources. Only one study employed activity data from an existing model [42]. Although the study showed good fitting, the use of modeled data to develop new QSAR model is prone to risk of circular reasoning, and therefore should be avoided in future study.

Data pretreatment usually takes a lot of time and effort. When activity information is available from multiple sources, a common approach is to use average data rather than all data [22, 33] to avoid the bias of diverse sources and to offer improved goodness-of-fit between predicted and measured data. Another issue in data pretreatment is to unify various sources of data with different units or experimental conditions. For example, short-term DBP formation potentials (DBPFPs) in water were converted to long-term DBPFP by a correcting factor [43]; and the hydrolysis kinetics of DBPs were standardized by either a pH-dependent equation (Equation 1) or the Arrhenius equation (Equation 2) [22].

$$r_H = -k_H \times [C] = -[C] \times (k_a \times [H^+] + k_n + k_b \times [OH^-]) \quad \text{Equation 1}$$

$$\ln \frac{k_{H,T_1}}{k_{H,T_2}} = \frac{E_A}{R}\left(\frac{1}{T_2} - \frac{1}{T_1}\right) \quad \text{Equation 2}$$

Note: see symbol descriptions in symbol table

2.1.3 Resources for activity data

Other than common academic databases, such as WOK, Scopus, Pubmed, Sciencedirect, Springer, and Taylor & Francis, professional websites, software, and programs are also important resources for collecting activity data; and many of these are free and publically accessible (Table 2). In DBP research, toxicity data resources are the most abundant followed by chemical properties and DBP transformation data; DBP formation data are the most limited (Table 2). Because many dedicated toxicity programs, such as TOPKAT, Derek, and MultiCase [44-47] are readily available already, the need of developing new QSAR models to estimate DBP toxicity is perhaps not as urgent as other activities. These tools can play an important role to prioritize DBPs in future research, as limited experimental toxicity data is currently available for many species.

The TOXNET website is probably the largest resource for chemical toxicity in the world as it integrates several other toxicity databases. The EU-sponsored report by Serafimova et al. [48] is a useful summary of QSAR toxicity models, and additional information regarding toxicity prediction is available from literature [49]. Furthermore, computer programs like EPI Suite, SPARC, and ACD/Lab and online databases like Chemspider provide many physical and chemical properties that may assist in the development of QSAR models for DBP properties. Regarding DBP removal, the National Institute of Standards and Technology (NIST) chemical kinetics database includes kinetic data, for example, on the ozonation of DBPs, which can be used to develop QSAR models for DBP reactivity. In addition, DBP biodegradability has been previously examined in a series of QSAR models [50]

2.1.4 Introduction to descriptors

Descriptors represent the structure and functionality of selected chemicals. With the advent of computer technology, thousands of descriptors are now available [28]. The descriptors can be classified into various groups, including 1) compositional descriptors (e.g., molecular weight [MW], type of halogen); 2) topological descriptors (e.g., Wiener index, Randic index); 3) quantum descriptors (e.g., energy of highest occupied molecular orbital [E_{HOMO}], energy of lowest unoccupied molecular orbital [E_{LUMO}]); 4) electronic parameters (e.g., local charge, polarity index, Hammett and Taft σ constants); 5) geometric parameters (e.g., steric constant E_s and electric constant σ), molecular volume, solvent contact surface area). Descriptors can also be classified as integral, such as E_{HUMO} and MW, which represent the whole compound, or substitutional, such as E_s and σ, which refer to specific substituent or functional group. Integral descriptor value can only be used for a specific compound, while substitutional descriptor values are applicable to multiple compounds; e.g., trichloromethane and trichloroacetic acid both have the same σ and E_s values for chlorine.

2.1.5 Existing descriptor data trends

To date, most QSAR-DBP studies have employed data from literature rather than from calculations (Table 1). The most commonly used descriptors are the compositional descriptors σ and Es, which were used 22 times. Their widespread acceptance can be attributed to their ease of interpretation and robustness in modeling. A few studies used certain chemical properties such as octanol-water partition coefficient (LogP) [51-53] and acid dissociation constant (pK_a) [52, 54, 55] as descriptors, indicating that an experimental property or activity was also used as a descriptor. This approach may be termed a property-property relationship rather than a structure-property relationship. Quantum descriptors like E_{LUMO} and E_{HOMO} were the second most frequently used descriptors, which is likely due to the gradual recognition of quantum

chemistry in recent decades. The most common programs used for quantum descriptor calculations were Spartan (4 times), AMPAC (3 times), e-Dragon (twice) and CLOGP (twice) (as shown on Table 1).

Via this review, several interesting practices were noticed in the use of substitutional descriptors. The first issue is associated with the type and number of substituents in DBPs. Often, activity changes were correlated with the changes in the number of different type of halogens (X, including F, Cl, Br, and I) [22] but rarely with the changes in functional groups (e.g., -COOH and $-NO_2$) or chain groups (e.g., $-CH_2-$ and $-CH_2-CH_2-$). This phenomenon suggests that early studies often examined only one group of DBPs at one time rather than considering multiple types together. A careful balance of specificity and generality of DBPs may need to be considered in the future [56]. The second interesting practice was the additivity methods that were used in descriptor calculations. Some researchers calculated the σ value of trichloro-compounds by simply tripling the σ value of a single chlorine [22, 57, 58]; but other researchers introduced correction factors [27]. A third observation relates to the choice of descriptor values. Sometimes multiple options are available for just one type of descriptor, and the optimal selection is uncertain [59]. For example, one study reported 0.45, 0.42, 0.45, and 0.42 [57] as the σ values for F, Cl, Br, and I, respectively, while another selected different values, i.e., 0.50, 0.47, 0.45, and 0.38, respectively, for the same descriptors [22]. In addition, compounds with ionizable substituents, such as –COOH (e.g. in the HAAs) may exist in either ionized or non-ionized forms with different electronic and steric values [59]. However, existing literature has only applied the non-charged form, even though HAAs are anions under typical pH conditions (pH 7-9) in drinking water treatment. Although these practices work well in a specific case, such inconsistencies may confuse those who are critical or less well versed in QSAR methodologies.

For this reason, caution should be exercised when making these decisions to optimize modeling outcomes, and protocols are needed to compare results from different scholars [60].

2.1.6 Resources for descriptors

Descriptor information is present in the scholarly literature, on websites, and in computer programs. Several classic handbooks list typical descriptor values for thousands of compounds [59, 61], and some literature use programs to compute descriptors [31, 62, 63]. Table 3 summarizes these resources. Of these, Dragon is perhaps the most popular software, which contains more than 5,000 molecular descriptors within 20 categories. For those who cannot afford the commercial product, a free online version of Dragon, called e-Dragon, allows the calculation of over 1,600 descriptors. In addition, the descriptor values of some compounds can be downloaded from an old Dragon version (version 5.4) at www.moleculardescriptors.eu/dataset/dataset.htm. For DBP scientists interested in the use of compositional descriptors, a series of functional group descriptor values for regulated and emerging DBPs, including carboxyl, carbonyl, nitrile, and halogens, are provided in Table 4, with more information available in the original references [59, 61].

2.2 Descriptor Filtration

2.2.1 Introduction to descriptor filtration

The purpose of descriptor filtration is to select appropriate and interpretable descriptors from the vast number of available options by either empirical judgment or computational analysis. This process is sometimes omitted in simple models. However, for more sophisticated computations, model development is more transparent if the descriptor filtration process is included. Because not all descriptors are useful in QSAR modeling and many descriptors often exhibit poor

correlation with activity or collinearity with other descriptors, using a descriptor filtration process can help reduce the complexity in dealing with too many descriptors and can lower the risk of over-training (i.e., spending excessive efforts to make negligible improvements) [64].

The most commonly used filtration processes include filter methods (e.g., factor analysis, Fisher score), wrapper methods (e.g., k-nearest neighbors [kNN], stepwise regression, replacement method [RM]), and hybrid methods. Filter methods are based on empirical judgment and relatively simple correlation analysis and can be further divided into two approaches: 1) internal filtration, which screens descriptors by collinearity of descriptors only without use of activity data; and 2) external filtration, which prioritizes descriptors using activity data. In contrast, wrapper methods entail a relatively complicated machine learning process (MLP) that requires trial-and-error iterations to filter out unnecessary descriptors [65]. To date, there are no protocols or standard practices regulating the descriptor filtration process; but in general, we believe that the following guidelines can be formulated:

1. The descriptors should be clearly informative of activity, and the definitions of descriptors should be as simple as possible. This guideline generally complies with the Organization for Economic Cooperation and Development (OECD) instruction to determine a "defined endpoint" [65].
2. The descriptor values should exhibit a reasonable trend for compounds with a range of chemical structures and/or composition. For example, descending σ values of halogens (e.g., 0.50, 0.47, 0.45, and 0.38 for F, Cl, Br, and I, respectively) are preferable than inconsistent sequence of halogen values (e.g., 0.45, 0.42, 0.45, and 0.42 for F, Cl, Br, and I, respectively) [66] although they can derive from the same data source [59].

3. The descriptors should be independent of each other or should exhibit no collinearity at a reasonable level to reduce information redundancy [41, 60]. For the multiple linear regression (MLR) algorithm, descriptors are expected to be linearly related to activity data (e.g., $p<0.05$); but for principal component regression (PCR), partial least square (PLS), and artificial neural network (ANN) algorithms, seemingly irrelevant descriptors may ultimately be chosen in final model because a combination of descriptors sometimes outperforms single descriptor in fitting results. As such, this rule is not always followed.
4. The number of descriptors should be as small as possible, with the number of descriptors usually ≤ 5 or $\leq 1/5$ of the number of activity data in a MLR model [67, 68]. Although the number of descriptors is unlimited for the PLS algorithm, a concise model is always preferred, no matter the conditions.

2.2.2 Existing descriptor filtration methods

Overall, 17 identified studies applied descriptor filtration methods in the development of QSAR models, whereas 14 references chose descriptors simply by experience. Nine studies applied filter methods, nine employed wrapper methods, and one used both filter and wrapper methods (note: the sum is greater than 17 because one literature applied both methods). The DBP transformation studies often lacked descriptor filtration processes, probably because they employed less activity data than other studies (Table 1). In contrast, research on DBP properties and toxicity often included filtration processes, with correlation analysis (CA) and principal component analysis (PCA) methods preferred.

The programs used to implement descriptor filtration included Microsoft Excel, statistical software (e.g., SPSS, Minitab), commercial QSAR programs (e.g., CODESSA), and

programming packages (e.g., FORTRAN) (Table 1). Nonetheless, descriptor filtration processes are seldom fully described in the literature, thus making the filtration process difficult to repeat.

2.2.3 Resources for descriptor filtration

Many articles describe descriptor filtration methods in detail [64, 65, 69]. These are not repeated here. Instead, the functionalities of programs commonly used for descriptor filtration are listed (Table 5). Statistical software (e.g., Minitab, SPSS, and SAS) encompass modules for descriptor filtration, such as enter-MLR, forward selection (FS), and backward elimination. It is noteworthy that SAS has nine descriptor filtration methods in additional to the maximum correlation and minimum correlation methods. Commercial QSAR programs have similar functionality to the statistical programs. For instance, CODESSA includes heuristic and best multi-linear regression methods for descriptor filtration. The online software OCHEM is equipped with PCA, CA, and FS processes [70]. Data-mining software such as Rapidminer also has descriptor filtration features like FS, backward elimination, and optimizing selection processes. These programs hence provide a diverse source of choices for descriptor filtration practices.

2.3 QSAR Models

2.3.1 Introduction of model algorithms

During model selection or formulation process, the task is to design, or more often, to select, an algorithm to mathematically describe the relations of descriptors and activities. Based on the level of complexity and the data availability, this process can be either quantitative or qualitative, with the former being the focus of this article. The classic algorithms used in QSAR models are the Hammett [71], Taft [72], and Hansch equations [73], which often apply electronic constant, steric constant, and hydrophobicity constants alone or in concert. As shown below in Equations

3-6, models can be made by single linear relationship (SLR), MLR, or multiple non-linear regression equations.

Hammett equation: $\mathrm{Log}(Y) = \rho \cdot \sigma + \text{constant}$ Equation 3

Taft equation: $\mathrm{Log}(Y) = \delta \cdot E_s + \rho \cdot \sigma + \text{constant}$ Equation 4

Hansch equation: $\mathrm{Log}(Y) = k \cdot \pi + k' \cdot \pi^2 + \delta \cdot E_s + \rho \cdot \sigma + \text{constant}$ Equation 5

$Y = b_0 + b_1 X_1^{\alpha} + b_2 X_2^{\beta} + \ldots + b_n X_n^{\gamma}$ Equation 6

Note: see symbol descriptions in the symbol list

Coupled with certain descriptor filtration processes, conventional MLR models can be upgraded to more advanced models like PCR, PLS, nonlinear PLS (n-PLS), support vector machine (SVM), and genetic algorithm (GA). MLR was ranked as the second most frequently used model in the Environmental Science Ecology area in recent decades due to its ease of use, understanding, and interpretation [28]. The top-ranked algorithm is PLS, which converts descriptors into new, latent variables first before applying them to a MLR algorithm. n-PLS, which modifies PLS to non-linear equation (e.g., Equation 6), is used occasionally [74, 75]. The use of PLS is especially suitable for complicated model development when the number of descriptors is greater than the number of activity data, such as for 3D-QSAR development with comparative molecular field analysis [28, 75]. Another popular QSAR method is ANN, which connects descriptors and activities with hidden neurons (called weighting factors). However, because no specific algorithm is obtained during the ANN process and it is impossible to explain the hidden layers explicitly, ANN has not been accepted widely, even though it often achieves better fitting than other types of algorithms.

In addition to using conventional MLR algorithms, another feasible option is to formulate new algorithms for QSAR models based upon Equation 6. Table 6 demonstrates several transformative algorithms of the general MLR. Although these algorithms have never been applied in existing DBP-QSAR models, the new options may improve future QSAR models.

2.3.2 Existing QSAR models

As some literature examined multiple types of models, a total of 40 models were investigated in this study. Most studies applied MLR (23 times) and SLR (8 times) in the QSAR models (Table 1), indicating that MLR is the most popular algorithm in existing QSAR-DBPs studies. By contrast, five studies used PLS and only two applied ANN and PCR, indicating that these methods in particular have been underexploited. It is noteworthy that there are two papers comparing the performance of different algorithms using the same activity data. In one example, Yoshida et al. used three different modeling methods (PLS, quadratic-PLS, and ANN) to model boiling points of 48 halomethanes, and identified ANN as the best fitting option [74]. In another example, Tang and colleagues demonstrated slight advantages of PLS over MLR and PCR [42]. These results thus confirm that PLS and ANN sometimes outperform MLR, even though the latter is more frequently applied.

The programs used to develop the abovementioned algorithms in existing literature sources were Excel, FORTRAN, SAS, Minitab, Origin, SPSS, and CODESSA (Table 1). Microsoft Excel appeared in 8 studies, suggesting that simple models were used in these cases. Dedicated statistical programs, such as Minitab and SPSS, were the second most popular choice [76, 77]. Two studies did not disclose which QSAR programs were used [78, 79].

2.3.3 Resources for QSAR modeling

Table 5 lists free and commercial programs for QSAR modeling. In general, the programs can be divided into three categories: commercial QSAR software, open-source data-mining software, and statistical analysis software. The commercial QSAR software, such as SYBYL-X, CODESSA, SciQSAR, and ISIDA/QSPR, are intended specifically for QSAR model development. These programs are preferred because they can be well integrated with other descriptor calculation programs. For example, CODESSA III is a state-of-the-art QSAR program that works directly with AMPAC, Gaussian, or AIMALL to generate models from quantum chemical, electrostatic, and topological descriptors. However, commercial QSAR software usually has fewer descriptor filtration options than statistical software. Data-mining programs such as WEKA, Rapidminer, and KNIME also have many descriptor filtration functions. In addition, data-mining programs are usually advantageous as their user-friendly, graphical operators can carry out feature selection, calibration, and validation processes in sequence. However, like commercial QSAR programs, data-mining programs have less statistical capability than dedicated statistical programs.

2.4 Model Validation

Validation is an integral part of the QSAR model development process; its purpose is to verify the predictivity of a model [80]. Validation is accomplished in two steps: data splitting (or allocation) and statistical analysis (Figure 2). In addition to goodness-of-fit, it is also necessary to identify the applicability domain (AD) and outliers. Although OECD has issued guidelines and protocols for statistical analysis [80], no consensus has been reached regarding preferred data splitting methods. Therefore, some general approaches will be briefly discussed here.

2.4.1 Data splitting

In this step, a database is typically divided into a training set and an external validation set. The training set can be further split into a calibration subset and an internal validation subset. The calibration set serves to calibrate model coefficients with the best statistical performances (namely, goodness-of-fit). The internal validation set serves to cross-validate the models developed by the calibration set. Unlike cross-validation, which may employ data derived from the same source, the external set only employs data never used during model development process. The importance and necessity of the external subset in the QSAR model development process have been underscored by the OECD guidelines [80].

The methods for data splitting include random data splitting, stratified random data splitting, and pseudo-stratified random data splitting [81]. The amount of external validation data usually accounts for > 30% of the total data [82], and both the calibration set and the internal validation set should have compounds representing the AD. Operations for data splitting include leave one out (LOO), leave many out (LMO, i.e., K-fold), and bootstrapping. LOO and LMO refer to a process of removing one or more data points (e.g., a certain percentage) from a database. The use of LOO is not preferred for a large pool of data because it may exaggerate the QSAR performance. By contrast, LMO is more popular because it has lower risk of inappropriate evaluation [83]. The number of samples included in LMO should be $\geq$ 2, and preferably large enough to enable sufficient rounds of training [84]. Bootstrapping offers an alternative allocation process to randomly group data for validation, and the process is usually repeated many times to verify the model [85]. Like LMO, bootstrapping is also an appropriate choice for studies with a large amount of activity data (e.g., >100) [86].

Another model validation technique is Y-randomization (or Y-scrambling). The process correlates randomly shuffled activity data with descriptors, and typically obtain lower correlation performance than the original model. If inappropriate descriptors, outliers, or algorithms are present, the Y-scrambling results will result in higher statistical performance that invalidates the model [87].

2.4.2 Statistical indicators

Statistical indicators serve as the criteria for evaluating the performance of a model. Guidelines available from the OECD can measure goodness-of-fit and predictivity of model in a standard formatted sheet [80]. The most common indicators include coefficient of determination (i.e., R^2), variance (e.g., RMSE, SD), data similarity (e.g., p value, student-t value), predictivity (i.e., Q^2), AD (e.g., h_i), and outlier detection, each of which may be determined by a range of parameters (Table 7).

With respect to thresholds, both R^2 and Q^2 were recommended to be greater than 0.6 [83]. Concordance correlation coefficient (CCC) was proposed to determine the external predictivity, with a recommended threshold of >0.85 [83]. The slopes k and k' were suggested to be near 1.0, and the ratio of $(R_i^2-R_0^2)/R_i^2$ to be close to zero [83, 88]. Standardized residual values were recommended to be less than 2x [89], 2.5x [90], or 3x [91] of the standard deviation, which are equivalent to 90%, 95%, or 99% levels of confidence, respectively. Another study suggested that a model may be considered acceptable if it satisfies the following conditions: $R^2>0.6$, $Q^2>0.5$, $(R_i^2-R_0^2)/R_i^2<0.1$ (or $(R_i^2-R_0^2)/R_i'^2<0.1$) and $0.85\leq m\leq 1.15$ (or $0.85\leq m'\leq 1.15$) [88]. In certain

circumstances, an adjusted R^2, R^2_{adj} may be used as an increase of variable numbers always leads to an increasing R^2 value [80].

2.4.3 Applicability domain (AD) and outliers

AD is defined as the paradigm set by representative compounds for model calibration. The purpose of AD analysis is to avoid inaccurate predictions for compounds outside the domain. Four methods are commonly used for identifying the AD [92, 93], with the density distribution-based method considered the most sophisticated [80]. Compounds with unexpected, poorly-fitted activity or compounds falling outside the AD in a QSAR model are considered outliers [94, 95]. The leverage "h" is a measure of the distance of the compound from the centroid of its training set. If a data point is greater than the warning leverage h*, it can be considered as either an outlier or an influential point [96]. Cook's distance and standardized residual (SDR) are two other statistical methods to detect outliers [33]. Cook's distance measures how much the total residual would change if a compound is excluded from the calibration. In one relevant study, the operational thresholds $D_i > 1$ and $|SDR_i| > 2.5$ were used [33]. Moreover, the distance-based method using Williams plot, as defined by SDRs versus leverage h, is often applied to identify outliers and influential points [43, 90].

2.4.4 Existing validation approaches

In Table 1, the column "no. & types of compounds" shows how many and what types of DBPs were selected for QSAR model development. All selected studies have employed DBPs in their training sets. The column "data splitting approach" indicates which models/literatures have applied validation methods and which have not. The criteria regarding if the model is validated or not is to see if extra data sets are applied to justify the model developed by training set.

Therefore, if it shows “NA (not available)” in this column, it means that the literature did not employ or disclose any validation method. From the literature review, we found 12 articles that apply validation method(s) during QSAR model development. Cross-validation and bootstrapping were used 14 and 5 times, respectively, with two references applying both. Within cross-validation methods, the LOO method was applied 8 times, and LMO was applied 6 times. External validation was implemented in 10 cases. However, there were also some studies that did not implement any data splitting processes, indicating this phase of method development is not fully implemented by some researchers.

The most commonly used statistical indicators were R^2, SD, SE, and F-ratio (Table 1). Q^2 and RMSE were the two most popular indicators used for predictivity analysis. Meanwhile, six references analyzed outliers and the AD, with the distance-based method favored in AD analysis. The leverage h and SDRs were the two parameters most frequently used to make the Williams plot. Cook’s distance, D, was applied twice to detect outliers whereas the density-based AD method was not observed in any literature. In additional to conventional indicators, selected studies have employed novel methods such as self-organizing map [78] for AD analysis.

2.4.5 Resources for validation

Table 5 lists some software programs (such as SPSS and SAS) that allow statistical analysis of model performance. These also serve as validation tools when validation is carried out automatically during modeling. For example, cross-validation and bootstrapping can be carried out by invoking the “ncvreg” and “bootstrap” modes in the statistical software “R”. Similarly, Minitab can implement many validation processes. Rapidminer, a data-mining software, also has validation operators for cross-validation and bootstrapping.

3. Future needs

In this section, we propose several needs in DBP research that may merit future exploration through the use of QSAR models. These ideas are intended to stimulate additional research and should not be deemed exclusive or complete.

The first, and probably the most immediate, need is to develop a DBP information harbor or platform for data-sharing and accumulation. Currently, activity data are sporadically and sparsely present in a large pool of sources, which requires intensive time and effort to be spent on data searching and pretreatment. To attain its important time-saving potential, the harbor should include not only original data, but also protocols for descriptor filtration and model development, which are often lacking in existing studies. In addition to encouraging more frequent use of QSAR techniques to study DBPs, this may also aid researchers to optimize QSAR models in the future, ultimately leading to better understanding of DBP issues.

The second need concerns the characterization of DBP precursors. To date, abundant information has been published on the experimental, bulk characteristics of DBP precursors[97], such as UV and specific UV absorbance (SUVA, an indicator of aromaticity), molecular size (or MW, by size-exclusion chromatography), polarity (i.e., hydrophobicity by adsorptive resin), functional groups (e.g., carboxyl group, by fluorescence excitation-emission matrix, nuclear magnetic resonance, and Fourier transform infrared spectroscopy analyses) [98, 99], and reactivity with certain disinfectants [100]. However, a QSAR model has not yet been developed to correlate the functional groups and properties of a comprehensive range of specific precursors or isolates with DBP formation. The complex and largely unknown identity of NOM makes this inherently

difficult, but becomes more feasible when considering model compounds, with defined molecular properties, such as that has already been achieved for chlorine demand [24]. It is anticipated that such QSAR models will offer more practical and timely recommendations to water providers by bridging the gap between DBPFP and precursor characteristics. Ideally, the DBP precursors would not be limited to a small number model compounds (e.g., amino acids) but would cover diverse and representative types of organic matter.

The third need concerns the stability and degradability of DBPs in order to alleviate the risks of DBPs to drinking water consumers and aquatic organisms. As discussed, early studies in DBP transformation often considered only halogens as the descriptors in the QSAR model and rarely integrated other substituents, such as nitrile (-CN), carboxyl (-COOH) functional group, or carbon chain, into the models. Although such models are successful for a single class of DBPs, they lack insights into the behavioral differences among multiple types of DBPs. Therefore, we suggest that future QSAR studies on DBP degradation incorporate more compounds with various types and numbers of halogens, functional groups, and chain structures. It is likely that a more sophisticated approach may lead to a more widely accepted model.

4. Conclusions

In an effort to facilitate greater use of QSAR models in DBP research, we summarized some trends in 31 studies that developed QSARs for DBP research, including the preferred methods for activity and descriptor collection, descriptor filtration, model selection, and validation. The most commonly used descriptors in previous DBP-QSAR studies were compositional and quantum chemical descriptors; MLR was the main algorithm; and correlation coefficients and standard deviations were the two most frequently applied statistical indicators. However, a

review of existing literature also revealed a lack of consensus in methods for handling data, descriptors, algorithms, and statistical indicators. Additionally, many studies displayed methodological weaknesses (e.g., the absence of cross-validation) and too narrow a focus (e.g., the degradation of a single class of DBPs rather than an ensemble of multiple DBPs), which may hinder a wide acceptance and holistic view of the DBP behaviors. Researchers should be aware of the implications and limitations of the decisions in each step of the QSAR development workflow. Typical concerns regarding QSAR model development processes are listed in Table 8.

Several barriers to the adoption of QSAR models in DBP research were also identified: 1) the limited availability of DBP data, with intensive effort required to collect limited data from a large pool of sources; 2) the lack of user-friendly QSAR software, which represents a major barrier to scientists unfamiliar with the field; and 3) the difficulty of designing QSAR models for DBP issues. One particular priority to aid future QSAR-DBP studies is to develop a platform for sharing DBP data and descriptors. In addition, benefits in using QSAR models can potentially be gained from the characterization of DBP precursors as well as the degradation of DBPs. Overall, although QSAR models will not replace in situ or laboratory experiments, they could play a more important role in supplementing, extending, and promoting our future understanding of DBP issues.

5. Acknowledgements

The study was supported by the National Natural Science Foundation of China (51278144), the Shenzhen Science & Technology R&D Funding (JCYJ20120613150442560), and the Funding for Returned Oversea Chinese from Shenzhen (KQCX20130627094615414). Thanks go to Drs. Paul Westerhoff and Yu Yang for assistance in accessing certain references, Dr. Peng Wang for

his critical comments, and James Stewart for providing a free copy of MOPAC program for evaluation.

6. References

[1] S.D. Richardson, Disinfection by-products and other emerging contaminants in drinking water, Trends in analytical chemistry, 22 (2003) 666-684.
[2] S.W. Krasner, H.S. Weinberg, S.D. Richardson, S.J. Pastor, R. Chinn, M.J. Sclimenti, G.D. Onstad, A.D. Thruston, Occurrence of a New Generation of Disinfection Byproducts, Environ Sci Technol, 40 (2006) 7175-7185.
[3] J.J. Rook, Formation of haloforms during chlorination of natural waters, Water Treatment Examination 23 (1974) 234-243.
[4] C.M. Villanueva, G. Castaño-Vinyals, V. Moreno, G. Carrasco-Turigas, N. Aragonés, E. Boldo, E. Ardanaz, E. Toledo, J.M. Altzibar, I. Zaldua, L. Azpiroz, F. Goñi, A. Tardón, A.J. Molina, V. Martín, C. López-Rojo, J.J. Jiménez-Moleón, R. Capelo, I. Gómez-Acebo, R. Peiró, M. Ripoll, E. Gracia-Lavedan, M.J. Nieuwenhujsen, P. Rantakokko, E.H. Goslan, M. Pollán, M. Kogevinas, Concentrations and correlations of disinfection by-products in municipal drinking water from an exposure assessment perspective, Environ Res, 114 (2012) 1-11.
[5] J. Grellier, J. Bennett, E. Patelarou, R.B. Smith, M.B. Toledano, L. Rushton, D.J. Briggs, M.J. Nieuwenhuijsen, Exposure to Disinfection By-products, Fetal Growth, and Prematurity: A Systematic Review and Meta-analysis, Epidemiology, 21 (2010) 300-313.
[6] M.J. Nieuwenhuijsen, M.B. Toledano, N.E. Eaton, J. Fawell, P. Elliott, Chlorination disinfection byproducts in water and their association with adverse reproductive outcomes: a review, Occupational and Environmental Medicine, 57 (2000) 73-85.
[7] S.D. Richardson, Disinfection By-Products: Formation and Occurrence in Drinking Water, in: J.O. Nriagu (Ed.) Encyclopedia of Environmental Health, Elsevier, Burlington, 2011, pp. 110-136.
[8] S.D. Richardson, M.J. Plewa, E.D. Wagner, R. Schoeny, D.M. DeMarini, Occurrence, genotoxicity, and carcinogenicity of regulated and emerging disinfection by-products in drinking water: A review and roadmap for research, Mutation Research/Reviews in Mutation Research, 636 (2007) 178-242.
[9] T. Bond, M.R. Templeton, N. Graham, Precursors of nitrogenous disinfection by-products in drinking water—A critical review and analysis, J. Hazard. Mater., 235–236 (2012) 1-16.
[10] T. Bond, J. Huang, M.R. Templeton, N. Graham, Occurrence and control of nitrogenous disinfection by-products in drinking water - A review, Water Res, 45 (2011) 4341-4354.
[11] S.W. Krasner, P. Westerhoff, B. Chen, B.E. Rittmann, G. Amy, Occurrence of Disinfection Byproducts in United States Wastewater Treatment Plant Effluents, Environ Sci Technol, 43 (2009) 8320-8325.
[12] H.S. Weinberg, Modern approaches to the analysis of disinfection by-products in drinking water, Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences, 367 (2009) 4097-4118.
[13] S.W. Krasner, W.A. Mitch, D.L. McCurry, D. Hanigan, P. Westerhoff, Formation, precursors, control, and occurrence of nitrosamines in drinking water: A review, Water Res, 47 (2013) 4433-4450.
[14] D.B. Jones, A. Saglam, H. Song, T. Karanfil, The impact of bromide/iodide concentration and ratio on iodinated trihalomethane formation and speciation, Water Res, 46 (2012) 11-20.
[15] D.B. Jones, A. Saglam, A. Triger, H. Song, T. Karanfil, I-THM Formation and Speciation: Preformed Monochloramine versus Prechlorination Followed by Ammonia Addition, Environ Sci Technol, 45 (2011) 10429-10437.
[16] J. Criquet, S. Allard, E. Salhi, C.A. Joll, A. Heitz, U. von Gunten, Iodate and Iodo-Trihalomethane Formation during Chlorination of Iodide-Containing Waters: Role of Bromide, Environ Sci Technol, 46 (2012) 7350-7357.
[17] M.J. Plewa, E.D. Wagner, P. Jazwierska, S.D. Richardson, P.H. Chen, A.B. McKague, Halonitromethane drinking water disinfection byproducts: Chemical characterization and mammalian cell cytotoxicity and genotoxicity, Environ Sci Technol, 38 (2004) 62-68.
[18] M.J. Plewa, E.D. Wagner, S.D. Richardson, A.D. Thruston, Y.-T. Woo, A.B. McKague, Chemical and Biological

Characterization of Newly Discovered Iodoacid Drinking Water Disinfection Byproducts, Environ Sci Technol, 38 (2004) 4713-4722.
[19] M.J. Plewa, M.G. Muellner, S.D. Richardson, F. Fasano, K.M. Buettner, Y.-T. Woo, A.B. McKague, E.D. Wagner, Occurrence, Synthesis, and Mammalian Cell Cytotoxicity and Genotoxicity of Haloacetamides: An Emerging Class of Nitrogenous Drinking Water Disinfection Byproducts, Environ Sci Technol, 42 (2008) 955-961.
[20] S.D. Richardson, F. Fasano, J.J. Ellington, F.G. Crumley, K.M. Buettner, J.J. Evans, B.C. Blount, L.K. Silva, T.J. Waite, G.W. Luther, A.B. McKague, R.J. Miltner, E.D. Wagner, M.J. Plewa, Occurrence and Mammalian Cell Toxicity of Iodinated Disinfection Byproducts in Drinking Water, Environ Sci Technol, 42 (2008) 8330-8338.
[21] J. Plewa Michael, D. Wagner Elizabeth, G. Muellner Mark, K.-M. Hsu, D. Richardson Susan, Comparative Mammalian Cell Toxicity of N-DBPs and C-DBPs, in: Disinfection By-Products in Drinking Water, American Chemical Society, 2008, pp. 36-50.
[22] B. Chen, Hydrolytic Stabilities of Halogenated Disinfection Byproducts: Review and Rate Constant Quantitative Structure–Property Relationship Analysis, Environ Eng Sci, 28 (2011) 385-394.
[23] T. Ye, B. Xu, Y.-L. Lin, C.-Y. Hu, S.-J. Xia, L. Lin, S.A. Mwakagenda, N.-Y. Gao, Formation of iodinated disinfection by-products during oxidation of iodide-containing water with potassium permanganate, J. Hazard. Mater., 241–242 (2012) 348-354.
[24] G.B. Luilo, S.E. Cabaniss, Quantitative Structure-Property Relationship for Predicting Chlorine Demand by Organic Molecules, Environ Sci Technol, 44 (2010) 2503-2508.
[25] D. Juretic, H. Kusic, D.D. Dionysiou, B. Rasulev, A. Loncaric Bozic, Modeling of photooxidative degradation of aromatics in water matrix; combination of mechanistic and structural-relationship approach, Chemical Engineering Journal, 257 (2014) 229-241.
[26] V. Gombar, Quantitative structure-activity relationships in toxicology: from fundamentals to applications, Adv. Mol. Toxicol, 10 (1997) 125-139.
[27] V. Glezer, B. Harris, N. Tal, B. Iosefzon, O. Lev, Hydrolysis of haloacetonitriles: linear free energy relationship, kinetics and products, Water Res, 33 (1999) 1938-1948.
[28] M. Dehmer, K. Varmuza, D. Bonchev, F. Emmert-Streib, Statistical modelling of molecular descriptors in QSAR/QSPR, John Wiley & Sons, 2012.
[29] C. Hansch, A. Leo, Exploring QSAR: fundamentals and applications in chemistry and biology, American Chemical Society, 1995.
[30] G.D. Veith, D.E. Konasewich, Structure-activity correlations in studies of toxicity and bioconcentration with aquatic organisms, in: Symposium on Structure-Activity Correlations in Studies of Toxicity and Bioconcentration with Aquatic Organisms (1975: Burlington, Ont.), Great Lakes Research Advisory Board, 1975.
[31] W.J. Lyman, W.F. Reehl, D.H. Rosenblatt, Handbook of Chemical Property Estimation Methods, McGraw-Hill, 1982.
[32] K. Tuppurainen, S. Lötjönen, R. Laatikainen, T. Vartiainen, U. Maran, M. Strandberg, T. Tamm, About the mutagenicity of chlorine-substituted furanones and halopropenals. A QSAR study using molecular orbital indices, Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 247 (1991) 97-102.
[33] G. Luilo, Quantitative structure-property relationships for predicting chlorine demand and disinfection byproducts formation in drinking water (Dissertation), in, The University of New Mexico, 2011.
[34] M.G. Muellner, E.D. Wagner, K. McCalla, S.D. Richardson, Y.-T. Woo, M.J. Plewa, Haloacetonitriles vs. Regulated Haloacetic Acids: Are Nitrogen-Containing DBPs More Toxic?, Environ Sci Technol, 41 (2007) 645-651.
[35] Y.-T. Woo, D. Lai, J.L. McLain, M.K. Manibusan, V. Dellarco, Use of mechanism-based structure-activity relationships analysis in carcinogenic potential ranking for drinking water disinfection by-prodcuts, Environ. Health Perspect., 110 (2002) 75-87.
[36] J.B. Richard, A.R. David, Use of Chemical Models and Structure-Activity Relationships to Identify Novel Disinfection By-Products of Potential Toxicological Concern, in: Disinfection By-Products in Drinking Water, American Chemical Society, 2008, pp. 51-64.
[37] W. Jin, J. Zhou, B. Chen, X. Zhu, C. Cui, Modeling volatilization and adsorption of disinfection byproducts in natural watersheds, J. Environ. Monit., 14 (2012) 2990-2999.
[38] J. Sohn, D. Gatel, G. Amy, Monitoring and modeling of disinfection by-products (DBPs), Environ. Monit. Assess., 70 (2001) 211-222.
[39] B. Chen, P. Westerhoff, Predicting disinfection by-product formation potential in water, Water Res, 44 (2010) 3755-3762.
[40] S. Chowdhury, P. Champagne, P.J. McLellan, Models for predicting disinfection byproduct (DBP) formation in

drinking waters: A chronological review, Sci. Total Environ., 407 (2009) 4189-4206.
[41] M.T.D. Cronin, T.W. Schultz, Pitfalls in QSAR, Journal of Molecular Structure: THEOCHEM, 622 (2003) 39-51.
[42] W.Z. Tang, F. Wang, Quantitative structure activity relationship (QSAR) of chlorine effects on ELUMO of disinfection by-product: Chlorinated alkanes, Chemosphere, 78 (2010) 914-921.
[43] G. Luilo, S. Cabaniss, QSPR for predicting chloroform formation in drinking water disinfection, SAR and QSAR in Environmental Research, 22 (2011) 489-504.
[44] N. Fjodorova, M. Vracko, M. Tusar, A. Jezierska, M. Novic, R. Kuhne, G. Schuurmann, Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses, Molecular Diversity, 14 (2010) 581-594.
[45] A.M. Richard, Commercial toxicology prediction systems: a regulatory perspective, Toxicol. Lett., 102–103 (1998) 611-616.
[46] E. Benfenati, R. Benigni, D.M. Demarini, C. Helma, D. Kirkland, T.M. Martin, P. Mazzatorta, G. OuÉDraogo-Arras, A.M. Richard, B. Schilter, W.G.E.J. Schoonen, R.D. Snyder, C. Yang, Predictive Models for Carcinogenicity and Mutagenicity: Frameworks, State-of-the-Art, and Perspectives, Journal of Environmental Science and Health, Part C, 27 (2009) 57-90.
[47] R.J. Kavlock, G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, E. Hubal, R. Judson, J. Rabinowitz, A. Richard, R.W. Setzer, I. Shah, D. Villeneuve, E. Weber, Computational Toxicology—A State of the Science Mini Review, Toxicol Sci, 103 (2008) 14-27.
[48] R. Serafimova, M. Fuart Gatnik, A. Worth, Review of QSAR models and software tools for predicting genotoxicity and carcinogenicity, JRC Scientific and Technical Report EUR 24427 EN-2010, (2010).
[49] Q. Bu, D. Wang, Z. Wang, Review of Screening Systems for Prioritizing Chemical Substances, Critical Reviews in Environmental Science and Technology, 43 (2012) 1011-1041.
[50] M. Pavan, A.P. Worth, Publicly-accessible QSAR software tools developed by the Joint Research Centre, Sar and Qsar in Environmental Research, 19 (2008) 785-799.
[51] E.S. Hunter, E. Rogers, M. Blanton, A. Richard, N. Chernoff, Bromochloro-haloacetic acids: Effects on mouse embryos in vitro and QSAR considerations, Reproductive Toxicology, 21 (2006) 260-266.
[52] A.M. Richard, E.S. Hunter, Quantitative structure-activity relationships for the developmental toxicity of haloacetic acids in mammalian whole embryo culture, Teratology, 53 (1996) 352-360.
[53] H. Yang, Z. Jiang, Y. Guan, W.Z. Tang, A.J. Pierotti, Chlorine Effect on Disinfect By Products Safety Lived In Drinking Water (in Chinese), Water & Wastewater Engineering, 26 (2000) 22-25.
[54] M. Yang, X. Zhang, Comparative Developmental Toxicity of New Aromatic Halogenated DBPs in a Chlorinated Saline Sewage Effluent to the Marine Polychaete Platynereis dumerilii, Environ Sci Technol, 47 (2013) 10868-10876.
[55] J. Liu, X. Zhang, Comparative toxicity of new halophenolic DBPs in chlorinated saline wastewater effluents against a marine alga: Halophenolic DBPs are generally more toxic than haloaliphatic ones, Water Res, 65 (2014) 64-72.
[56] R. Benigni, A.M. Richard, QSARS of mutagens and carcinogens: Two case studies illustrating problems in the construction of models for noncongeneric chemicals, Mutation Research/Genetic Toxicology, 371 (1996) 29-46.
[57] S.O. Pehkonen, Z. Qi, The Degradation of Organophosphorus Pesticides in Natural Waters:A Critical Review, Critical Reviews in Environmental Science & Technology, 32 (2002) 17.
[58] Y. Xiao, R. Fan, L. Zhang, J. Yue, R.D. Webster, T.-T. Lim, Photodegradation of iodinated trihalomethanes in aqueous solution by UV 254 irradiation, Water Res, 49 (2014) 275-285.
[59] C. Hansch, A. Leo, D. Hoekman, A. Leo, Exploring QSAR: hydrophobic, electronic, and steric constants, American Chemical Society Washington, DC, 1995.
[60] A. Cherkasov, E.N. Muratov, D. Fourches, A. Varnek, I.I. Baskin, M. Cronin, J. Dearden, P. Gramatica, Y.C. Martin, R. Todeschini, V. Consonni, V.E. Kuz'min, R. Cramer, R. Benigni, C. Yang, J. Rathman, L. Terfloth, J. Gasteiger, A. Richard, A. Tropsha, QSAR Modeling: Where Have You Been? Where Are You Going To?, J. Med. Chem., 57 (2014) 4977-5010.
[61] R. Todeschini, V. Consonni, Handbook of molecular descriptors, John Wiley & Sons, 2008.
[62] R.S. Boethling, D. Mackay, Handbook of property estimation methods for chemicals environmental and health sciences, Lewis Publishers, 2000.
[63] P.H. Howard, W.M. Meylan, Handbook of Physical Properties of Organic Chemicals, Lewis Publisher, Boca Raton, FL, 1997.
[64] M. Goodarzi, Y.V. Heyden, S. Funar-Timofei, Towards better understanding of feature-selection or reduction techniques for Quantitative Structure–Activity Relationship models, TrAC, Trends Anal. Chem., 42 (2013) 49-63.

[65] M. Shahlaei, Descriptor selection methods in quantitative structure–activity relationship studies: a review study, Chem. Rev., 113 (2013) 8093-8103.
[66] X. Zhang, R.A. Minear, Decomposition of trihaloacetic acids and formation of the corresponding trihalomethanes in drinking water, Water Res, 36 (2002) 3665-3673.
[67] J.G. Topliss, R.P. Edwards, Chance factors in studies of quantitative structure-activity relationships, J. Med. Chem., 22 (1979) 1238-1244.
[68] R.A. Francis, J.M. Vanbriesen, M.J. Small, Bayesian Statistical Modeling of Disinfection Byproduct (DBP) Bromine Incorporation in the ICR Database, Environ Sci Technol, 44 (2010) 1232-1239.
[69] Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, bioinformatics, 23 (2007) 2507-2517.
[70] I. Sushko, S. Novotarskyi, R. Korner, A.K. Pandey, M. Rupp, W. Teetz, S. Brandmaier, A. Abdelaziz, V.V. Prokopenko, V.Y. Tanchuk, R. Todeschini, A. Varnek, G. Marcou, P. Ertl, V. Potemkin, M. Grishina, J. Gasteiger, C. Schwab, Baskin, II, V.A. Palyulin, E.V. Radchenko, W.J. Welsh, V. Kholodovych, D. Chekmarev, A. Cherkasov, J. Aires-de-Sousa, Q.Y. Zhang, A. Bender, F. Nigsch, L. Patiny, A. Williams, V. Tkachenko, I.V. Tetko, Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information, J. Comput. Aided Mol. Des., 25 (2011) 533-554.
[71] L.P. Hammett, Some Relations between Reaction Rates and Equilibrium Constants, Chem. Rev., 17 (1935) 125-136.
[72] R.W. Taft, Steric effects in organic chemistry, J. Am. Chem. Soc., 58 (1956) 2436.
[73] A.J. Leo, C. Hansch, Linear free energy relations between partitioning solvent systems, The Journal of Organic Chemistry, 36 (1971) 1539-1544.
[74] H. Yoshida, Y. Miyashita, S. Sasaki, Modelling of halomethanes using neural networks, Chemometrics and intelligent laboratory systems, 32 (1996) 193-199.
[75] S.J. Bang, S.J. Cho, Comparative molecular field analysis (CoMFA) and comparative molecular similarity index analysis (CoMSIA) study of mutagen X, Bull. Korean Chem. Soc., 25 (2004) 1525-1530.
[76] E.J. Delgado, Predicting aqueous solubility of chlorinated hydrocarbons from molecular structure, Fluid Phase Equilib., 199 (2002) 101-107.
[77] Y. Wang, J.H. Qu, R.C. Wu, P.J. Lei, The electrocatalytic reduction of nitrate in water on Pd/Sn-modified activated carbon fiber electrode, Water Res, 40 (2006) 1224-1232.
[78] D. Libotean, J. Giralt, R. Rallo, Y. Cohen, F. Giralt, H.F. Ridgway, G. Rodriguez, D. Phipps, Organic compounds passage through RO membranes, J. Membr. Sci., 313 (2008) 23-43.
[79] P.A. Neale, B.I. Escher, K.-U. Goss, S. Endo, Evaluating dissolved organic carbon–water partitioning using polyparameter linear free energy relationships: Implications for the fate of disinfection by-products, Water Res, 46 (2012) 3637-3645.
[80] OECD, Guidance document on the validation of (quantitative) structure-activity relationships [(Q) SAR] models, in: OECD Series on Testing and Assessment No. 69. ENV/JM/MONO (2007) 2, 2007.
[81] F.E. Harrell, Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis, Springer, 2001.
[82] P. Gramatica, Principles of QSAR models validation: internal and external, QSAR & Combinatorial Science, 26 (2007) 694-701.
[83] N. Chirico, P. Gramatica, 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 (2011) 2320-2335.
[84] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, in: International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1145.
[85] R.D. Cramer, J.D. Bunce, D.E. Patterson, I.E. Frank, Cross validation, Bootstrapping, and Partial Least Squares Compared with Multiple Regression in Conventional QSAR Studies, Quant. Struct.-Act. Relat., 7 (1988) 18-25.
[86] A. Isaksson, M. Wallman, H. Goransson, M.G. Gustafsson, Cross-validation and bootstrapping are unreliable in small sample classification, Pattern Recognition Letters, 29 (2008) 1960-1965.
[87] C. Rücker, G. Rücker, M. Meringer, y-Randomization and Its Variants in QSPR/QSAR, Journal of Chemical Information and Modeling, 47 (2007) 2345-2357.
[88] A. Tropsha, P. Gramatica, V.K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, QSAR & Combinatorial Science, 22 (2003) 69-77.
[89] A. Pérez-Garrido, M.P. González, A.G. Escudero, Halogenated derivatives QSAR model using spectral moments

to predict haloacetic acids (HAA) mutagenicity, Biorg. Med. Chem., 16 (2008) 5720-5732.
[90] G.B. Luilo, S.E. Cabaniss, Predicting total organic halide formation from drinking water chlorination using quantitative structure-property relationships, SAR QSAR Environ Res, 22 (2011) 667-680.
[91] F. Xiao, J.S. Gulliver, M.F. Simcik, Predicting aqueous solubility of environmentally relevant compounds from molecular features: A simple but highly effective four-dimensional model based on Project to Latent Structures, Water Res, 47 (2013) 5362-5370.
[92] T.I. Netzeva, A.P. Worth, T. Aldenberg, R. Benigni, M.T.D. Cronin, P. Gramatica, J.S. Jaworska, S. Kahn, G. Klopman, C.A. Marchant, G. Myatt, N. Nikolova-Jeliazkova, G.Y. Patlewicz, R. Perkins, D.W. Roberts, T.W. Schultz, D.T. Stanton, J.J.M. van de Sandt, W.D. Tong, G. Veith, C.H. Yang, Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships - The report and recommendations of ECVAM Workshop 52, Atla-Alternatives to Laboratory Animals, 33 (2005) 155-173.
[93] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of Different Approaches to Define the Applicability Domain of QSAR Models, Molecules, 17 (2012) 4791-4810.
[94] E. Furusjö, A. Svenson, M. Rahmberg, M. Andersson, The importance of outlier detection and training set selection for reliable environmental QSAR predictions, Chemosphere, 63 (2006) 99-108.
[95] G.M. Maggiora, On outliers and activity cliffs why QSAR often disappoints., Journal of chemical information and modeling, 46 (2006) 1535-1535.
[96] R.P. Verma, C. Hansch, An approach toward the problem of outliers in QSAR, Biorg. Med. Chem., 13 (2005) 4597-4621.
[97] T. Bond, E.H. Goslan, S.A. Parsons, B. Jefferson, A critical review of trihalomethane and haloacetic acid formation from natural organic matter surrogates, Environmental Technology Reviews, 1 (2012) 93-113.
[98] A. Matilainen, E.T. Gjessing, T. Lahtinen, L. Hed, A. Bhatnagar, M. Sillanpää, An overview of the methods used in the characterisation of natural organic matter (NOM) in relation to drinking water treatment, Chemosphere, 83 (2011) 1431-1442.
[99] A.T. Chow, S. Gao, R.A. Dahlgren, Physical and chemical fractionation of dissolved organic matter and trihalomethane precursors: A review, J Water Supply Res T, 54 (2005) 475-507.
[100] P. Westerhoff, P. Chao, H. Mash, Reactivity of natural organic matter with aqueous chlorine and bromine, Water Res, 38 (2004) 1502-1513.
[101] K. Tuppurainen, S. Lotjonen, R. Laatikainen, T. Vartiainen, Structural and electronic-properties of mx compounds related to ta100 mutagenicity - a semiempirical molecular-orbital QSAR study, Mutation research, 266 (1992) 181-188.
[102] K. Tuppurainen, QSAR approach to molecular mutagenicity. A survey and a case study: MX compounds, Journal of Molecular Structure: THEOCHEM, 112 (1994) 49-56.
[103] K.L. Rule, V.R. Ebbett, P.J. Vikesland, Formation of chloroform and chlorinated organics by free-chlorine-mediated oxidation of triclosan, Environ Sci Technol, 39 (2005) 3176-3185.
[104] H. Gallard, U. von Gunten, Chlorination of Phenols: Kinetics and Formation of Chloroform, Environ Sci Technol, 36 (2002) 884-890.
[105] W. Chu, N. Gao, Y. Deng, Stability of newfound nitrogenous disinfection by-products haloacetamides in drinking water, Chin. J. Org. Chem, 29 (2009) 1569-1574.
[106] B. Chen, W. Lee, P.K. Westerhoff, S.W. Krasner, P. Herckes, Solar photolysis kinetics of disinfection byproducts, Water Res, 44 (2010) 3401-3409.
[107] S. Sudhakaran, G.L. Amy, QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification, Water Res, 47 (2013) 1111-1122.
[108] A.T. Balaban, N. Joshi, L.B. Kier, L.H. Hall, Correlations between chemical structure and normal boiling points of halogenated alkanes C1-C4, Journal Of Chemical Information And Computer Sciences, 32 (1992) 233-237.
[109] R.Z. Chen, Development, validation and uncertainty analysis of quantitative structure and activity relationship models for Log P of disinfection by-products (Disseration), in, Florida International University, 2009, pp. 1-176.
[110] E. Benfenati, G. Gini, N. Piclin, A. Roncaglioni, M.R. Vari, Predicting log P of pesticides using different software, Chemosphere, 53 (2003) 1155-1164.
[111] C. Helma, Lazy structure-activity relationships (lazar) for the prediction of rodent carcinogenicity and Salmonella mutagenicity, Molecular Diversity, 10 (2006) 147-158.
[112] C. Tebby, E. Mombelli, P. Pandard, A.R.R. Pery, Exploring an ecotoxicity database with the OECD (Q)SAR Toolbox and DRAGON descriptors in order to prioritise testing on algae, daphnids, and fish, Sci. Total Environ., 409 (2011) 3334-3343.

[113] J.C. Dearden, In silico prediction of drug toxicity, J. Comput. Aided Mol. Des., 17 (2003) 119-127.
[114] R.D. Combes, Is phenylbutazone a genotoxic carcinogen? A weight-of-evidence assessment, Alternatives to laboratory animals : ATLA, 41 (2013) 235-248.
[115] H. Sanderson, M. Thomsen, Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action, Toxicol. Lett., 187 (2009) 84-93.
[116] R.J. Bull, D.A. Reckhow, X. Li, A.R. Humpage, C. Joll, S.E. Hrudey, Potential carcinogenic hazards of non-regulated disinfection by-products: haloquinones, halo-cyclopentene and cyclohexene derivatives, N-halamines, halonitriles, and heterocyclic amines, Toxicology, 286 (2011) 1-19.
[117] S.M. Evans, A. Casartelli, E. Herreros, D.T. Minnick, C. Day, E. George, C. Westmoreland, Development of a high throughput in vitro toxicity screen predictive of high acute in vivo toxic potential, Toxicology in Vitro, 15 (2001) 579-584.
[118] R. Loomba, A. Rowley, R. Wesley, T.J. Liang, J.H. Hoofnagle, F. Pucino, G. Csako, Systematic review: The effect of preventive lamivudine on hepatitis B reactivation during chemotherapy, Annals of Internal Medicine, 148 (2008) 519-528.
[119] H.P.H. Arp, C. Niederer, K.U. Goss, Predicting the partitioning behavior of various highly fluorinated compounds, Environ Sci Technol, 40 (2006) 7298-7304.
[120] I. Linkov, M.R. Ames, E.A. Crouch, F.K. Satterstrom, Uncertainty in octanol-water partition coefficient: Implications for risk assessment and remedial costs, Environ Sci Technol, 39 (2005) 6917-6922.
[121] A. Mulero, M.I. Parra, I. Cachadina, The Somayajulu correlation for the surface tension revisited, Fluid Phase Equilib., 339 (2013) 81-88.
[122] A. Lagunin, A. Zakharov, D. Filimonov, V. Poroikov, QSAR Modelling of Rat Acute Toxicity on the Basis of PASS Prediction, Molecular Informatics, 30 (2011) 241-250.
[123] L.S. Villata, A.M. Berkovic, M.C. Gonzalez, D.O. Mártire, One-electron oxidation of antioxidants: A kinetic-thermodynamic correlation, Redox Report, 18 (2013) 205-209.
[124] H. Tsutsui, T. Maeda, J.Z. Min, S. Inagaki, T. Higashi, Y. Kagawa, T. Toyo'oka, Biomarker discovery in biological specimens (plasma, hair, liver and kidney) of diabetic mice based upon metabolite profiling using ultra-performance liquid chromatography with electrospray ionization time-of-flight mass spectrometry, Clin. Chim. Acta, 412 (2011) 861-872.
[125] L. Vezzaro, L. Benedetti, V. Gevaert, W. De Keyser, F. Verdonck, B. De Baets, I. Nopens, F. Cloutier, P.A. Vanrolleghem, P.S. Mikkelsen, A model library for dynamic transport and fate of micropollutants in integrated urban wastewater and stormwater systems, Environmental Modelling & Software, 53 (2014) 98-111.
[126] N.P. Moore, D.J. Andrew, D.L. Bjerke, S. Creton, D. Dreher, T. Holmes, P. Prieto, T. Seidle, T.G. Rowan, Can acute dermal systemic toxicity tests be replaced with oral tests? A comparison of route-specific systemic toxicity and hazard classifications under the Globally Harmonized System of Classification and Labelling of Chemicals (GHS), Regul. Toxicol. Pharm., 66 (2013) 30-37.
[127] S.M. Simonich, P. Sun, K. Casteel, S. Dyer, D. Wernery, K. Garber, G. Carr, T. Federle, Probabilistic analysis of risks to us drinking water intakes from 1,4-dioxane in domestic wastewater treatment plant effluents, Integrated Environmental Assessment and Management, 9 (2013) 554-559.
[128] N.S. Sipes, M.T. Martin, D.M. Reif, N.C. Kleinstreuer, R.S. Judson, A.V. Singh, K.J. Chandler, D.J. Dix, R.J. Kavlock, T.B. Knudsen, Predictive Models of Prenatal Developmental Toxicity from ToxCast High-Throughput Screening Data, Toxicol Sci, 124 (2011) 109-127.
[129] D.R. Ownby, M.C. Newman, Advances in quantitative ion character-activity relationships (QICARs): Using metal-ligand binding characteristics to predict metal toxicity, Qsar & Combinatorial Science, 22 (2003) 241-246.
[130] K. Hund, W. Traunspurger, Ecotox-evaluation strategy for soil bioremediation exemplified for a PAH-contaminated site, Chemosphere, 29 (1994) 371-390.
[131] I.V. Tetko, S. Novotarskyi, I. Sushko, V. Ivanov, A.E. Petrenko, R. Dieden, F. Lebon, B. Mathieu, Development of Dimethyl Sulfoxide Solubility Models Using 163 000 Molecules: Using a Domain Applicability Metric to Select More Reliable Predictions, Journal of Chemical Information and Modeling, 53 (2013) 1990-2000.
[132] J.K. Wegner, H. Frohlich, A. Zell, Feature selection for descriptor based classification models. 2. Human intestinal absorption (HIA), Journal of Chemical Information and Computer Sciences, 44 (2004) 931-939.
[133] D. Ballabio, A. Manganaro, V. Consonni, A. Mauri, R. Todeschini, Introduction to MOLE DB - on-line Molecular Descriptors Database, Match-Communications in Mathematical and in Computer Chemistry, 62 (2009) 199-207.

[134] A. Valencia, J. Prous, O. Mora, N. Sadrieh, L.G. Valerio, A novel QSAR model of Salmonella mutagenicity and its application in the safety assessment of drug impurities, Toxicol. Appl. Pharmacol., 273 (2013) 427-434.
[135] M. Arnaldos, K. Pagilla, Effluent dissolved organic nitrogen and dissolved phosphorus removal by enhanced coagulation and microfiltration, Water Res, 44 (2010) 5306-5315.
[136] C.S. Song, X.L. Ma, A.D. Schmitz, H.H. Schobert, Shape-selective isopropylation of naphthalene over mordenite catalysts: Computational analysis using MOPAC, Applied Catalysis a-General, 182 (1999) 175-181.
[137] P. Gramatica, S. Cassani, P.P. Roy, S. Kovarich, C.W. Yap, E. Papa, QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae, Molecular Informatics, 31 (2012) 817-835.
[138] J. Junker, Statistical filtering for NMR based structure generation, Journal of Cheminformatics, 3 (2011).
[139] K. Singh, P. Rai, P. Pandey, S. Sinha, Modeling and optimization of trihalomethanes formation potential of surface water (a drinking water source) using Box–Behnken design, Environmental Science and Pollution Research, 19 (2012) 113-127.
[140] S. Van Dorpe, V. Vergote, A. Pezeshki, C. Burvenich, K. Peremans, B. De Spiegeleer, Hydrophilic interaction LC of peptides: Columns comparison and clustering, J Sep Sci, 33 (2010) 728-739.
[141] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, Journal of Chemical Information and Computer Sciences, 43 (2003) 1882-1889.
[142] C.D. Ruark, C.E. Hack, P.J. Robinson, P.E. Anderson, J.M. Gearhart, Quantitative structure-activity relationships for organophosphates binding to acetylcholinesterase, Arch. Toxicol., 87 (2013) 281-289.
[143] A.R. Katritzky, I.B. Stoyanova-Slavova, D.A. Dobchev, M. Karelson, QSPR modeling of flash points: An update, Journal of Molecular Graphics & Modelling, 26 (2007) 529-536.
[144] E. Papa, F. Villa, P. Gramatica, Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in Pimephales promelas (fathead minnow), Journal of Chemical Information and Modeling, 45 (2005) 1256-1266.
[145] N. Flores-Holguin, A. Aguilar-Elguezabal, L.M. Rodriguez-Valdez, D. Glossman-Mitnik, Theoretical study of chemical reactivity of the main species in the alpha-pinene isomerization reaction, Journal of Molecular Structure-Theochem, 854 (2008) 81-88.
[146] G. Tugcu, M.T. Sacan, M. Vracko, M. Novic, N. Minovski, QSTR modelling of the acute toxicity of pharmaceuticals to fish, Sar and Qsar in Environmental Research, 23 (2012) 297-310.
[147] C. Hansch, A.Leo, R.W.Taft, A survey of Hammett substituent constants and resonance and filed parameters, Chem. Rev., 91 (1991) 165-195.
[148] P.D. James G. Speight, Lange's Handbook of Chemistry, Sixteenth Edition ed., The McGraw-Hill Companies, Inc., the United States of America, 2005.
[149] A.K. Banerjee, N. Harikrishna, J.V. Kumar, U.S. Murty, Towards classifying organisms based on their protein physicochemical properties using comparative intelligent techniques, Applied Artificial Intelligence, 25 (2011) 426-439.
[150] M. Graczyk, T. Lasota, B. Trawinski, Comparative Analysis of Premises Valuation Models Using KEEL, RapidMiner, and WEKA, in: N.T. Nguyen, R. Kowalczyk, S.M. Chen (Eds.) Computational Collective Intelligence: Semantic Web, Social Networks and Multiagent Systems, Springer-Verlag Berlin, Berlin, 2009, pp. 800-812.
[151] D.Z. Pavlinic, J. Gersak, J. Demsar, I. Bratko, Predicting seam appearance quality, Textile Research Journal, 76 (2006) 235-242.
[152] A. Varnek, N. Kireeva, I.V. Tetko, Baskin, II, V.P. Solov'ev, Exhaustive QSPR studies of a large diverse set of ionic liquids: How accurately can we predict melting points?, Journal of Chemical Information and Modeling, 47 (2007) 1111-1122.
[153] M.J. Vainio, M.S. Johnson, McQSAR: A multiconformational quantitative structure-activity relationship engine driven by genetic algorithms, Journal of Chemical Information and Modeling, 45 (2005) 1953-1961.
[154] L. Villanueva-Alcojol, F. Monje, R. Gonzalez-Garcia, Hyperplasia of the Mandibular Condyle: Clinical, Histopathologic, and Treatment Considerations in a Series of 36 Patients, Journal of Oral and Maxillofacial Surgery, 69 (2011) 447-455.
[155] A.R. Katritzky, M. Kuanar, S. Slavov, D.A. Dobchev, D.C. Fara, M. Karelson, W.E. Acree, V.P. Solov'ev, A. Varnek, Correlation of blood-brain penetration using structural descriptors, Biorg. Med. Chem., 14 (2006) 4888-4917.
[156] J.F. Contrera, Improved in silico prediction of carcinogenic potency (TD50) and the risk specific dose (RSD) adjusted Threshold of Toxicological Concern (TTC) for genotoxic chemicals and pharmaceutical impurities, Regul. Toxicol. Pharm., 59 (2011) 133-141.

[157] V.K. Vyas, M. Ghate, CoMFA and CoMSIA Studies on Aryl Carboxylic Acid Amide Derivatives as Dihydroorotate Dehydrogenase (DHODH) Inhibitors, Current Computer-Aided Drug Design, 8 (2012) 271-282.
[158] J.A. Lazzus, Neural Network Based on Quantum Chemistry for Predicting Melting Point of Organic Compounds, Chinese Journal of Chemical Physics, 22 (2009) 19-26.
[159] A. Gurses, M. Yalcin, C. Dogar, Electrocoagulation of some reactive dyes: a statistical investigation of some electrochemical variables, Waste Manage. (Oxford), 22 (2002) 491-499.
[160] D. Ajloo, A.A. Saboury, N. Haghi-Asli, G. Ataei-Jafarai, A.A. Moosavi-Movahedi, M. Ahmadi, K. Mahnam, S. Namaki, Kinetic, thermodynamic and statistical studies on the inhibition of adenosine deaminase by aspirin and diclofenac, Journal of Enzyme Inhibition and Medicinal Chemistry, 22 (2007) 395-406.
[161] M. Ghadamgahi, D. Ajloo, Calculation and prediction of rate and equilibrium constants for aggregation of porphyrin by molecular dynamics, Docking and QSPR methods, J. Porphyrins Phthalocyanines, 15 (2011) 240-256.
[162] G. Tavernier, G. Fletcher, I. Gee, A. Watson, G. Blacklock, H. Francis, A. Fletcher, T. Frank, P. Frank, C.A. Pickering, R. Niven, IPEADAM study: Indoor endotoxin exposure, family status, and some housing characteristics in English children, Journal of Allergy and Clinical Immunology, 117 (2006) 656-662.
[163] Y. Tominaga, Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN, Chemometrics and Intelligent Laboratory Systems, 49 (1999) 105-115.
[164] A. Khare, S. Trivedi, H. Rajak, R.S. Pawar, U.K. Patil, P.K. Singour, Hansch analysis of novel pyrimidine derivatives as highly potent and specific COX-2 inhibitors, Med. Chem. Res., 21 (2012) 672-680.
[165] L. Eriksson, et al., Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and regression-based QSARs, Environ. Health Perspect., 111 (2003).
[166] A. Golbraikh, A. Tropsha, Beware of q2!, J. Mol. Graphics Modell., 20 (2002) 269-276.