Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201300180

Global Performance and Trend of QSAR/QSPR Research: A Bibliometric Analysis

Li Li,[a] Jianxin Hu,[a] and Yuh-Shan Ho*[b, c]

Abstract: A bibliometric analysis based on the *Science Citation Index Expanded* was conducted to provide insights into the publication performance and research trend of quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) from 1993 to 2012. The results show that the number of articles per year quadrupled from 1993 to 2006 and plateaued since 2007. *Journal of Chemical Information and Modeling* was the most prolific journal. The internal methodological innovations in acquiring molecular descriptors and modeling stimulated the articles' increase in the research fields of drug design and synthesis, and chemoinformatics; while the external regulatory demands on model validation and reliability fueled the increase in environmental sciences. "Prediction endpoints", "statistical algorithms", and "molecular descriptors" were identified as three research hotspots. The articles from developed countries were larger in number and more influential in citation, whereas those from developing countries were higher in output growth rates.

Keywords: Quantitative structure-activity relationship (QSAR) · Quantitative structure-property relationship (QSPR) · Research trend · SCI-EXPANDED · Scientometrics

1 Introduction

During the last decades, quantitative structure-activity relationship (QSAR) and quantitative structure-property relationship (QSPR) have been explored and exploited in a wide range of scientific fields, such as structure-based drug and pesticide design,[1] environmental toxicology,[2] molecular biology,[3] along with the control and management of industrial chemicals.[4] Based on a set of analog chemicals with determined activities or properties, QSAR/QSPR relates a certain endpoint such as pharmacological activity, physicochemical property, biological toxicity and environmental parameter, with the molecular structure features that are mimed by categorical or numerical descriptors and, thus, facilitates the prediction of the untested chemicals' endpoints fed with the molecular structure information alone. In recent years, QSAR/QSPR has gained particular favor as a cost-effective alternative to avoid substantial laboratory experiments and in vivo animal test, especially under REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) of the European Union.[1,5]

The earliest QSAR/QSPR research dates back to the 19th century when Brown-Crum and Fraser first hypothesized a function linking physiological action and chemical constitution in a pharmacological study;[6] the clear point of its commencement is still under debate.[7] With the development of physical organic chemistry, in the 1930s, Hammett proposed the linear free energy relationships (LFER) theory, in which the rate constants and the equilibrium constants of a set of substituted reactants were associated by introducing the electronic Hammett constant of specific substituents.[8,9] For the first time, the LFER theory projected the macroscopic thermodynamic similarity in a quantitative manner on the microcosmic substituent similarity.[10,11] Following the substituent-based perspective, a series of new descriptors like biological Hammett constant,[12] steric substituent constant,[13] and hydrophobic constant[14] were produced in succession, and novel theoretical models like the linear solvation energy relationships (LSER)[15] and the theoretical linear solvation energy relationships (TLSER)[16] were raised in subsequent researches.

The contemporary QSAR/QSPR technique started off in late 1960s due to the emergence of two most common modeling approaches, the Hansch–Fujita analysis and Free–Wilson analysis.[7] As a whole-molecule approach, the Hansch–Fujita analysis attributed the substance endpoints, by first introducing the multivariate equations, to a linear

[a] *L. Li, J. Hu*
State Key Joint Laboratory for Environmental Simulation and Pollution Control, College of Environmental Sciences and Engineering, Peking University
Beijing 100871, People's Republic of China

[b] *Y.-S. Ho*
Trend Research Centre, Asia University
Taichung 41354, Taiwan
**e-mail: ysho@asia.edu.tw*

[c] *Y.-S. Ho*
Department of Environmental Engineering, Peking University
Beijing 100871, People's Republic of China
tel: +886 4 2332 3456 x 1797; fax: +886 4 2330 5834.

Supporting Information for this article is available on the WWW under www.molinf.com.

combination of physicochemical parameters describing hydrophobic, steric, and electrostatic properties.[17–19] In contrast, as a substructure approach, the Free–Wilson analysis attributed the substance endpoints, by first introducing the concept of additivity and symmetry, to the sum of contributions from molecular fragments such as a single atom, functional group and substituent.[20] Instead of just using experimentally determined or theoretically calculated physicochemical parameters in QSAR/QSPR modeling before, the Free-Wilson approach involved the molecular structure descriptors, which inspired a bewildering variety of theoretical methods that capture molecular structure features and interpret them as numeric descriptors. In traditional 2D-QSAR/QSPR domain, typical ones such as topological,[21] electrotopological,[22,23] geometrical,[24] and quantum-chemical[25] descriptors were brought forth in succession. When goes to hyperspace, multidimensional descriptors were generated to meet the needs of fingerprinting and delineating sophisticated pharmacophore features in interactions between receptor and drug/contaminants. Such modeling approaches include but are not limited to the comparative molecular field analysis (CoMFA),[26] the comparative molecular similarity indices analysis (CoMSIA),[27] and the grid-independent descriptors (GRIND).[28]

With massive molecular descriptors involved, traditional multiple linear regression has exposed its deficiencies in dealing with numerous collinear independent variables in small sample size situation;[29] therefore, proposed in the following decades were more reliable and robust algorithms, among which the partial least-squares (PLS) regression is the most applied.[29–31] Furthermore, given that QSAR/QSPR is more often used for data mining in legislation and regulatory practices, the standardized and verifiable procedures for selecting descriptors and generating models became pressing issues to both the academia and policy makers. These demands led to the proposal of *OECD principles for the validation of (Q)SAR models* by Organization for Economic Co-operation and Development (OECD) in 2004, in which five indispensable components were called for, including a defined endpoint, an unambiguous algorithm, a limited applicability domain, appropriate assessment for internal performance and external predictivity, and possible mechanistic interpretation.[32] Unlike the previous situation that most of efforts had been made on acquiring molecular descriptors and modeling, the main focus has shifted to assessment of effectiveness, reliability and predictivity of QSAR/QSPR models in the recent decade. Several crucial efforts were made to address these problems, such as proposing new evaluation methodologies to overcome the shortcomings of traditional validation technology,[33–35] and measuring and scaling chemical space that QSAR/QSPR can yield reliable results.[36]

Although expanding efforts have been devoted to QSAR/QSPR research, few attempts have been made on a systematic evaluation of its global performance and trend, in particular a bibliometric analysis. Based on the document records from Science Citation Index Expanded (SCI-EXPANDED), the bibliometric analysis quantifies publication performance, including the distribution pattern of subjects, journals and keywords;[37–39] as well as citation behaviors such as inter-article and inter-annual variations in number of citations and co-citations,[40–43] providing new insights on the global publication productivity, future orientation, research frontier and hotspots.[44] In the previous studies, Willett and his co-workers provided a preliminary glimpse of the global QSAR/QSPR research trend based on discrete bibliometric analysis on individual journals, like *Quantitative Structure-Activity Relationships*,[45] *Journal of Chemical Information and Computer Sciences*,[46] and *Journal of Computer-Aided Molecular Design*.[47] However, in order to depict a more comprehensive QSAR/QSPR research image among disciplines of the whole scientific world, it is required to look at broader journal sources and analyze deeper content containing more detail information such as author keywords, words in title, and *KeyWords Plus*.[38,42,48]

In this study, a bibliometric analysis based on SCI-EXPANDED database was carried out to evaluate the global QSAR/QSPR publication performance and research trend from 1993 to 2012. The documents were analyzed and summarized to determine the quantitative characteristics like annual outputs, source journals, categories, research fields, and geographic distribution. Moreover, research hotspots were mapped to gain insights into the structure of QSAR/QSPR research, by clustering phrases in the article front page[49] and ranking the high-impact articles.

2 Methodology

The analyzed data of documents published from 1993 to 2012 were retrieved from the Science Citation Index Expanded (SCI-EXPANDED) database of Web of Science (WoS) from Thomson Reuters (updated on June 7th, 2014). Explicit QSAR/QSPR-relevant phrases, including "quantitative structure property relationship", "quantitative structure activity relationship", QSAR, QSARs, QSPR, and QSPRs; and the typical phrases relevant to the emerging 3D-QSAR/QSPR techniques, including CoMFA, CoMSIA, "comparative molecular similarity indices analysis" and "comparative molecular field analysis" were searched in terms of topic. The retrieved "CoMFA"-containing documents were checked to exclude those that are irrelevant to QSAR/QSPR topic despite being in accordance with the retrieval criteria, because "ComFA" and "COMFA" also often refer to a protein name in molecular biology or the thermal comfort in biometeorology, respectively. Other possible alternative terms, such as "chemometrics", "structure descriptor" and "molecular descriptor" were not involved because over half of the retrieved documents were not relevant to QSAR/QSPR topic according to the pretest results; it is unable to exactly differentiate from the massive retrieved results, although many of them admittedly dealt with QSAR/QSPR problems.

The documents with the retrieving words in title, abstract, author keywords, and *KeyWords Plus* were extracted. Here, *KeyWords Plus* is an additional index term in Web of Science. It is derived from the words in title or author keywords but describes the article's contents with greater depth and variety than title and author keywords.[50] However, introducing *KeyWords Plus* often artificially augments the content range of the retrieved documents;[49] therefore, the final filter in this study was set to be "article front page", which means only the articles with the search keywords in the article title, abstract, and author keywords were singled out.[49] The procedural steps for data retrieval and processing are detailed in Figure S1 of Supporting Information.

The trend analysis was performed using Microsoft Excel 2007 to trace the time-wise article outputs, authorships, citations and word distribution. The following points are beforehand elaborated to avoid possible misunderstanding in the subsequent analysis: (1) Articles originating from England, Scotland, Northern Ireland, and Wales were reclassified as being from the United Kingdom (UK);[51] articles from Yugoslavia and Serbia were reclassified as being from Serbia; articles from Hong Kong and Macao were incorporated into China's category;[39] (2) Collaboration type was determined by the address of authors. The term "single country article" was assigned if the all researchers' affiliations were from the same country/territory, and the term "internationally collaborative article" was designated to those articles with at least one co-author from different country/territory;[52] (3) For journal with one or more former names, or journal incorporating other previous journals, only the present journal name and impact factor (for the year 2012) were presented in this study, and article outputs were counted under the name of present journal.

The document co-citation analysis (DCA) was performed on the software *Citespace* ver. 3.6 to find out the most influential intellectual propellants on QSAR/QSPR research during the entire period. Co-citation means that a pair of documents (in any document types not necessarily published within the period 1993–2012) are in common cited by the same article retrieved in this study. The co-citation theory believes that a cluster of co-cited documents constitutes the common intellectual base of a single citing article, thus, the evolution of the research focuses (or referred as research fronts), which consist of the newly surged articles in individual sub-periods, can be monitored by the transient but successive emergence of the most prominent cluster of co-cited documents. In this sense, research focus marks the influential breakthroughs in each stage in the history of QSAR/QSPR development. By tracing the appearance order and content of the dominant co-cited documents, it is possible to find out the time that a research hotspot flourished and the landmark articles triggered the flourish.[41,53,54] The procedural steps of *Citespace* are sketched in Figure S2 (Supporting Information). Briefly, all cited documents were grouped into ten sliced 2-year segments from 1993 to 2012 to compose the co-citation spaces, according to the year first cited (time slicing). Some of the top-cited documents that meet the following selection thresholds[53] were picked out to construct individual co-citation networks for respective time slices (thresholding): minimum citation counts (*c*), minimum co-citation strength in each time slice (*cc*), and normalized co-citation cosine coefficient (*ccv*) were assigned (5, 3, 0.15), (10, 4, 0.19), and (12, 5, 0.20) for the earliest, middle and last slice, and the linear interpolated thresholds were assigned for the remaining slices. Co-citation network is represented as the node-and-link clusters, with the tree-ring nodes denoting individual co-cited documents and closely spaced links denoting the co-citations. Given the clutter and overlap in the preliminary co-citation networks, redundant link-crossings needed to be removed using minimum spanning tree algorithms (pruning) and all networks were integrated as one for a panoramic overview (merging), with the purpose of creating more clear network visualization.[53] A final co-citation network with 277 nodes and 1285 lines was obtained, and subjected to interpretation. The parameterization of time slicing, threshold settings and modeled results are tabulated in Table S1 (Supporting Information).

3 Performance of Publication

3.1 Publication Overview

To obtain an overview of QSAR/QSPR research performance, 12767 retrieved documents, with the retrieving words only explicit in their front pages, were subjected to analysis of document type and language. Fourteen document types identified by WoS were found. Original articles (11020; including articles and proceedings papers pursuant to WoS's definition) composed the most frequently used document type accounting 86% for all publications, followed by meeting abstracts (806; 6.3%), reviews (761; 6.0%), proceedings papers (692; 5.4%), and the remainder having less proportion, were editorial materials (62), corrections (51), letters (19), notes (18), book chapters (13), addition corrections (7), news items (4), biographical-items (3), bibliographies (2), and one for discussion. Noticing that the original article is the major peer-reviewed document type directly proposing novel concepts and presenting substantive findings, 11 020 articles were subjected to further analysis, whereas all others were discarded despite their equal value in comparison with articles. Ninety-six percent of all the articles were published in English. Fourteen other languages were also used, including Chinese (317), Portuguese (16), Spanish (12), Russian (11), Rumanian (8), French (6), Japanese (6), German (5), Czech (3), Korean (3), and one in Turkish, Hungarian, Serbian, and Afrikaans respectively.

The evolution of the number of articles per year and the proportion in all SCI-EXPANDED journals' articles are illustrated in Figure 1. The quantitative characteristics of authorship, cited reference, and document page are summar-

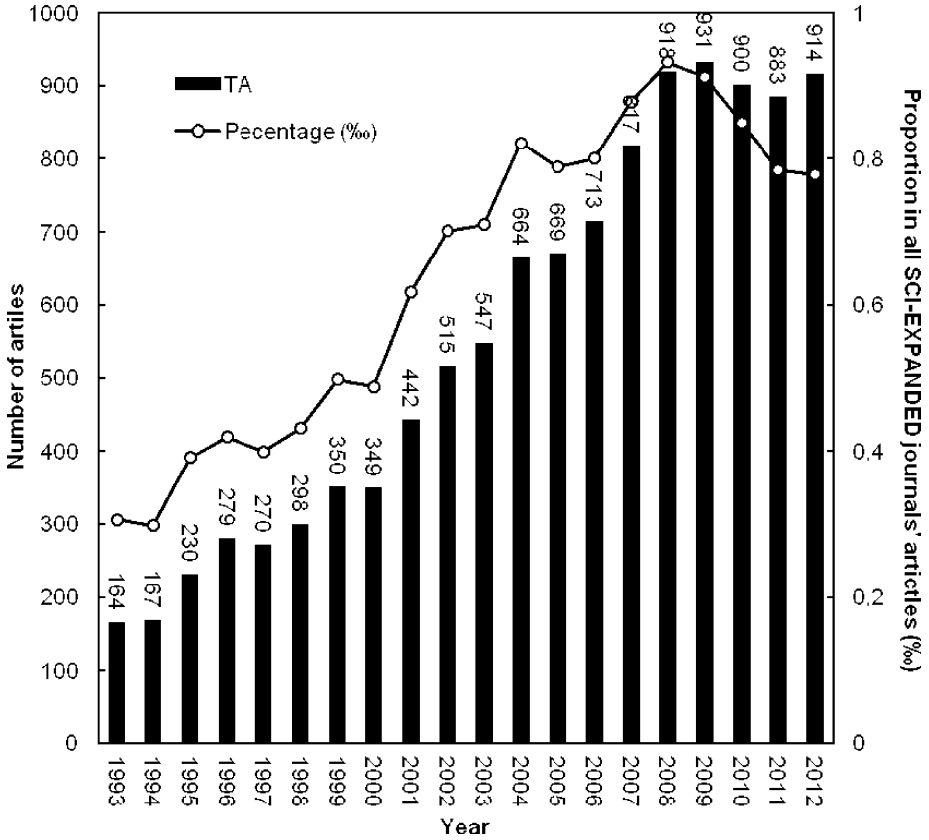

Figure 1. The number of articles per year on QSAR/QSPR research and the proportion in all SCI-EXPANDED journals' articles (‰) from 1993 to 2012.

ized in Table S2 (Supporting Information). Briefly, the number of articles per year quadrupled linearly from 1993 (164) to 2006 (713), then plateaued around 900 after 2007. The article productivity on QSAR/QSPR research boomed faster than the average level of the entire science and technology, as demonstrated by the continuous increasing trend of its share in all SCI-EXPANDED journals' articles to the peak of 0.93‰ in 2008. The average number of co-authors per article slightly rose from 1993 (4.0) to 2004 (4.2), and remained asymptotic to 4.4 since then. The plateaued multiple authorship after the first rise is also reported in a previous bibliometric analysis of experimental biology.[55] The average cited references per article almost doubled from 1993 (25) to 2012 (43) owing to more convenient access to increasing larger database. The document size changed a little from 9.6 in 1993 to 11.4 in 2012.

The number of articles in peer-reviewed SCI-EXPANDED journals and symposium series books was counted to assess the performance of publication sources. All 11 020 articles resided in 913 journals or symposium series, and 9 core journals (>200 articles) published 3848 (35%) articles (Table 1). *Journal of Chemical Information and Modeling* (previously entitled *Journal of Chemical Information and Computer Sciences* before 2005) was the most prolific journal with the total articles of 718 (6.5%), followed by *Bioorganic & Medicinal Chemistry* (578; 5.2%) and *Molecular Informatics* (previously entitled *QSAR & Combinatorial Science* before 2003, and *Quantitative Structure-Activity Relationship* for 2003–2009) (564; 5.1%). It is also noticed that there were 401 (44% of total 913) journals with only one QSAR/QSPR-related article and even 756 (83%) journals with no more than ten articles. The lack of a single preponderant journals covering the lion's share of articles suggests the low publication centralization in QSAR/QSPR research and the wide range of interests from multidisciplinary angle.[42] The temporal evolution of article outputs from the top nine journals is illustrated in Figure S3 (Supporting Information). In general, almost all selected journals increased in article outputs before 2007. However, the increasing trend continued for the journals on environmental sciences (*SAR and QSAR in Environmental Research* and *Chemosphere*); whereas the fairly stable trends on molecular biology (*Bioorganic & Medicinal Chemistry* and *Journal of Computer-Aided Molecular Design*) and the descending trends on medicinal chemistry (e.g. *Journal of Medicinal Chemistry* and *European Journal of Medicinal Chemistry*). Moreover, some specialist journals, such as *Molecular Informatics* (564; 5.1%) and *SAR and QSAR in Environmental Research* (324; 2.9%), paid particular attention to QSAR/QSPR application in spe-

Table 1. The ten most prolific journals with the number of articles, impact factor, and Web of Science category. *TA*: total articles; *IF2012*: impact factor in *Journal Citation Reports* 2012 published by Thomson Reuters.

Journal	*TA*(%)	*IF2012*	Web of Science category
Journal of Chemical Information and Modeling [a]	718 (6.5)	4.304	multidisciplinary chemistry information systems computer science interdisciplinary applications computer science
Bioorganic & Medicinal Chemistry	578 (5.2)	2.903	biochemistry & molecular biology medicinal chemistry organic chemistry
Molecular Informatics [b]	564 (5.1)	2.338	medicinal chemistry mathematical & computational biology
Journal of Medicinal Chemistry	506 (4.6)	5.614	medicinal chemistry
European Journal of Medicinal Chemistry	395 (3.6)	3.499	medicinal chemistry
SAR and QSAR In Environmental Research	324 (2.9)	1.667	multidisciplinary chemistry interdisciplinary applications computer science environmental sciences mathematical & computational biology toxicology
Journal of Computer-Aided Molecular Design	295 (2.7)	3.172	biochemistry & molecular biology biophysics interdisciplinary applications computer science
Bioorganic & Medicinal Chemistry Letters	257 (2.3)	2.338	medicinal chemistry organic chemistry
Chemosphere	211 (1.9)	3.137	environmental sciences

[a] previously entitled Journal of Chemical Information and Computer Sciences before 2005; [b] previously entitled QSAR & Combinatorial Science before 2002 and Quantitative Structure-Activity Relationship for 2003–2009

cific areas, indicating the professionalization of QSAR/QSPR research.

3.2. Publication Patterns: Subject Categories and Research Fields

To demonstrate the discipline distribution of QSAR/QSPR research, the WoS categories, at its 2012 version where all journals are categorized into 251 disciplines, were used to assign the individual articles. All 11 020 articles originated from 147 WoS categories. The seven most prolific categories produced more than 1000 articles each (upper panel of Figure 2). Multidisciplinary chemistry contributed the most (3388; 31%), followed by medicinal chemistry (3308; 30%) and interdisciplinary computer science (2190; 20%). Meanwhile, 83 (56% of total 147) categories produced no more than ten articles each, and 25 (17%) categories produced only one article. Diversity of ten least ones (lower panel of Figure 2) suggests the breadth of QSAR/QSPR application. To facilitate further interpretation, the several most prolific WoS categories, which covered 10 018 articles belonging to 612 journals, were grouped into three "research fields" according to their common research goal and similar research methodology: drug design and synthesis (7620 articles, including multidisciplinary chemistry, medicinal chemistry, biochemistry and molecular biology, and pharmacology and pharmacy), chemoinformatics (4649 articles, including interdisciplinary applications of computer science, information systems of computer science, mathematical and computational biology, and organic/physical/analytical chemistry), and environmental sciences (1447 articles, including environmental sciences and toxicology). Here exists an overlap that 3050 and 324 articles (published in 47 and one journals, respectively) belong to two and three research fields, as 361 of all 612 selected journals reside in more than one WoS categories. Figure 3 shows the trends of article outputs of the three research fields. The field of drug design and synthesis (*TA* = 7620) was more fruitful than the other two (*TA* = 4649 and 1447). Both fields of drug design and synthesis and chemoinformatics have accelerated their increase in article output since 1997, but obviously dropped from 2008 onwards (downhill fields); whereas environmental sciences kept steady increase at relatively lower speed throughout the period, and remains continuous promising trend after 2008 (uphill field). Possible reasons will be given later.

The co-citations analysis were conducted and visualized by *Citespace*, allowing to understand the emerging evolutional work that serve as the common intellectual base for the retrieved articles between 1993 and 2012. A total of 277 highly co-cited documents by our retrieved articles were selected with reference to chosen thresholds (Figure S2, Supporting Information) and represented as discrete, chronologically-located citation tree-ring-like nodes (Figure 4), among which are the nine most prominent ones published after 1985 representing landmark works. Solid dots mark the node centers; references are given aside. Meanwhile, information for the top 11 most co-cited articles is given in Table S3 (Supporting Information). In Figure 4, the size of node is proportional to the co-citation

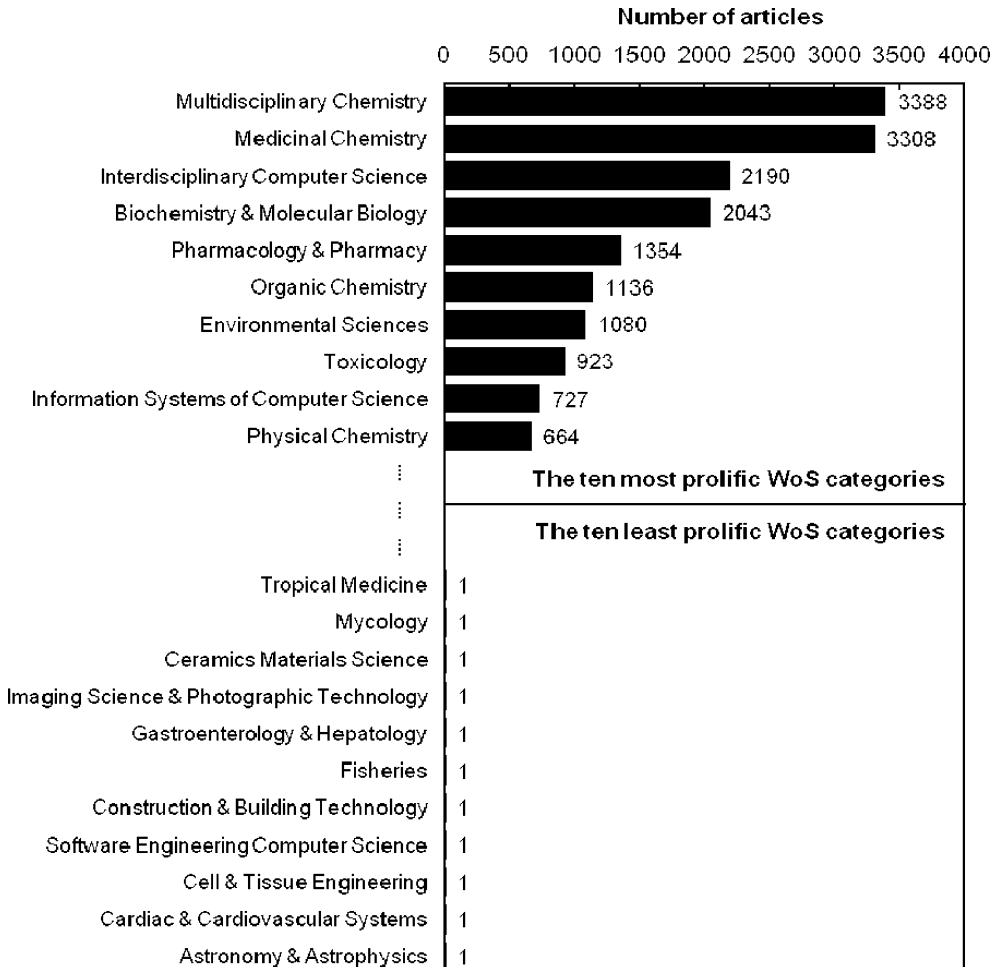

Figure 2. The number of articles of the ten most and least prolific Web of Science categories from 1993 to 2012.

counts throughout the entire time interval; the larger a co-citation tree-ring is, the more co-citation and higher influence a document has. The thickness of heterochromatic concentric rings in a node represents the relative amount of co-citations in a given time slice, and the color of concentric rings denotes time slices that co-citations reside in. The color of the connecting line between a pair of tree-rings denotes the year of the first co-citation for the chosen thresholds.[41,53] Two interesting conclusions are drawn from the co-citation analysis (Figure 4 and Table S3 (Supporting Information)): First, the contemporary QSAR/QSPR research during the 1993–2012 period was more influenced by the latest scientific findings appeared after the mid-1980s. Second, the co-cited documents are still active in influencing the most recent QSAR/QSPR research, as almost all tree-rings have salmon and bright red edges and represent the co-citations later than 2009.

It is interesting to see that all the five most influential works before 2002 deal with molecular descriptors and statistical algorithms (modeling-related articles), such as AM1,[56] CoMFA,[26] CoMSIA,[27] genetic function approximation,[29] and PLS regression;[30] while all the four after 2001 focus on model validation and reliablity assessment (validation-related articles), such as discussions on the internal and external validation.[35] More interestingly, the modeling-related articles got their first co-citation almost between 1997 and 1999, and their highest co-citations before 2007; whereas validation-related articles got their first co-citation almost between 2007 and 2009, and their highest co-citations after 2009, as indicated by the respective prevalent green and salmon connecting lines and concentric rings in Figure 4. This implies the academia has shifted its research emphasis from molecular descriptors acquisition and modeling methodologies to model optimization and validation. Furthermore, this reveals the different impetuses of "downhill fields" and "uphill field" mentioned thereinbefore, which is explainable for their different publication performances. For research fields of drug design and synthesis and chemoinformatics, the research evolution depends more on the internal impetus, i.e., the scientific methodological innovations in descriptor acquisition and model generation, because the time before 2009 that their article outputs increased (Figure 3) is exactly when the five modeling-related articles got their highest co-citations (Figure 4) and, with

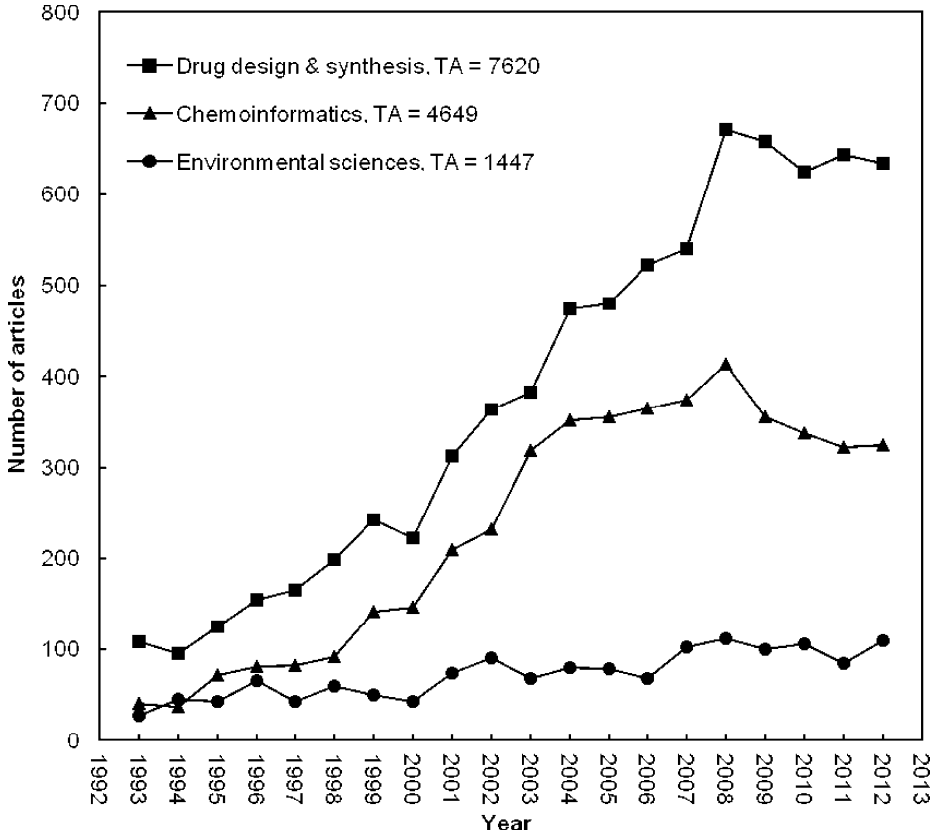

Figure 3. The annual article outputs in the top three research fields from 1993 to 2012. (*TA*: total articles)

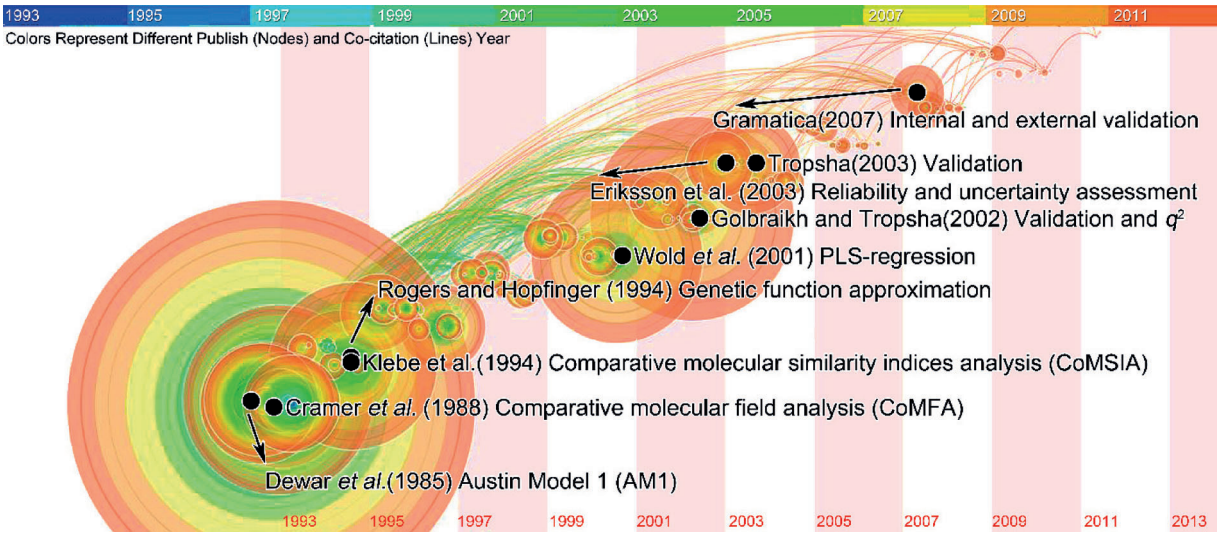

Figure 4. Citation tree-rings of co-cited documents by the articles during 1993–2012.

less influential methodological innovations in the most recent decade, the internal impetus seems not as strong as before (Figure 3). For example, as shown by the prevalent light blue and green concentric rings, around 70% of co-citations of Cramer et al.,[26] the most prominent tree-ring in the bottom left of Figure 4, occurred between 1997 and 2005. During this period surged a series of influenced breakthroughs in structure-based drug design, such as synthesizing antitumor[57] and anti-HIV drugs,[58] concatenating new field and descriptors into conventional 3D-QSAR,[59,60] developing appropriate variable selection approach,[61,62] and exploring more effective and advanced statistical

methods or algorithms.[63,64] In contrast, for research field of environmental sciences, its research evolution relies more on the external impetus, i.e., the growingly stricter regulatory requirement on model validation and reliability in environmental sound management of chemicals, as it continued increasing after 2007 when a series of validation-related articles got the most cited. In environmental sciences area, QSAR/QSPR was used by the USEPA very early for filling data gaps in new and existing chemical evaluations.[65] Some integrated software, such as the Estimation Program Interface Suite by the USEPA and QSAR Toolbox by OECD, were also released for the better access to QSAR calculations. QSAR/QSPR has not received great promoted until 2003, when it was recommended an alternative methods to conventional laboratory in vivo animal testing under the legal framework of REACH of the European Union,[5] a new regulation aiming for systematic assessment of both human health and environmental risks of over 30 000 chemical substances available in European market.[66] For legislative purposes, a QSAR/QSPR model is required to be transparent, unambiguous and explainable; therefore, many efforts were made to investigate model validation,[35,67] which led to the articles' increase in environmental sciences.

3.3 Research Tendencies and Hotspots

In terms of the content analysis, research tendencies and hotspots are reflected by the frequently emerging key phrases obtained by "word cluster analysis".[38] Meanwhile, in terms of the citation analysis, they are reflected by the research emphasis in the highly-cited articles that impose profound effects on sequent researches.[68]

The flow diagram of the word cluster analysis[38] is shown in Figure S4 (Supporting Information). Briefly, author keywords and *KeyWords Plus*, along with the substantives in title, were extracted from the 11 020 retrieved articles and complied as the "word sources", which contain almost all QSAR/QSPR-related information that authors of the retrieved articles meant to express. Next, "supporting words", some most-appeared synonymic or near-synonymic words or phrases in the word sources, were picked up and grouped as the "word clusters" by authors' specialized knowledge. The research hotspots are generalized from the articles with these supporting words in their front pages. The change in the article count of a certain cluster represents the flourish or decline of a research hotspot.[38] In this study, three domains were classified with their trends illustrated in Figure 5.

Cluster one pertains to "prediction endpoints" (4962 articles), composed of the supporting words including toxicity, lipophilicity, hydrophobicity, pharmacophore, cytotoxicity, and inhibitor(s). The term of endpoint refers to a particular response to the substance's molecular structure such as pharmacological activity, physicochemical property, biological toxicity and environmental nature. Cluster one took the leading role in article count over the two decades, and the case studies merely using existing computational methods or conventional QSAR/QSPR models composed the most of them. QSAR/QSPR is the most used in predicting the pharmacological effects. For example, there were 1127 of all 4962 articles dealing with the pharmacological endpoints, denoted by the prefix or suffix of "pharm" and "inhibitor" in author keywords; but 711 and 376 articles dealing with the biological and physicochemical endpoints, denoted by the prefix or suffix of "toxic", as well as "phobic", "philic", and "partition" in their author keywords. In terms of the articles related to the pharmacological endpoints (1127 articles), "pharmacophore" (139; 12%) was the most frequently used author keywords. Here, the pharmacophore means the specific structural features onto molecule exhibiting the pharmacological activity with given 3D structure and positions of functional groups and atoms.[69] The dominance of the "pharmacophore" concept suggests it has infiltrated into the most mainstream domains such as virtual screening, de novo molecular design and multi-target drug design.[69] Meanwhile, in terms of the articles related to the biological endpoints (711 articles), more frequent than others appear the toxicological keywords such as "cytotoxicity" (88; 12%) and "acute toxicity" (52; 7.3%), because QSAR/QSPR was favor of data mining of scientific or regulatory application, especially in the decision support system of chemical management.[70] Meanwhile, a few studies focused on data selection, evaluation and refinement pursuant to OECD QSPR principles,[70] in order to avoid ambiguous even conflicting QSAR/QSPR models resulted from the unclearly "defined" endpoints.

Cluster two deals with the "statistical algorithms" adopted in QSAR/QSPR modelling (3855 articles), including the supporting words of genetic algorithm(s), genetic function approximation, regression analysis, multiple linear regression, MLR, partial least square(s), PLS, principal component analysis, (artificial) neural network(s), support vector machine, SVM, classification, and heuristic method. The term of algorithm refers to the mathematical representations linking endpoints and molecular descriptors.[71] Generally, classic, mature and reliable algorithms gained more popularity than novel ones; one possible reason is that the most articles were devoted to the mere QSAR/QSPR application in case studies for data acquisition purpose. For instance, the retrieving words of "neural network" and "ANN" appeared the most frequently in articles' author keywords (603; 16%), followed by "partial least square" and "PLS" (557; 14%), and "multiple linear regression" and "MLR" (348; 9.0%). By contrast, some novel machine-learning algorithms were applied in just a few cases: thirteen articles on gene expression programming,[72] seven on local lazy regression,[73] three on iterative double least square,[74] and one on project pursuit regression.[75]

Cluster three is a set of molecular descriptors and the corresponding modeling approaches (3802 articles), including the supporting words of molecular descriptor(s), com-

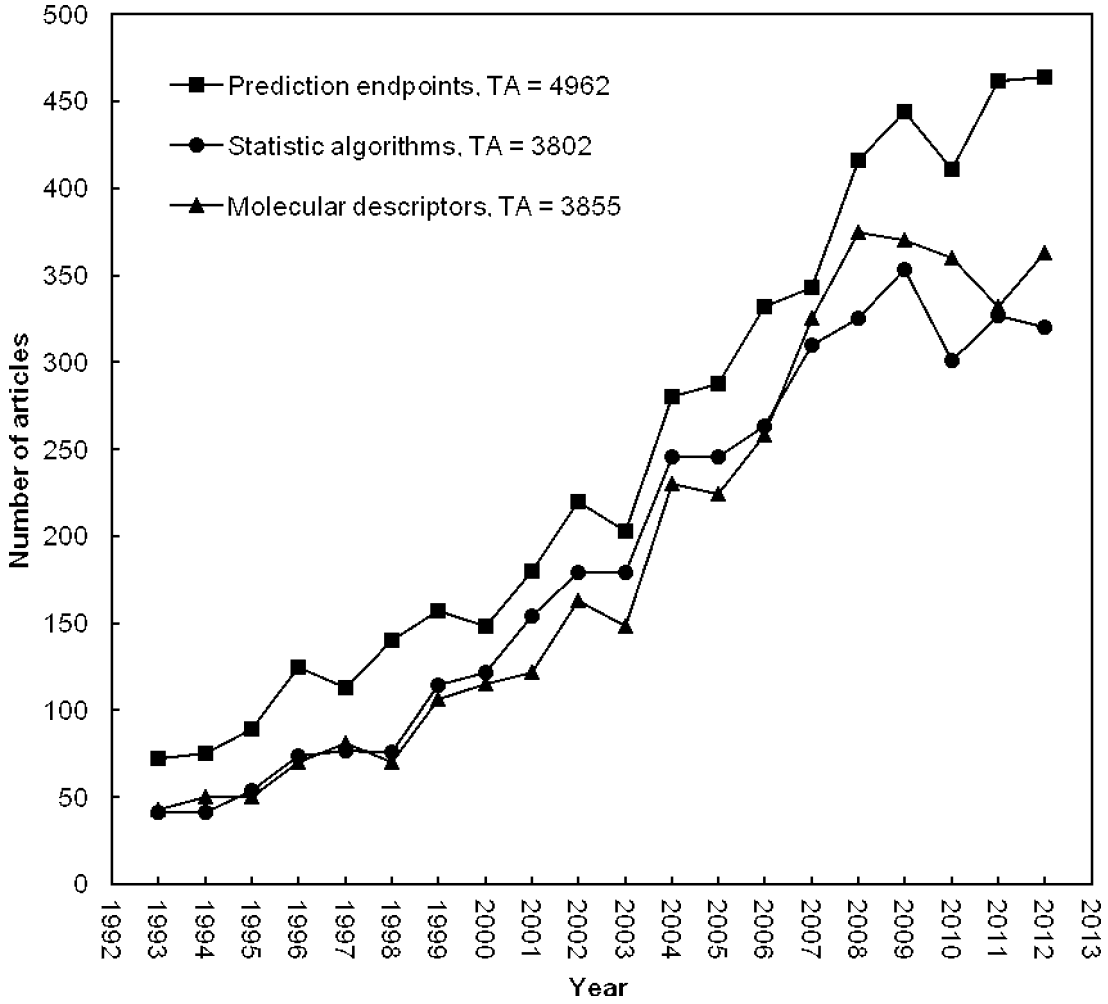

Figure 5. Comparison of article output trends in the three hot topics from 1993 to 2012. (*TA*: total articles)

parative molecular field analysis, CoMFA, comparative molecular similarity indices analysis, CoMSIA, density functional theory, DFT, quantum chemical descriptor(s), and topological index (indices). Cluster three represents the selection and employment of appropriate molecular descriptors to depict molecular features and natures. 3D-QSAR/QSPR technique was the most used descriptors acquisition technology, as over half of articles (1967; 52%) contained the representative words of "3D", "CoMFA", "comparative molecular field analysis", "CoMSIA", or "comparative molecular similarity indices analysis" in their front pages. Of the 1967 3D-QSAR related articles, 748 (38%) belonged to the WoS category of medicinal chemistry, suggesting 3D-QSAR/QSPR was primarily adopted in drug design. This is because traditional 2D-QSAR/QSPR technology failed to describe the ligand-receptor interactions in the pharmacophore of drug molecule.[76] However, 549 of the remaining articles belonged to multidisciplinary chemistry, indicating the interdisciplinary background of 2D-QSAR/QSPR. For instance, there were a total of 351 articles in environment sciences, but 266 (76%) belongs to non-3D-QSAR/QSPR topics. Furthermore, it is interesting found that the major methodological innovations took place in the first decade (1993–2002) or earlier, such as the proposal of CoMSIA[27] and electrotopological state index approach;[22,77] whereas the major successful modifications and applications were achieved in the second decade (2003–2012), such as the applications of CoMFA and CoMSIA in quantifying the enzyme inhibition activities[78–80] and toxicological properties[81,82] of chemicals.

In citation analysis, total citation count from the initial publication of the article up to the end of 2012 (*TC2012*) is calculated for each document.[43,68] The total citation count is believed to be a valuable indicator to evaluate the quality and popularity of scientific work[83] as cutting-edge and high-quality articles often attract more citations from the scientific community;[84] therefore, the most frequently cited articles wield the higher influence (high-impact articles), and the topics involved are identified as the hot issues.[43,68,85] In this study, thirteen high-impact articles were extracted by a filter of $TC2012 \geq 250$ to analyze the research emphasis (Table 2). Here, the paper entitled "The importance of being earnest: Validation is the absolute essen-

Table 2. The high-impact articles with *TC2012* ≥ 250.

Title	Authors (year)	*TC2012*	Author keywords	Country of first author/corresponding author
PLS-regression: a basic tool of chemometrics	Wold et al. (2001)[30]	1283	PLS; PLSR; two-block predictive PLS; latent variables; multivariate analysis	Sweden/Sweden
Beware of q^2!	Golbraikh and Tropsha (2002)[33]	1088	QSAR modeling; LOO cross-validation; training and test sets; kNN QSAR	USA/USA
Molecular similarity indexes in a comparative-analysis (CoMSIA) of drug molecules to correlate and predict their biological-activity	Klebe et al. (1994)[85]	954	N/A	Germany/Germany
Application of genetic function approximation to quantitative structure-activity-relationships and quantitative structure-property relationships	Rogers and Hopfinger (1994)[29]	646	N/A	USA/USA
Generating optimal linear PLS estimations (GOLPE)- an advanced chemometric tool for handling 3D-QSAR problems	Baroni et al. (1993)[86]	331	GOLPE; PLS; variable selection; COMFA; 3D-QSAR	Italy/Italy
Neighborhood behavior: A useful concept for validation of "molecular diversity" descriptors	Patterson et al. (1996)[87]	309	N/A	USA/USA
Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*)	Russom et al. (1997)[88]	309	quantitative structure-activity relationships; expert systems; toxic action mode; aquatic toxicology; *Pimephales promela*	USA/USA
Three-dimensional quantitative structure-activity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor Xa	Böhm et al. (1999)[89]	282	N/A	Germany/Germany
Random forest: A classification and regression tool for compound classification and QSAR modeling	Svetnik et al. (2003)[31]	267	N/A	USA/USA
Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 2. Prediction of blood-brain barrier penetration	Clark (1999)[90]	264	N/A	UK/UK
GRid-INdependent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors	Pastor et al. (2000)[28]	261	N/A	Italy/Italy
Quantitative structure-activity relationship analysis of phenolic antioxidants	Lien et al. (1999)[91]	253	antioxidants; E_{HOMO}; E_{LUMO}; flavonoids; heat of formation; one-electron redox potential; phenols; quantitative structure-activity relationship (QSAR); TEAC; vitamin E derivatives; free radical	USA/USA
Electrotopological state indexes for atom types - a novel combination of electronic, topological, and valence state information	Hall and Kier (1995)[92]	250	N/A	USA/USA

N/A: not available; *TC2012*: total citations from the initial publication of the article up to the end of 2012.

tial for successful application and interpretation of QSPR models"[34] with *TC2012* = 553 was excluded, as it was granted the document type of review by its original journal but misclassified as an article (both proceedings paper and article) in WoS database due to database bias. Inspecting the words/phrases appearing in title and author keywords, it is easy to conclude that almost all the high-impact articles are relevant to methodological innovations other than simple case studies. This is in agreement with the finding in analyzing the top-cited papers in the journal *Quantitative Structure-Activity Relationships*.[45] The topics fall into three major groups: (1) proposing or optimizing the reliable, unambiguous and chemically interpretable mathematical algorithms in QSAR/QSPR modeling;[29–31,86] (2) designing novel, comprehensible and rapid algorithms for identifying mathematical descriptions to represent molecular features;[23,28,85,90,92] and (3) assessing and validating the robustness and predictability of the built QSAR/QSPR model.[33,87]

Table 3. The top 20 most prolific countries based on the total number of articles published. *TA*: total articles; *TAR* (%): rank and percentage of total articles; *SAR* (%): rank and the percentage of single country articles; *CAR* (%): rank and the percentage of international collaborative articles; *FAR* (%): rank and the percentage of first-authored articles; *RPR* (%): rank and the percentage of the corresponding-authored articles; *SA*: single country articles; *SA/TA R*(%): rank and percentage of single country articles in total articles.

Country	*TA*	*TAR* (%)	*SAR* (%)	*CAR* (%)	*FAR* (%)	*RPR* (%)	*SA/TA R* (%)
USA	2310	1 (21)	2 (17)	1 (35)	1 (17)	1 (17)	16 (64)
China	1873	2 (17)	1 (18)	5 (12)	2 (16)	2 (16)	4 (84)
India	1336	3 (12)	3 (14)	9 (6.0)	3 (12)	3 (12)	2 (89)
Italy	731	4 (6.6)	4 (4.5)	2 (15)	4 (5.0)	4 (5.0)	27 (53)
UK	679	5 (6.2)	6 (3.9)	3 (14)	5 (4.3)	5 (4.3)	35 (50)
Germany	577	6 (5.2)	8 (3.0)	4 (13)	7 (3.4)	7 (3.4)	36 (45)
Spain	485	7 (4.4)	9 (2.5)	6 (11)	8 (3.1)	8 (3.1)	39 (43)
Iran	431	8 (3.9)	5 (4.0)	16 (3.6)	6 (3.5)	6 (3.5)	5 (80)
France	411	9 (3.7)	12 (2.1)	7 (9.7)	10 (2.3)	11 (2.2)	42 (44)
Japan	366	10 (3.3)	7 (3.1)	13 (4.3)	9 (2.7)	9 (2.7)	11 (72)
Brazil	275	11 (2.5)	11 (2.3)	20 (3.1)	11 (2.2)	10 (2.2)	9 (73)
South Korea	233	12 (2.1)	10 (2.4)	39 (1.2)	12 (2.0)	12 (2.0)	3 (88)
Canada	230	13 (2.1)	15 (1.5)	11 (4.3)	14 (1.7)	13 (1.7)	24 (55)
Romania	216	14 (2.0)	14 (1.7)	22 (3.0)	13 (1.7)	13 (1.7)	14 (67)
Poland	214	15 (1.9)	13 (1.7)	26 (2.8)	15 (1.5)	15 (1.5)	13 (69)
Russia	205	16 (1.9)	16 (1.4)	15 (3.7)	16 (1.4)	17 (1.3)	19 (58)
Netherlands	200	17 (1.8)	18 (1.0)	10 (4.7)	18 (1.2)	18 (1.2)	37 (45)
Cuba	193	18 (1.8)	34 (0.40)	8 (6.7)	19 (1.1)	19 (1.1)	64 (18)
Sweden	192	19 (1.7)	17 (1.3)	18 (3.4)	17 (1.3)	16 (1.3)	20 (57)
Switzerland	162	20 (1.5)	21 (0.8)	14 (3.8)	22 (1.0)	22 (1.0)	41 (44)

Interestingly, almost all authors of the high-impact articles possess the *h*-index[93] lower than 20, with the only two exceptions of Tropsha, A (*h*-index = 26) and Hopfinger, AJ (*h*-index = 23), suggesting the most of high-impact articles came from the authors who are either less prolific or less influential. Meanwhile, the high-impact articles received not many citations from the authors with *h*-index higher than 20 (Table S4 Supporting Information): apart from some self-citations, four researchers Roy, K (*h*-index = 24), Gramatica, P (*h*-index = 23) and Yao, XJ (*h*-index = 21) cited the five high-impact articles.

3.4 Geographic Distribution of Publications

Analyzing the publication distributions is illustrative to evaluate the scientific contribution of a certain country/territory to our overall knowledge. Of the 11 004 articles with author affiliation information, 8623 (78%) were single country articles and 2381 (22%) were internationally collaborative articles. The approximate 8 : 2 ratios between single country and internationally collaborative articles were also observed in many topics relevant to biology and environmental sciences, such as drinking water,[94] photosynthesis,[95] estuary pollution,[96] and climate change,[97] implying that the global collaboration status is of low relevance to the issue scope whether it is at local or global level. All the articles come from 88 countries/territories. The indicators related to the number of single country and internationally collaborative articles, first authored and corresponding authored articles from the top 20 most prolific countries are given in Table 3, and the temporal trends of the top five countries, developed and developing countries groups are displayed in Figure S5, Supporting Information. The United States, China, and India come out in front in terms of the total article count. The United States contributed the most internationally collaborative articles (823; 35%), first-authored articles (1816; 17%) and corresponding-authored articles (1851; 17%); China contributed the most single country articles (1579; 18%). In general, the major body of article outputs came from developed countries: 61% of all articles were produced from 34 OECD member countries. In contrast, the article outputs in developing countries increased at higher growth rates: the article output from BRIICS (Brazil, Russia, India, Indonesia, China, and South Africa) countries increase remarkably faster (26 folds) than those from the United States (2.8 folds), Italy (5.4 folds) and the United Kingdom (2.8 folds) during 1993–2012 (Figure S5 Supporting Information). Back to Table 2, besides the total article count, developed countries independently contributed almost all the high-impact articles, among which the United States ranked first with seven of thirteen corresponding-authored and first-authored high-impact articles. It can be inferred that developed countries made greater contributions to methodological innovation in QSAR/QSPR research given the contents high-impact articles, which is similar to the previous general conclusions in analyzing overall high-impact SCI-EXPANDED journals' articles.[43] Meanwhile, the SA/TA ratio, the percentage of single country articles in total articles which characterizes the research independence of a country/territory, is remarkably high in middle-income developing countries like China (84%), India (89%), and Iran (80%); intermediate in developed countries like

United States (64%), Italy (53%) and Germany (45%) (Table 3); but the lowest in small countries like Iraq (9.1%; rank 67), Saudi Arabia (9.1%; rank 67), and Cyprus (5.6%; rank 69). This suggests the researchers from developed countries often pursue broader international cooperation in research, while those from developing countries confine themselves within domestic efforts. Table S5 (Supporting Information) presents the international collaboration matrix among the top ten most prolific countries, where normalized collaborative coefficients are calculated and colored to demonstrate the relative association strength. The results show that collaborations between the United States and China, the United States and the United Kingdom, together with China and France were most close and prolific in global QSAR/QSPR research.

3.5 Limitation and Perspective

Nowadays, the exploration and exploitation of QSAR/QSPR have already penetrated a diverse range of disciplines, and different scholars from diverse backgrounds favor on various alternative keywords to present their work, thus making it impossible to capture all QSAR/QSPR publications through a simple query fed with the limited retrieving words. Furthermore, it is a common practice in bibliometric analysis that a retrieving word should be excluded if its presence brings over half of retrieved publications irrelevant to the desired topic, unless the overall statistical conclusions might be biased without this word. Therefore, a dilemma was confronted that broadening retrieving words incurs irrelevant publications while narrowing leads to omission, which suggests the inherent limitations of bibliometric analysis. In this sense, the retrieved documents here are just a subset of the world's overall QSAR/QSPR publications, or in a statistical context, one sample from the unquantifiable QSAR/QSPR population.

By using only the core terms with stringent appearance of QSAR/QSPR, our study provided very preliminary insights to global QSAR/QSPR research performance and hotspots. With regard to individual publications (especially high-impact articles), ones dealing with the following kinds of topics may be inevitably omitted if our retrieving words are not contained in their front pages.

3.5.1 QSAR/QSPR-Rooted Subjects or Disciplines

Publications relevant to QSAR/QSPR-rooted subjects or disciplines, such as "chemometric(s)", "chem(o)informatic(s)", and "chemical informatics". QSAR/QSPR plays the most important roles in its subordinative subjects like chemometrics, chemoinformatics and information management, hence these subjects are often used as synonyms for referring to QSAR/QSPR research in many cases. For example, another 3671 articles can be retrieved if an additional term "chemometrics" was included into retrieving words, among them were some high-impact articles, like PLS package in R language programmed by Mevik and Wehrens[98] (*TC2012* = 134) that has been frequently adopted in QSAR/QSPR and other chemometric research. Despite the omission, excluding "chemometrics" from retrieving words is the last resort. This is because 2792 (76%) and 126 (3.4%) of the retrieved articles concentrated within analytical chemistry (e.g. spectroscopic and chromatographic topics) and automation & control system (e.g. process monitoring and fault detection) issues; even the remaining 753 articles were not all QSAR/QSPR relevant, if a detail analysis was carried out on their author keywords distribution (Table S6 Supporting Information).

3.5.2 Techniques Adopted in QSAR/QSPR

Publications dealing with specific techniques adopted in QSAR/QSPR research, like molecular (structural) descriptor, statistical algorithm and model validation technology. For example, Frank and Frideman[99] (*TC2012* = 637), a classical statistics-topic article that gained wide favor from QSAR/QSPR researchers and received 55 citations by our retrieved articles. This is because that many techniques used in QSAR/QSPR research were borrowed from chemometric, statistical and engineering literature, although they had not been exclusively designed for QSAR/QSPR at their birth; in fact, they gained as much popularity in other scientific domains as in QSAR/QSPR research. For instance, apart from QSAR/QSPR field, the above article[99] also received 323 (51%), 112 (18%), and 110 (17%) citations from probability statistics, artificial intelligence computer science and analytical chemistry fields, respectively.

3.5.3 Technical or Linguistic Varieties of QSAR/QSPR

Publications related to various technical or linguistic forms of QSAR/QSPR, for example, "quantitative structure-toxicity relationship (QSTR)",[100] and "quantitative structure-retention relationship (QSRR)",[101] even simply as "quantitative analysis" and "prediction". As the varieties of QSAR/QSPR, these techniques keep almost identical in principle, methodology and modeling procedurals; thus, they can be regarded as alias in specific areas.

Nevertheless, with regard to the statistical conclusions on global publication performance and overall research tendency, the validness is assured as negligible changes will be observed if additional terms are also included retrieving words in our pretests (data not shown), given the size of our retrieved article database is large enough (11 020 articles). This confirms our retrieving words seem not to reach the misleading conclusions in general, albeit they might lead to underestimation in number to some extent. However, more accurate retrieving words are still warranted to broaden the range of retrieved articles but reduce the irrelevant inclusion in the future bibliometric analysis. Besides, consistent, concerted and widely-accept-

ed common terminology, names, and acronyms are also needed in future QSAR/QSPR research.

4 Conclusion

The global QSAR/QSPR research trend and performance were assessed in this bibliometric study by combing the trend analysis and co-citation analysis. A total of 11020 original articles of 12767 documents were published during the period 1993–2012. The number of articles per year quadrupled from 1993 to 2006 and plateaued from 2007 onwards. *Journal of Chemical Information and Modeling* was the most active journal. Articles originated from 147 WoS categories, among which the research fields of drug design and synthesis, and environmental sciences were identified as the most prolific and the most promising regimes, respectively. The internal methodological innovations in acquiring molecular descriptors and modeling stimulated the publication increase in the research fields of drug design and synthesis, and chemoinformatics; while the external regulatory demands on model validation and reliability fueled the increase in environmental sciences. "Prediction endpoints", "statistical algorithms", and "molecular descriptors" were three pressing issues. The articles from developed countries were more abundant and influential, whereas those from developing countries increased at higher growth rates.

References

[1] H. Kubinyi, *Drug Discov. Today* **1997**, *2*, 457–467.
[2] T. W. Schultz, M. T. D. Cronin, J. D. Walker, A. O. Aptula, *Theochem.-J. Mol. Struct.* **2003**, *622*, 1–22.
[3] C. Hansch, D. Hoekman, A. Leo, D. Weininger, C. D. Selassie, *Chem. Rev.* **2002**, *102*, 783–812.
[4] 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, *Altern. Lab. Anim.* **2005**, *33*, 155–173.
[5] W. Lilienblum, W. Dekant, H. Foth, T. Gebel, J. Hengstler, R. Kahl, P.-J. Kramer, H. Schweinfurth, K.-M. Wollin, *Arch. Toxicol.* **2008**, *82*, 211–236.
[6] A. Brown-Crum, T. R. Fraser, *Trans. Roy. Soc. Edinb.* **1868**, *25*, 151–203.
[7] H. Kubinyi, *Quant. Struct.-Act. Relat.* **2002**, *21*, 348–356.
[8] L. P. Hammett, *J. Am. Chem. Soc.* **1937**, *59*, 96–103.
[9] L. P. Hammett, *Trans. Faraday Soc.* **1938**, *34*, 156–165.
[10] L. P. Hammett, in *Advances in Linear Free Energy Relationships* (Eds: N. B. Chapman, J. Shorterm), Plenum, London, **1972**, p. vii
[11] S. Wold, M. Sjöström, *Acta Chem. Scand.* **1998**, *52*, 517–523.
[12] O. R. Hansen, *Acta Chem. Scand.* **1962**, *16*, 1593–1600.
[13] R. W. Taft Jr, *J. Am. Chem. Soc.* **1952**, *74*, 3120–3128.
[14] C. Hansch, R. M. Muir, T. Fujita, P. P. Maloney, F. Geiger, M. Streich, *J. Am. Chem. Soc.* **1963**, *85*, 2817–2824.
[15] M. J. Kamlet, R. W. Taft, *J. Chem. Soc. Perkin* **1979**, *2*, 337–341.
[16] G. R. Famini, C. A. Penski, L. Wilson, *J. Phys. Org. Chem.* **1992**, *5*, 395–408.
[17] C. Hansch, P. P. Maloney, T. Fujita, R. M. Muir, *Nature* **1962**, *194*, 178–180.
[18] C. Hansch, T. Fujita, *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
[19] C. Hansch, *Acc. Chem. Res.* **1969**, *2*, 232–239.
[20] S. M. Free, J. W. Wilson, *J. Med. Chem.* **1964**, *7*, 395–399.
[21] L. B. Kier, L. H. Hall, W. J. Murray, M. Randić, *J. Pharm. Sci.* **1975**, *64*, 1971–1974.
[22] L. B. Kier, L. H. Hall, *Pharm. Res.* **1990**, *7*, 801–807.
[23] L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.
[24] M. L. Connolly, *J. Am. Chem. Soc.* **1985**, *107*, 1118–1124.
[25] M. Karelson, V. S. Lobanov, A. R. Katritzky, *Chem. Rev.* **1996**, *96*, 1027–1044.
[26] R. D. Cramer, D. E. Patterson, J. D. Bunce, *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
[27] G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130–4146.
[28] M. Pastor, G. Cruciani, I. McLay, S. Pickett, S. Clementi, *J. Med. Chem.* **2000**, *43*, 3233–3243.
[29] S. Wold, M. Sjostrom, L. Eriksson, *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
[30] D. Rogers, A. J. Hopfinger, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.
[31] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
[32] *OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models*, Organization for Economic Co-operation and Development, **2004**; available at http://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (assessed: Sep. 3rd, **2013**).
[33] A. Golbraikh, A. Tropsha, *J. Mol. Graph.* **2002**, *20*, 269–276.
[34] A. Tropsha, P. Gramatica, V. K. Gombar, *QSAR Comb. Sci.* **2003**, *22*, 69–77.
[35] P. Gramatica, *QSAR Comb. Sci.* **2007**, *26*, 694–701.
[36] J. Jaworska, N. Nikolova-Jeliazkova, T. Aldenberg, *Altern. Lab. Anim.* **2005**, *33*, 445–459.
[37] S. Xie, J. Zhang, Y.-S. Ho, *Scientometrics* **2008**, *77*, 113–130.
[38] N. Mao, M.-H. Wang, Y.-S. Ho, *Hum. Ecol. Risk Assess.* **2010**, *16*, 801–824.
[39] H. Fu, Y.-S. Ho, *J. Inform.* **2013**, *7*, 210–222.
[40] C. Chen, *J. Am. Soc. Inf. Sci. Technol.* **2006**, *57*, 359–377.
[41] H. Fu, M.-H. Wang, Y.-S. Ho, *J. Colloid Interf. Sci.* **2012**, *379*, 148–156.
[42] Y.-S. Ho, *Scientometrics* **2013**, *94*, 1297–1312.
[43] E. Garfield, *Nature* **1970**, *227*, 669–671.
[44] P. Willett, *QSAR Comb. Sci.* **2009**, *28*, 1231–1236.
[45] R. A. Jishi, P. Willett, *J. Chem. Inf. Model.* **2010**, *50*, 1915–1923.
[46] P. Willett, *J. Comput. Aided Mol. Des.* **2012**, *26*, 153–157.
[47] L. Li, G. Ding, N. Feng, M.-H. Wang, Y.-S. Ho, *Scientometrics* **2009**, *80*, 39–58.
[48] M.-H. Wang, Y.-S. Ho, *Arch. Environ. Sci* **2011**, *5*, 1–10.
[49] E. Garfield, *Current Contents* **1990**, *32*, 5–9.
[50] W.-T. Chiu, Y.-S. Ho, *Scientometrics* **2005**, *63*, 3–23.
[51] M.-H. Wang, T.-C. Yu, Y.-S. Ho, *Scientometrics* **2010**, *84*, 813–820.
[52] C. Chen, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5303–5310.
[53] C. Chen, F. Ibekwe-SanJuan, J. Hou, *J. Am. Soc. Inf. Sci.* **2010**, *61*, 1386–1409
[54] B. L. Clarke, *Science* **1964**, *143*, 822–824.

[55] M. J. Dewar, E. G. Zoebisch, E. F. Healy, J. J. Stewart, *J. Am. Chem. Soc.* **1985**, *107*, 3902 – 3909.
[56] Y. Fan, L. M. Shi, K. W. Kohn, Y. Pommier, J. N. Weinstein, *J. Med. Chem.* **2001**, *44*, 3254 – 3263.
[57] J. K. Buolamwini, H. Assefa, *J. Med. Chem.* **2002**, *45*, 841 – 852.
[58] B. Silverman, D. E. Platt, *J. Med. Chem.* **1996**, *39*, 2129 – 2140.
[59] G. Bravi, E. Gancia, P. Mascagni, M. Pegna, R. Todeschini, A. Zaliani, *J. Comput. Aided Mol. Des.* **1997**, *11*, 79 – 92.
[60] M. Pastor, G. Cruciani, S. Clementi, *J. Med. Chem.* **1997**, *40*, 1455 – 1464.
[61] K. Hasegawa, T. Kimura, K. Funatsu, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 112 – 120.
[62] H. Chen, J. Zhou, G. Xie, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 243 – 250.
[63] I. V. Tetko, V. V. Kovalishyn, D. J. Livingstone, *J. Med. Chem.* **2001**, *44*, 2411 – 2420.
[64] M. Zeeman, C. M. Auer, R. G. Clements, J. V. Nabholz, R. S. Boethling, *SAR QSAR Environ. Res.* **1995**, *3*, 179 – 201.
[65] European Commission, *Regulation concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency*, Regulation (EC) No. 1907/2006, **2006**; available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2006R1907:20110505:en:PDF (accessed: Nov. 24th, **2013**)
[66] M. Tichý, M. Rucki, *Interdiscip. Toxicol.* **2009**, *2*, 184 – 186.
[67] M. A. Khan, Y.-S. Ho, *Sci. Total Environ.* **2012**, *431*, 122 – 127.
[68] O. Guner, *Curr. Top. Med. Chem.* **2002**, *2*, 1321 – 1332.
[69] J. S. Jaworska, M. Comber, C. Auer, C. Van Leeuwen, *Environ. Health Perspect.* **2003**, *111*, 1358 – 1360.
[70] P. Liu, W. Long, *Int. J. Mol. Sci.* **2009**, *10*, 1978 – 1998.
[71] H. Si, T. Wang, K. Zhang, Z. Hu, B. Fan, *Bioorg. Med. Chem.* **2006**, *14*, 4834 – 4841.
[72] R. Guha, D. Dutta, P. C. Jurs, T. Chen, *J. Chem Inf. Model.* **2006**, *46*, 1836 – 1847.
[73] Q. Du, R. Huang, K. Chou, *Curr. Protein Pept. Sci.* **2008**, *9*, 248 – 259.
[74] H. Du, J. Wang, Z. Hu, X. Yao, X. Zhang, *J. Agric. Food Chem.* **2008**, *56*, 10785 – 10792.
[75] M. Akamatsu, *Curr. Top. Med. Chem.* **2002**, *2*, 1381 – 1394.
[76] L. H. Hall, B. Mohney, L. B. Kier, *Quant. Struct.-Act. Relatsh.* **1991**, *10*, 43 – 51.
[77] R. A. Pearlstein, R. J. Vaz, J. Kang, X.-L. Chen, M. Preobrazhenskaya, A. E. Shchekotikhin, A. M. Korolev, L. N. Lysenkova, O. V. Miroshnikova, J. Hendrix, *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829 – 1835.
[78] J. M. Sanders, A. O. Gómez, J. Mao, G. A. Meints, E. M. Van Brussel, A. Burzynska, P. Kafarski, D. González-Pacanowska, E. Oldfield, *J. Med. Chem.* **2003**, *46*, 5171 – 5183.
[79] M. T. Makhija, R. T. Kasliwal, V. M. Kulkarni, N. Neamati, *Bioorg. Med. Chem.* **2004**, *12*, 2317 – 2333.
[80] Y. Wang, H. Liu, C. Zhao, H. Liu, Z. Cai, G. Jiang, *Environ. Sci. Technol.* **2005**, *39*, 4961 – 4966.
[81] Z. Wang, X. Han, L. Wang, *Chin. J. Struct. Chem.* **2005**, *24*, 851 – 857.
[82] P. L. Gross, E. Gross, *Science* **1927**, *66*, 385 – 389.
[83] E. Garfield, *Curr. Contents* **1973**, *39*, 5 – 6.
[84] K.-Y. Chuang, M.-H. Wang, Y.-S. Ho, *Malays. J. Libr. Inf. Sci.* **2013**, *18*, 47 – 63.
[85] G. Klebe, U. Abraham, T. Mietzner, *J. Med. Chem.* **1994**, *37*, 4130 – 3146
[86] M. Baroni, G. Costantino, G. Cruciani, D. Riganelli, R. Valigi, S. Clementi, *Quant. Struct.-Act. Relat.* **1993**, *12*, 9 – 20
[87] D. E. Patterson, R. D. Cramer, A. M. Ferguson, R. D. Clark, L. E. Weinberger, *J. Med. Chem.* **1996**, *39*, 3049 – 3059.
[88] C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister, R. A. Drummond, *Environ. Toxicol. Chem.* **1997**, *16*, 948 – 957.
[89] M. Böhm, J. Sturzebecher, G. Klebe, *J. Med. Chem.* **1999**, *42*, 458 – 477.
[90] D. E. Clark, *J. Pharm. Sci.* **1999**, *88*, 815 – 821.
[91] E. J. Lien, S. J. Ren, H. Y.H. Bui, R. B. Wang, *Free Radic. Biol. Med.* **1999**, *26*, 285 – 294.
[92] L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039 – 1045
[93] J. E. Hirsch, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 16569 – 16572
[94] H. Fu, M.-H. Wang, Y.-S. Ho, *Sci. Total Environ.* **2013**, *443*, 757 – 765.
[95] J. Yu, M.-H. Wang, M. Xu, Y.-S. Ho, *Photosynthetica* **2012**, *50*, 5 – 14.
[96] J. Sun, M.-H. Wang, Y.-S. Ho, *Mar. Pollut. Bull.* **2012**, *64*, 13 – 21.
[97] J. Li, M.-H. Wang, Y.-S. Ho, *Glob. Planet. Change* **2011**, *77*, 13 – 20.
[98] B. H. Mevik, R. Wehrens, *J. Stat. Softw.* **2007**, *18*, 1 – 23.
[99] I. E. Frank, J. H. Frideman, *Technometrics* **1993**, *35*, 109 – 135.
[100] A. G. Siraki, T. S. Chan, P. J. O'Brien, *Toxicol. Sci.* **2004**, *81*, 148 – 159.
[101] R. Kaliszan, *J. Chromatogr. A* **1993**, *656*, 417 – 435.

Received: December 20, 2013
Accepted: August 4, 2014
Published online: September 10, 2014

