Abstract:

Organic pollutants that resist degradation in the environment can accumulate in body tissues and cause unavoidable intoxications to organisms in wild life as well as humans. The possible effects, usually increasing with the cumulative exposure to such chemicals, are not always addressed adequately in risk assessment procedures evaluating long and short-term contact hazard. Thus, chemicals accumulation, degradation and environmental fate are of prime concern for REACH when defining side effects due to chronic exposure.

Characteristics and behavior of organic pollutants have been investigated experimentally during the last decades by use of various methods of trace analysis. However, the available data still contains several gaps. In this aim, REACH promotes the use of alternative methods to reduce the number of animal tests and suggests *in-silico* methods such as Quantitative Structure-Activity Relationships (QSARs) to fill the lack of knowledge.

The goal of this thesis, in the framework of the ECO-ITN project, was to build QSAR models with high reliability based on good experimental data for optimal estimation of environmental endpoints of interest for REACH. New molecular descriptors and feature selection techniques have been tested paying particular attention to the validation steps and applicability domain definition.

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

MARIE CURIE ACTIONS

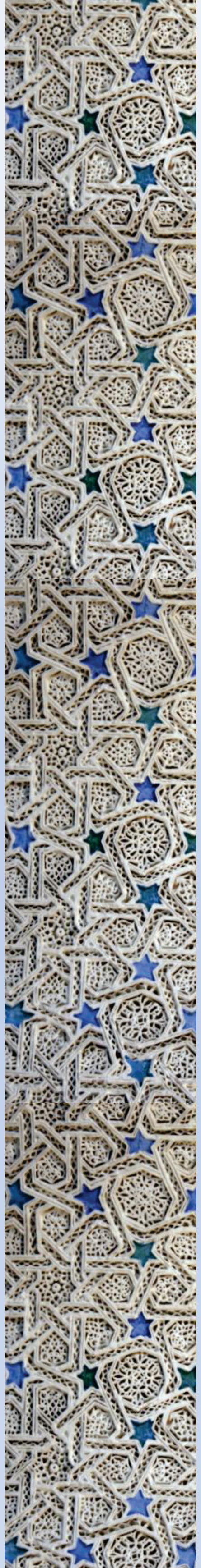
eco
environmental chemoinformatics
marie curie initial training network

Cover illustration:
Arabesque pattern, Bardo Palace Tunis.

PhD thesis

Kamel MANSOURI

2012/2013



New molecular descriptors for estimating degradation and fate of organic pollutants by QSAR/QSPR models within REACH

Kamel MANSOURI

University of Milano-Bicocca

Department of Environmental Sciences

Ph.D. Thesis Cycle: XXV

New molecular descriptors for estimating degradation and fate of organic pollutants by QSAR/QSPR models within REACH

Kamel MANSOURI

Tutor: Prof. Roberto TODESCHINI

Co-tutors: Dr. Viviana CONSONNI
Dr. Davide BALLABIO

2012/2013

This thesis was defended at the University of Milano-Bicocca, on the 27th of June 2013, in front of the jury composed by:

Prof. Willie PEIJNENBURG
Institute of Environmental Sciences
University of Leiden

Prof. Lutgarde BUYDENS
Faculty of Sciences
Radboud University Nijmegen

Coordinator:
Prof. Marco VIGHI
Department of Enivironmental Sciences and Earth
University of Milano-Bicocca

Acknowledgments:

First of all I would like to acknowledge my supervisor Prof. Roberto Todeschini for his guidance during this work.

Special thanks to my co-tutors Viviana and Davide not only for their precious scientific suggestions but also for their help with the administrative aspects and their ability to solve the problems we faced.

I want to thank also Prof. Tomas Öberg and Dr. Igor Tetko for the opportunity of making internships in their labs.

I am grateful for Dr. Igor Tetko and Dr. Eva Schlosser for coordinating and managing the ECO project.

Many thanks to all the Milano Chemometrics Group members I worked with during these 3 years. Especially Andrea, Alberto, Matteo, Francesca, and the ECO fellows Faizan, Tine and Eva who contributed to this work.

I'm also thankful for my family and all my dear friends especially my best friend Aymen who always supported me in all situations.

الى أمي

Contents

List of Figures

List of Tables

Preface

The ECO project

This thesis was carried out in the framework of the Environmental ChemOinformatics project (ECO) which is a Marie Curie Initial Training Network, Funded by the European Commission under FP7 - People Program. The project started on 01/10/2009 and planned to end on the 30/09/2013 [1].

The aim of the Marie Curie Initial Training Networks (ITN) is to enhance the career of young researchers in Europe. The ECO-ITN project aimed at training the fellows in the field of environmental Chemoinformatics and to contribute to the implementation of the REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) EU regulation. The primary objective of this ITN was to contribute to the education of environmental chemo-informaticians in both environmental sciences and computational *in-silico* methods. The fellows of the network were then expected to apply their knowledge for the implementation of REACH in particular with respect to the replacement, refinement and reduction of animal tests by alternative (*in-silico* and *in-vitro*) methods.

The project involved seven academic institutions from five EU countries (Germany, The Netherlands, Spain, Sweden and Italy).

The expertise of the ECO partners consists of both experimental and computational chemistry including traditional analytical techniques, modern bio-screening methods, molecular mechanics, semi-empirical and ab-initio quantum chemical calculations, in addition to the commonly used Chemoinformatic and Chemometric techniques. During the project, several endpoints of interest for REACH were evaluated by means of both experimental and computational approaches. Studies on physico-chemical properties, toxicological and complex problems of metabolism and biodegradation were carried out. Properties of complex mixtures, fate modeling as well as exposure assessment of nanomaterials were also addressed.

Thesis goals and structure

The main goal of this thesis was to contribute in filling the lack of knowledge about chemicals for regulatory reasons of specific endpoints of interest to the European legislation REACH. The study was focused on specific molecular properties related to biodegradation and environmental fate of chemicals. Methods in agreement with the scope of REACH, in avoiding animal testing, such as QSAR modeling were developed in order to predict the endpoints of interest. A particular attention was paid to molecular descriptors and their relationships to the modeled endpoints.

The thesis was structured in three parts. In the first part, a general introduction about Persistent Organic Pollutants (POPs), their physicochemical properties, pathways to the environment and their acute effects on humans and wild life is given. The REACH legislation is then introduced, as well as the role of QSARs as a tool of trust to provide the missing information about the chemical substances with the desired reliability.

In the second part of the thesis, the different steps required for QSAR modeling and the related methods used in this study are introduced. Since the predictions of a QSAR model are influenced by the experimental values used as response to be predicted, it is fundamental to filter the available information and ensure a high quality initial dataset. Methods and algorithms used for this purpose are explained. Then, classical and recent advances in variable selection methods are elucidated, since the selection of a proper set of molecular descriptors is usually an important step for QSAR modeling. Once the models were built using the suitable regression/classification methods, it had to be validated and its accuracy measured then its domain of applicability defined.

The third part of the thesis showed how the previously defined methods have been used in order to build and validate the QSAR models. It presented the preliminary results of the conducted studies and summaries of the published articles. The selected endpoints of interest to the project were the octanol-water partition coefficient, bioaccumulation factors and the ready biodegradability of chemicals. The obtained results were evaluated in comparison with the literature and the selected molecular descriptors were discussed in relation to the studied endpoints. In addition to the modeling results, a comparison study on different applicability domain approaches was carried out and a study on the activity cliffs in the QSAR datasets was introduced and the first obtained results are discussed.

Part I: Introduction

1. POPs and pathways to the environment

The rapid technological and industrial development during the last decades aimed to increase welfare in most parts of the globe. However, it has also led to side impacts on human health and the environment. That was due to the fact that chemicals production grows roughly in line with the economies especially in the developed countries, releasing toxic substances to the environment. From the several hundreds of million tons of chemicals produced every year, Europe has by far the largest part accounting for 38% of the total [2]. About 2% of Europe's GDP and 7% of its employment are provided by chemical industry. The 33% of world-wide chemicals production are detained by western Europe, of which Germany provides 26%, France 19%, while UK and Italy 12% each [3].

Since hundreds of new substances are marketed each year, the total number of chemicals available on the market is possibly exceeding the 100,000 chemicals that were registered in the European Inventory of Existing Commercial Chemical Substances (EINECS) in 1981 [4]. The rising quantities and variety of substances released in the environment increase the potential damage to humans and biota. However, about 75% of these substances are associated with insufficient toxicity and eco-toxicity data [4].

Potentially dangerous marketed chemicals were developed and used for different applications, such as polychlorinated biphenyls (PCBs) as insulating fluids in electrical equipment, hexachlorobenzene (HCB) to protect crops and wood from fungi, and polybrominated diphenyl ethers (PBDEs) to reduce the risk of fires. Such substances are often associated with high degree of halogenations and turned out to be persistent in the environment as well as toxic for living organisms. They are called persistent organic pollutants (POPs).

Evidence of POP toxicity has been mounted by associating them with chronic and acute effects deriving from long term exposure. In addition, POPs can also cause cancer, allergies, diseases of the immune system, damage to nervous systems, developmental disorders, reproductive disorders as well as damage to wildlife [5–7].

Rapid progress is being made to reduce the releases of POPs. Also, the production of such substances is being gradually phased out by installing alternative industrial processes and cleaning equipment. However, POPs continue to pose risk to the environment long periods after their production and use because of their slow degradation. In fact, due to their persistency, these chemicals were also detected in different areas far from their original site of production [8,9].

To reduce the risks associated with POPs, an agreement has been adopted by the European countries under the Convention on Long-Range Trans-boundary Air Pollution at the fourth European conference of environment in June 1998 (Aarhus, Denmark). Soon after in Montreal, the global community started negotiations about a worldwide treaty for safety from chemicals which can be released in one part of the globe and distributed in vast geographical areas. In 2001, the Stockholm convention on POPs was adopted and entered into force in 2004 [10,11].

In the framework of the European Commission's stock-taking legislative instruments to govern chemical substances, risk assessment is used to identify potential harm caused by different exposure levels. Further knowledge about these toxic chemicals and their pathways to the environment is needed to fill the huge data gaps and prevent their toxicity effects.

1.1. General properties of POPs

The concept of POP is associated with the Stockholm Convention (SC), the global treaty developed under the United Nation Environmental Program [12]. The SC intent was to identify the chemicals which have to be reduced or eliminated from the intentional/unintentional production and use chain. The three properties typically used to identify POPs are persistency, bioaccumulating potential and toxicity (PBT) [13,14]. Initially, the set of POPs consisted of twelve chlorinated chemicals, called "the dirty dozen", fulfilling the PBT and long range environmental transport criteria. Later in 2009, the list was updated by adding nine substances including few polybrominated diphenyl ethers (PBDEs) [11].

POPs are substances that resist degradation in the environment and poorly dissolve in water (hydrophobic). Such compounds often have a carbon backbone with halogen substituents, for instance, bromine for PBDEs and chloride for PCBs. POPs with the same backbone structure but different halogen numbers and positioning are called congeners. Usually, congeners are associated with different physicochemical properties that are likely affecting their fate and transport in the environment [10,15].

POPs tend to partition to organic matter in soil and sediments or particles in suspension in water, while in biota these compounds accumulate in lipids. Their solubility is known to be similar in lipids while it exhibits large variations in water. Therefore, one of their major physicochemical differences can be expressed in terms of hydrophobicity [16]. The most common measurements of hydrophobicity is the octanol-water partition coefficient expressed in log values (log KOW, log POW or log P) and calculated by the ratio between the concentration in water and 1-octanol at equilibrium [17]. Their hydrophobicity degree was demonstrated to be correlated with the number of halogens [17–19].

Their long range atmospheric transport ability is due to their volatility allowing them to have repeated evaporation and deposition cycles [20]. They can also be attached to particles that can be transported for long distances in air and water [21].

The persistency of a chemical do not depends only on its physicochemical properties, but also on the environmental conditions including the types of microbes living in the sediments and the concentration of hydroxyl radicals in the atmosphere [16].

Even if anaerobic dehalogenation is a possible way of degradation, POPs half life is very long and can reach, in the case of PCDD/Fs, several decades to centuries [22–24]. The hydrophobic property in addition to persistency, enable a POP to bioaccumulate and reach high concentrations in biota [14].

Bioaccumulation and bioconcentration factors (BAF and BCF, respectively) are two important measurements for the accumulation of chemicals in organisms. These factors are calculated as by the ratio between the concentrations in the organism and the surrounding media such as water or sediments [25]. BAF takes in consideration all uptake routes, including respiratory, dermal and gastrointestinal systems. While for BCF calculation, only the passive ways such as respiratory and dermal system are considered [25]. Due to their accumulating effect, the acute toxicity of the POPs is mainly manifested in the top predators of the food chain and particularly in fish-eating organisms [26,27].

1.2. Pathways to the environment

Chemical substances usually find their way into the environment via industrial waste and emissions, agricultural production and consumer uses. Once in the environment, they can interact with the hosting media to break down into other compounds with different properties or persist for long periods. For effective risk assessment of chemicals, it is essential to track their environmental fate and their exposure implications from manufacture to marketing and use. For each chemical compound, transport through air and water as well as its deposition into soil and sediments should be investigated. Multimedia fate models are also used to estimate the potential exposure to chemicals by assessing the inputs and outputs in a given geographical region [16].

Air is likely to be the main way most volatile POPs travel through. Due to the "grasshopper" effect, substances released in one part of the world can be transported to very far regions. This fact explains the origin of the POPs found in the Arctic or on high mountains [28].

Since water covers about 70% of the Earth's surface, it is highly probable that POPs are transported attached to particles and organic matter in suspension and, subsequently, end up to deposit in sediments [29]. However, the highest concentration of POPs in sediments is always detected close to the original sources [30–33].

Even with decreased emissions from the sources, due to their persistency, POPs can continuously contaminate the aquatic environment by dispersion to biota living in the sediments [34,35].

Once in living organisms, these pollutants can increase concentration in tissues of animals and accumulate at the highest levels of the food chain including humans. This process is called biomagnification. Thus, the complexity of the multiple exposure modes of these substances requires more knowledge about all chemicals to be marketed. To avoid the dangerous effects of direct contact or long term accumulation, only safe chemicals should be authorized to be manufactured.

2. Regulation of chemicals in Europe

The regulation process of chemicals in Europe started in 1976 and it restricted the marketing or use of only few hundreds of substances classified as carcinogenic, mutagenic or toxic to reproduction [36].

For a more safe manufacture and use of chemicals available in the European market, the implementation of a new legislation was required. The new regulated procedure aiming at evaluating the physico-chemical properties of both new and existing chemicals and their adverse effects on humans and the environment. Thus, the new regulation (REACH) was made aiming at assessing the existing substances within a process of eleven years.

It is known that most of the manufactured chemicals are missing information about toxicity [37,38]. In order to bridge this huge gap of knowledge on chemicals without increasing the actual numbers of animals used in the required tests, the European Commission made suggestions about alternatives to animal testing. This new system encourages the refinement of replacement strategies such as the development of new *in-vitro* methods but also the use of the validated *in-silico* techniques including computational predictive models.

2.1. REACH, the European legislation about chemicals

REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) is the new European Community regulation on chemical substances and their safe use starting from the 1st of June 2007 [39].

REACH aimed to protect humans, wild life and the environment by assessing the risks that can be caused by chemical substances in a gradual process. The most dangerous chemicals are going to be progressively substituted as soon as suitable alternatives are found. These goals should be achieved in transparency without altering the innovative capability and competitiveness of the chemical industry.

REACH is expected to have a gradual positive impact on health by restricting substances of high concern that can be linked to cancers, skin irritation, respiratory diseases, vision disorders, asthma, endocrine disrupting, inter alia.

According to World Bank estimates and other prudent assumptions, REACH would result in a 10% reduction of diseases caused by chemicals [40]. Assuming that these diseases account for about 1% of the overall burden of all types of disease in Europe, the reduction of 0.1% would be equivalent to avoiding 4500 deaths every year [36].

The implementation of the REACH legislation will also increase the information on hazards of chemicals and thus improve the quality of the environment. It aims to improve the assessment of persistent, bio-accumulative and toxic substances so as to prevent them from polluting the air, water and soil.

According to REACH, providing safety information and assessing risk of chemicals is responsibility of manufacturers or importers. The required properties of the substances should be gathered before dealing it in the market. This necessary information for the safe handling of chemicals should be registered in the central database managed by the European Chemicals Agency (ECHA, Helsinki).

2.2. The European Chemicals Agency (ECHA)

The role of ECHA within REACH is to ensure the proper implementation of the legislation and build credibility with all stakeholders by managing the technical, scientific and administrative aspects of the regulation at Community level [41]. The central point that the Agency acts can be summarized as following: management of the registration process, evaluation of the dossiers, taking decisions about the suspicious chemicals and coordinating between consumers and professionals by running databases of the available hazard information.

Another important role of ECHA is to enable sharing of the public information about chemicals at the pre-registration stage by means of substance information exchange forums set-up for the purpose. Such forums are useful to fill the lack of sufficient experimental and predicted information about chemicals in order to avoid testing on vertebrate animals and costs accordingly.

2.3. Mode of action within REACH

The idea behind REACH is that chemicals should be tested for any harm to humans or the environment by manufacturers or importers before putting them on the European market. This is pushing the industries to acquire more knowledge about their products and assess any potential risk. Thus, the only task left for the authorities is to make sure industries are compliant with all the requirements about substances of high concern.

A registration dossier should be submitted to ECHA for each substance manufactured or imported in quantities of 1 ton or above per year otherwise the product will not be allowed in the European markets [36]. The dossiers of substances potentially harmful to human health or the environment are prioritized. According to REACH, the dangerous substances are classified into: carcinogenic, mutagenic or toxic to reproduction, persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB). Dossiers of such suspected substances should contain additional physicochemical properties and relevant eco-toxicological information.

For the chemicals exceeding the quantity of 10 tons per year, a Chemical Safety Report (CSR) is needed. This report should include an assessment of the potential hazards as well as a classification to PBT or vPvB substances. The CSR is also supposed to include an exposure scenario for potentially dangerous substances.

According to REACH requirements, new experimental testing is allowed only if there are no alternatives to provide information about the substance. The use of existing information or techniques such as *in-vitro*, quantitative structure-activity relationships (QSARs) and read across are, therefore, prioritized.

3. QSARS for regulatory purposes

3.1. QSARs and REACH

One of the central principles of REACH legislation is to keep animal testing as the last resort to provide the required information about the submitted substances. Alternatives to animal testing are therefore promoted and special mechanisms were built-in for the purpose. QSARs are particularly encouraged and their use is recognized within the regulation's legal text by detailing special guidance documents [42].

QSARs are used to predict the behavior of chemicals from their structures, leading to better understanding of the adverse effects of the studied substances in cells and tissues. These modeling techniques make use of existing experimental data to predict new chemicals. The conceptual basis of QSARs is that similar structures are expected to exhibit similar biological behavior. The appropriate theoretical descriptors calculated from structural information are used to train the models and predict the biological activity of the chemicals. Thus, the environmental and eco-toxicological endpoints of interest could be assessed complying with the regulatory requirements for human health and minimizing, at the same time, the need for animal testing.

Different principles and guidelines for QSARs have been established by the REACH authorities in order to harmonize the models used for predictions.

Even being a highly valuable tool, any inappropriate use of these methods could cause a failure at REACH compliance check. Subsequently, a move forward animal testing can be made, which is in disagreement with reducing the costs and waiving animal test requirements.

3.2. OECD Principles for the Validation of QSARs

Five principles to establish the validity of QSAR models for use in regulatory purposes and assessment of chemical safety have been adopted at the 37th Meeting of Chemicals Committee and Working Party on Chemicals, Pesticides & Biotechnology, held in Paris on 17-19 November by the OECD Member Countries [43,44].

In this work, attention was paid to these principles during the QSAR modeling procedure. The evaluation of each of the five principles is an important condition in order to propose models to be applied for the regulatory purposes of REACH , which was the aim of this thesis.

The OECD principles intended to be considered in QSAR model validation for regulatory purposes within REACH, are as follows:

Principle 1: Defined Endpoint

Since experimental protocols and conditions determining the same endpoint may vary from a laboratory to another, it is therefore important to ensure clarity in the endpoint that a given model is predicting. To avoid any misleading ambiguity regarding the interpretation of the defined endpoint, guidelines have been developed to meet the information requirements of a given regulatory purpose and in the same time, the scientific sense of defined endpoint referring to a specific effect on a specific tissue/organ under precise conditions.

Principle 2: Unambiguous Algorithm

Transparency is essential in the used algorithm for building the model and generating the predictions for a chemical's specific endpoint from its structure and/or physicochemical properties. This information is useful to independently establish the performance and the reproducibility of the predictions of a given model. Any missing information about the used algorithm, which is usually the case in commercially-developed models, could rise ambiguity and represent a barrier for regulatory acceptance of the model.

Principle 3: Defined Domain of Applicability

Since the reliability of predictions by QSAR models is usually associated with limited types of chemical structures, physicochemical properties and mechanisms of action, a defined applicability domain is needed. It is the duty of QSARs developers to define the needed information and the appropriate methods for establishing the applicability domains of their models.

Principle 4: Appropriate Measures of Goodness-of-Fit, Robustness and Predictivity

The intent of this principle is to include all the three steps of the development of a QSAR model. Proper techniques to measure the degree of fitting of the studied endpoint to the structures of the used chemicals should be applied. The robustness of a model is determined in the validation step to avoid any over-fitting , while its predictive ability could be checked by an external test set of compounds that were not included in the fitting step.

Principle 5: Mechanistic Interpretation if possible

It is known that is not always easy to provide a mechanistic interpretation of QSARs from a scientific point of view, it could also happen that a multitude of interpretations are possible for a unique model. Thus, such information is not mandatory for a model to be accepted in a regulatory context. The intent of this fifth principle is to encourage documenting any attempt to associate the significance of the used descriptors to the endpoint that the model aimed to predict.

References

1. ECO-ITN Environmental ChemOinformatics http://www.eco-itn.eu/ (accessed Apr 21, 2013).

2. United Nations Economic Commission for Europe http://www.unece.org/ (accessed Apr 21, 2013).

3. Cefic | European Chemical Industry Council http://www.cefic.org/ (accessed Apr 21, 2013).

4. Allanou, R.; G. Hansen, B.; Van der Bilt, Y. *Public Availability of Data on EU High Production Volume Chemicals*; European Commission, Joint Research Centre, Institute for Health and Consumer Protection, European Chemicals Bureau: Ispra (VA), 21020, Italy, 1999.

5. Jacobson, J. L.; Jacobson, S. W. Intellectual Impairment in Children Exposed to Polychlorinated Biphenyls in Utero. *New England Journal of Medicine* **1996**, *335*, 783–789.

6. White, S. S.; Birnbaum, L. S. An Overview of the Effects of Dioxins and Dioxin-like Compounds on Vertebrates, as Documented in Human and Ecological Epidemiology. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* **2009**, *27*, 197–211.

7. IARC International Agency for Research on Cancer. Monographs on the Evaluation of Carcinogenic Risks to Humans, Polychlorinated dibenzo-para-dioxins and polychlorinated dibenzofurans. In; IARC Press ; Distributed by the World Health Organization Distribution and Sales, 1997; Vol. 69.

8. AMAP. Arctic Monitoring and Assessment Programme *Arctic pollution 2009*; Arctic Monitoring and Assessment Programme: Oslo, Norway, 2009.

9. Ballschmiter, K.; Hackenberg, R.; Jarman, W. M.; Looser, R. Man-made chemicals found in remote areas of the world: The experimental definition for POPs. *Environmental Science and Pollution Research* **2002**, *9*, 274–288.

10. Persistent Organic Pollutants(POPs) http://www.chem.unep.ch/pops/ (accessed Apr 21, 2013).

11. STOCKHOLM CONVENTION ON PERSISTENT ORGANIC POLLUTANTS http://www.pops.int/documents/meetings/dipcon/meetingdoclist_en.htm (accessed Apr 21, 2013).

12. United Nations Environment Programme (UNEP) - Home page http://www.unep.org/ (accessed Apr 21, 2013).

13. Aronson, D.; Howard, P. H. Evaluating potential POP/PBT compounds for environmental persistence. *Final Report Prepared under Contract to the Chemical Manufacturer's Association* **1999**.

14. Gobas, F. A. P. C.; De Wolf, W.; Burkhard, L. P.; Verbruggen, E.; Plotzke, K. Revisiting bioaccumulation criteria for POPs and PBT assessments. *Integrated*

Environmental Assessment and Management **2009**, *5*, 624–637.

15. Brown, F. R.; Winkler, J.; Visita, P.; Dhaliwal, J.; Petreas, M. Levels of PBDEs, PCDDs, PCDFs, and coplanar PCBs in edible fish from California coastal waters. *Chemosphere* **2006**, *64*, 276–286.

16. Mackay, D. *Multimedia environmental models: the fugacity approach*; 2nd ed.; Lewis Publishers: Boca Raton, 2001.

17. Hawker, D. W.; Connell, D. W. Octanol-water partition coefficients of polychlorinated biphenyl congeners. *Environ. Sci. Technol.* **1988**, *22*, 382–387.

18. Åberg, A.; MacLeod, M.; Wiberg, K. Physical-Chemical Property Data for Dibenzo-p-dioxin (DD), Dibenzofuran (DF), and Chlorinated DD/Fs: A Critical Review and Recommended Values. *Journal of Physical and Chemical Reference Data* **2008**, *37*, 1997–2008.

19. Braekevelt, E.; Tittlemier, S. A.; Tomy, G. T. Direct measurement of octanol-water partition coefficients of some environmentally relevant brominated diphenyl ether congeners. *Chemosphere* **2003**, *51*, 563–567.

20. Wania, F.; Mackay, D. Tracking the distribution of persistent organic pollutants. *Environmental Science and Technology* **1996**, *30*, 390A–397A.

21. Scheringer, M.; Jones, K. C.; Matthies, M.; Simonich, S.; Van De Meent, D. Multimedia partitioning, overall persistence, and long-range transport potential in the context of pops and pbt chemical assessments. *Integrated Environmental Assessment and Management* **2009**, *5*, 557–576.

22. Adriaens, P.; Fu, Q.; Grbic-Galic, D. Bioavailability and Transformation of Highly Chlorinated Dibenzo-p-Dioxins and Dibenzofurans in Anaerobic Soils and Sediments. *Environ. Sci. Technol.* **1995**, *29*, 2252–2260.

23. Brown, J. F.; Bedard, D. L.; Brennan, M. J.; Carnahan, J. C.; Feng, H.; Wagner, R. E. Polychlorinated Biphenyl Dechlorination in Aquatic Sediments. *Science* **1987**, *236*, 709–712.

24. Kjeller, L.-O.; Rappe, C. Time Trends in Levels, Patterns, and Profiles for Polychlorinated Dibenzo-p-dioxins, Dibenzofurans, and Biphenyls in a Sediment Core from the Baltic Proper. *Environ. Sci. Technol.* **1995**, *29*, 346–355.

25. Schwarzenbach, R. P.; Gschwend, P. M.; Imboden, D. M. *Environmental organic chemistry*; John Wiley & Sons: Hoboken, N.J., 2003.

26. Armitage, J. M.; Gobas, F. A. P. C. A terrestrial food-chain bioaccumulation model for POPs. *Environmental Science and Technology* **2007**, *41*, 4019–4025.

27. Bernes, C. *Persistent Organic Pollutants: A Swedish View of an International Problem*; Swedish Environmental Protection Agency, 1998.

28. CEC - Publications: Continental Pollutant Pathways: An Agenda for Cooperation to Address Long-Range Transport of Air Pol http://www.cec.org/Page.asp?PageID

=30101&ContentID=16645&SiteNodeID=477 (accessed Apr 21, 2013).

29. Tanabe, S. PCB problems in the future: Foresight from current knowledge. *Environmental Pollution* **1988**, *50*, 5–28.

30. Bremle, G.; Larsson, P. Long-Term Variations of PCB in the Water of a River in Relation to Precipitation and Internal Sources. *Environ. Sci. Technol.* **1997**, *31*, 3232–3237.

31. Isosaari, P.; Kankaanpää, H.; Mattila, J.; Kiviranta, H.; Verta, M.; Salo, S.; Vartiainen, T. Spatial Distribution and Temporal Accumulation of Polychlorinated Dibenzo-p-dioxins, Dibenzofurans, and Biphenyls in the Gulf of Finland. *Environ. Sci. Technol.* **2002**, *36*, 2560–2565.

32. Salo, S.; Verta, M.; Malve, O.; Korhonen, M.; Lehtoranta, J.; Kiviranta, H.; Isosaari, P.; Ruokojärvi, P.; Koistinen, J.; Vartiainen, T. Contamination of River Kymijoki sediments with polychlorinated dibenzo-p-dioxins, dibenzofurans and mercury and their transport to the Gulf of Finland in the Baltic Sea. *Chemosphere* **2008**, *73*, 1675–1683.

33. Oberg, T. Halogenated aromatics from steel production: results of a pilot-scale investigation. *Chemosphere* **2004**, *56*, 441–448.

34. Weber, R.; Tysklind, M.; Gaus, C. Dioxin - contemporary and future challenges of historical legacies. *Environmental Science and Pollution Research* **2008**, *15*, 96–100.

35. Weber, R.; Gaus, C.; Tysklind, M.; Johnston, P.; Forter, M.; Hollert, H.; Heinisch, E.; Holoubek, I.; Lloyd-Smith, M.; Masunaga, S.; Moccarelli, P.; Santillo, D.; Seike, N.; Symons, R.; Torres, J. P. M.; Verta, M.; Varbelow, G.; Vijgen, J.; Watson, A.; Costner, P.; Woelz, J.; Wycisk, P.; Zennegg, M. Dioxin- and POP-contaminated sites--contemporary and future relevance and challenges: overview on background, aims and scope of the series. *Environ Sci Pollut Res Int* **2008**, *15*, 363–393.

36. European Commission, Environment Directorate General *REACH in brief*; 2007.

37. Pease, W. *Toxic ignorance: the continuing absence of basic health testing for top-selling chemicals in the...*; Diane Pub Co: [S.l.], 1997.

38. Toxicity Testing: Strategies to Determine Needs and Priorities http://www.nap.edu/openbook.php?isbn=0309034337 (accessed Apr 21, 2013).

39. REACH - Environment - European Commission http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed Apr 21, 2013).

40. Data | The World Bank http://data.worldbank.org/ (accessed Apr 21, 2013).

41. ECHA European Chemicals Agency http://echa.europa.eu/ (accessed Apr 21, 2013).

42. Ap, W.; A, B.; A, G.; Ti, N.; G, P.; M, P.; I, T.; M, V.; Ap, W.; A, B.; A, G.; Ti, N.; G, P.; M, P.; I, T.; M, V. *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*.

43. OECD Quantitative Structure-Activity Relationships Project [(Q)SARs] http://www.oecd.org/env/ehs/risk-assessment/oecdquantitativestructure-activityrelationshipsprojectqsars.htm (accessed Apr 21, 2013).

44. OECD *Guidance Document on the Validation of (Quantitative) Structure Activity Relationship (Q)SAR Models.*; OECD Environment Health and Safety Publications. Series on Testing and Assessment No. 69.; Organisation for Economic Cooperation and Development: Paris, France., 2007.

Part II: Tools and Methods

1. Introduction

Computer-based tools are increasingly employed in most fields of scientific research. The use of computer technologies to process chemical data resulted in the relatively new discipline called Chemoinformatics, which combines the use of theoretical chemistry and mathematical algorithms. In the fields of environmental and life sciences, Chemoinformatics represents a link between chemistry and biology. QSAR modeling is an important tool in Chemoinformatics and it exploits this theoretical connection. In fact, the investigation of the structure-activity relationships (SARs) is mainly based on the premise that biological activity (or property in the case of QSPR) of a given chemical can be predicted from its molecular structure since it depends mainly on its intrinsic nature. The conceptual basis of QSARs is the congenericity principle which states that compounds with similar structures are assumed to be associated with similar properties. Thus, the biological activity of chemicals can be inferred from the properties of the compounds with known experimental responses. This explains the relevance of the computational predictive models that can be used to fill the lack of knowledge on chemicals for scientific as well as regulatory purposes.

However, QSAR models should first demonstrate high predictive ability in order to be useful for regulatory applications. For this reason, general guidelines of good practice have been published in the literature [1]. In addition, REACH requires a set of 4 conditions in alignment with the OECD principles to be fulfilled for QSAR modeling [2]:

- the model is scientifically valid;
- the model is applicable to the chemical of interest;
- the prediction is relevant for the regulatory purpose; and
- the method and results are appropriately documented.

This chapter explains the conceptual basis of QSAR/QSPR as well as the methodologies used in this thesis, from data acquisition and preparation, through calculation of molecular descriptors, application of appropriate machine learning methods till the model validation and the assessment of its domain of applicability.

2. Data acquisition and curing

The development of a predictive QSAR model is a process of several steps. Initially, the gathering and screening of experimental data is required. This step is fundamental to providing reliable data for subsequent QSAR models. Therefore, it is one of the most important steps of the analysis, since all the results will depend on data quality.

2.1. Data sources

Collection of experimental data requires a deep investigation in the scientific literature to extract the appropriate data from reliable sources. Moreover, QSAR models should be based on datasets that present good coverage of a wide range of the chemical space. Unfortunately, a single published experimental study does not always present a sufficient amount of data needed for QSAR analysis. It also occurs that the experimental conditions and/or the used test protocol are not explicitly available. This condition can be misleading especially for specific and similar endpoints such as BioConcentration Factor (BCF) and BioAccumulation Factor (BAF), which differ only by the ways of uptake. Thus, merging experimental data from different sources for modeling purposes could be a time demanding process.

However, data collection can be facilitated by the use of experimental data collected in publicly available databases. There are several online databases which store information on chemical compounds including physicochemical properties, toxicological/eco-toxicological and environmental fate endpoints. Examples of these databases are ChemSpider [3], PubChem [4,5], ChemExper [6]. These databases have useful searching options, such as chemical name, CAS-RN (Chemical Abstract Registration Number) [7,8], PubMed ID [9] and/or structure representations such as SMILES and INCHI codes [10].

In addition to the information about chemicals, other online sources provide also access to modeling tools designed for QSAR, such as VCCLAB [11], OCHEM (Online Chemical Modeling Environment) [12], OpenTox [13], QSARdb [14], SPARC [15] and PBT profiler [16], inter alia.

Moreover, some QSAR modeling software allow access to their databases. One example is the OECD QSAR toolbox, a huge database of referenced entries accessible through a user-friendly interface enabling a rich list of features such as multi search options for 2D structures, a large number of physico-chemical properties and endpoints for a wide range of chemicals [17]. Another relevant data source for QSAR is the online freely available database of the United States Environmental Protection Agency (US-EPA) [18]. The datasets used to build the physicochemical and environmental fate models implemented in EPI (Estimation Program Interface) Suite are available online [19]. It can also store QSAR models and provide literature references.

2.2. Data curing

The online QSAR datasets and those included in the software databases may contain different types of errors. One of the commonly encountered errors is the presence of duplicates of molecules. Duplicates can be perfect copies, and in this case the error can be solved by keeping only one of the database entries. However, in most of the cases, it is not easy to deal with duplicates. This usually happens in merged datasets from different sources and/or experimental conditions, which can give different results for the same compound. Nevertheless, it also occurs that different entries can be merged resulting in "false" duplicates when compounds have the same identifier but different structures and vice versa. This problem can be avoided by using more than one identifier (e. g. CAS-RN, INCHI, chemical name, molecular formula) in addition to the internal identifier of the database. Matching all of these identifiers during queries and making them available with the published QSAR model can remove ambiguity for the users.

Another source of errors in the databases is related to the structure representations. This type of errors can highly affect the quality of the model since the chemical structures are used to calculate the molecular descriptors. Storing the structures in two-dimensional (2D) format rather than 3D can facilitate their use and the database management as well as the subsequent modeling steps. The commonly used 2D formats are SMILES (simplified molecular-input line-entry system) [10], or unique SMILES [20].

However, several errors in the SMILES notations can be faced during the structures checks [21,22]. The most common are related to stereochemistry, valence and charge.

Other ambiguities could occur when experimental results are reported in different units. Thus, all values should be converted to the appropriate unit before merging them and proceeding with the modeling step. As an example, several endpoints should be given in molar units rather than weight or concentrations. This can be explained by the fact that biological activity usually depends on the number of present molecules and not on their weight [1].

Since the comprehensive assessment of QSAR data requires checks for errors and self consistency, dealing with it manually is a hard task especially in the case of huge databases.

Several Chemoinformatic tools and data-mining software are available to eradicate the inconsistency of experimental data. The main tools employed in this work were ChemBioFinder and KNIME.

2.1.1. ChemBioFinder

A complete set of tools for database management is available in ChemBioFinder software (CambridgSoft) [23]. It allows storing of chemical information including identifiers, physicochemical properties, notes, tables of data and charts. The data can be imported and exported easily in different formats. The obtained database is searchable by querying a multitude of field combinations. The searching methods can be based on text, numbers, full structures or sub-structures for an exact match, similarity or tautomerism specifying the desired stereochemistry. This chemical database manager performs also searches for duplicates, errors and other special searches.

This tool is part of the ChemBioOffice software that is a modeling suite for chemists and biologists [24]. It performs structure activity relationships calculations, clustering, statistics, physicochemical and bioavailability properties predictions, viewing and editing the small molecules and peptide structures in addition to database management.

This software suite was used during this project (under a license provided by the University of Strasbourg) to analyze a big dataset of compounds for log P prediction.

2.1.2. KNIME

Another powerful tool extensively used during this work is the data-mining software KNIME (Konstanz Information Miner) [25]. It is a user-friendly graphical workbench for the entire data analysis process starting from the initial data access, transformation and investigation until the predicting analytics, visualization and reporting steps. Over 1000 modules, called nodes, are provided by its open integration platform including the contribution of the users' community and partner network. The desktop version of KNIME is a free and open-source, released under the GNU General Public License (GPL) [26].

Once KNIME has been started, the installed extensions such as WEKA, R and MATLAB integrations and other additional nodes for data analysis are loaded and initialized. Then, the workbench is opened showing the platform of the tools for data-mining. It is intuitively organized in different sections and mainly consists of the workflow editor, the node repository and the node description.

To build a new workflow, the nodes are dragged from the node repository to the workflow editor. The selected nodes are, then connected according to the desired order through their input/output ports and configured to perform the needed tasks. In the end the workflow is executed, following the right order of the nodes or in parallel if possible.

The repository contains all the installed nodes organized in categories and subcategories. By default, KNIME offers different features of preinstalled nodes for Chemoinformatics as well as other fields. It has nodes for integrated scripting languages (Perl, Python, R, MATLAB) and packages of basic input/output and advanced data processing operations.

KNIME workflows can interact with any software installed on the computer by using the "External tool" node. To interact with online sources, KNIME has the "Generic Web-service Client" node. During this work, this tool was particularly useful for retrieving and/or checking the chemical structures from online databases that provide SOAP web-services. ChemSpider database gives free access for academic users to its APIs services for searching and retrieving chemical information through automated workflows such as KNIME or Pipeline Pilot [27]. OCHEM also offers several API services for uploading data as well as creating and applying QSAR models [28]. The newly developed node named CIR (Chemical Identifier Resolver) have been used in order to exploit CACTUS the online service of the NCI/NIH for checking chemical structures and converting different formats [29,30].

There is a wide range of nodes developed by the users' community and KNIME partners. These packages are continuously improved and updated while new ones are being released with every version. In the field of

Chemoinformatics, there are several useful tools that have been included in the node repository, such as: ChemAxon tools, the Chemistry Development Kit CDK, PaDel and many others that allow performing all steps of data gathering and curing as well as modeling and predicting of new chemicals. The developers of KNIME have recently published a book entitled "Guide to Intelligent Data Analysis" to explain many data-mining techniques giving examples of how it can be applied using KNIME workflows [31].

3. Molecular descriptors

3.1. Introduction

Structure–activity relationships (SARs) are theoretical models relating structural features of chemicals to their experimental activity/property. These models are used in order to predict physicochemical, biological or fate properties of a given molecule on the basis of its chemical structure.

The complexity of a molecular structure is due to the fact that most of its properties cannot be derived from the summation of the properties of its single atoms [32]. Hence, it is a holistic system that depends on the atomic connections and interactions. Consequently, a molecular structure has not a unique representation but several possible models depending on the theoretical approach adopted and the degree of approximation.

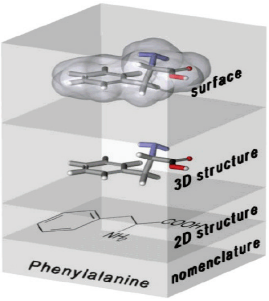


Figure 1: *different levels of structural representation.*

As shown in Figure 1, different "symbolic" representations for the same molecule are possible. It can vary from the simple nomenclature or molecular formula to the 2D representation based on the graph theory and the more complex 3D conformations [33,34]. However, these representations, offering different aspects of the chemical information, are usually not derivable from each other.

These different levels of representations are used by scientific researchers to retrieve the corresponding theoretical information encoded in the molecular structure in order to establish the desired relationships between the studied structures and the experimentally demonstrated properties. This information is converted to a significant number called molecular descriptor.

By definition: "*The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number or the result of some standardized experiment*" [32].

For the key role they are playing in many fields of scientific research, a special interest is given to the development of molecular descriptors. Thousands of descriptors have been proposed in the literature. Their list is being continuously updated and their number increasing with the complexity of the investigated chemical systems. This is enhanced by the fast increase of the computational speed enabling the rapid calculation of molecular orbital and quantum mechanical descriptors such as charges, dipole moments and energy levels.

Molecular descriptors are required to encode the hydrophobic, electronic and steric aspects of a molecule in order to be able to describe the biological activity of a chemical in a living organism.

As for structural representations, molecular descriptors are classified in five dimensions equivalent to different levels of "complexity" according to the encoded chemical information:

- The 0D corresponds to the molecular formula. At this level, the retrieved information is independent from any structural representation and can be referred to as weighting schemes, atom type counters or constitutional indices. The UIPAC International chemical Identifier (InChI) is also used as a descriptor to predict properties of chemicals [35].

- In the 1D class, only partial knowledge of the structure concerning functional groups and fragments is needed. Such groups of adjacently connected atoms in a molecule are typically used in substructural analysis. The presence of biological activity related to a substructure is called structural alert [36].
- The 2D class of descriptors is based on graph theory. These descriptors are mainly topological and connectivity indices. Recently, the 2D molecular representations, such as SMILES, were also used as descriptors for QSPR models [37].
- The 3D descriptors are derived from the geometrical representations of the molecules and they encode information about the size and shape of a studied conformation of the molecule.
- Finally, the 4D descriptors take into consideration the flexibility aspect of the 3D structural representation of the molecule used in 4D- or Dynamic-QSAR. This class of descriptors also includes the stereo-electronic representations characterizing the electronic interactions of a molecule with its surrounding environment. This concept is the basis of the grid-based QSAR techniques such as the Comparative Molecular Field Analysis (CoMFA) [38–40].

A comprehensive review of molecular descriptors has been published by Todeschini and Consonni [32].

Since the models developed in this research work were aimed to be used in regulatory purposes within the new European legislation on chemicals (REACH), care has been taken in the choice of molecular descriptors to be included in the models. Only interpretable and reproducible descriptors have been considered. Thus, descriptors based on 3D representations were excluded in order to avoid the irreproducible geometrical optimization of molecular conformers.

3.2. Analysis of new molecular descriptors

In this work, in addition to the classical molecular descriptors a set of new descriptors has been evaluated. In particular, the recently developed spectral indices, derived from different graph matrices, have been analyzed for the first time and used later in the QSAR models [41]. Moreover, this analysis focused on some other topological descriptors which have never been used to model environmental endpoints and other string representations which are relatively new descriptors for QSAR modeling, being only used in database searching.

3.2.1. Spectral indices

Spectral indices are molecular descriptors based on the eigenvalues of graph theoretical matrices. Since they can be derived from any graph-theoretical molecular matrix, there is a large number of combinatorial possibilities of these indices [32,42,43]. Besides the adjacency (**A**), Laplacian (**L**), Barysz (**Dz**) and Burden (**B**) matrices, some other matrices to derive spectral indices are the distance-path matrix, Szeged matrix, distance valency matrices, geometry matrix, resistance distance matrix and conductance matrix [42,44–47]. However, not all of the combinations that can be derived from such matrices have already been evaluated and used as molecular descriptors for QSAR/QSPR studies.

Using a molecular matrix $\mathbf{M}(A \times A)$ with a weighting scheme w, the most commonly used indices are calculated as following:

$$SpAbs(\mathbf{M}, w) = \sum_{i=1}^{A} |\lambda_i|$$

$$SpPos(\mathbf{M}, w) = \sum_{i=1}^{A^+} (\lambda_i^+)$$

$$SpMax(\mathbf{M}, w) = max_i\{\lambda_i\}$$

$$SpMaxA(\mathbf{M}, w) = max_i\{|\lambda_i|\}$$

where λ_i are the eigenvalues of the matrix or spectrum.

$SpAbs$ is the sum of the A absolute eigenvalues of the molecular matrix. When derived from the adjacency matrix, this entity is called the graph energy (E) [48–50]. It is also called the Laplacian graph energy when it's calculated from the Laplacian matrix [51,52]. $SpPos$ is the sum of the A positive eigenvalues of the weighted matrix. $SpMax$ is the leading eigenvalue of the spectrum corresponding to the Lovasz-Pelikan index when it's derived from the adjacency matrix [53]. $SpMaxA$ is the maximum absolute value of the spectrum [32].

The spectral moments are a similar class of molecular descriptors. Applied on the weighted graph-theoretical matrix $(\mathbf{M}, w)$, the spectral moments are defined in terms of the *k*th power of eigenvalues [32]. These descriptors are calculated as following:

$$\mu^k(\mathbf{M}, w) = \sum_{i=1}^{n} \lambda_i^+$$

where $k = 1, \dots, n$ define the order of the spectral moment.

The spectral moments were extensively used by E. Estrada in the QSAR/QSPR studies [54–57].

Although being largely investigated, due to their large number, spectral indices and spectral moments have not been fully investigated tested and used in the literature of QSAR modeling. In this work, some of these descriptors have been successfully included in the QSAR models for predicting biodegradability of chemicals [58].

Two new families of spectral indices have been recently developed and published in the literature [41]. These indices are calculated on the same basis as the previously defined spectral indices, using any graph-theoretical matrix $\mathbf{M}(w)$, its eigenvalues λ_i and their average $\bar{\lambda}$.

The sum of absolute deviations from the average eigenvalue:

$$SpAD(\mathbf{M}, w) = \sum_{i=1}^{n} |\lambda_i - \bar{\lambda}|$$

The mean absolute deviation which is size independent:

$$SpMAD(\mathbf{M}, w) = \frac{\sum_{i=1}^{n} |\lambda_i - \bar{\lambda}|}{n}$$

Tested in some univariate models, these indices showed interesting properties and modeling ability [64]. In this work, $SpMAD$ indices have been used to model the bioaccumulation of polybrominated diphenyl ethers in aquatic species [59].

These descriptors have several useful features for QSAR/QSPR studies. Even though these indices are extracted from relatively complicated matrices, their decomposition and interpretation could lead to some relevant correlation that describes the physicochemical and/or biological properties of the investigated molecular structures [54]. The contribution of such descriptors to the studied properties can be described by means of known properties such as molecular mass, branching or steric features of the structures [60]. In addition to QSAR analysis, these descriptors can also be useful in similarity/dissimilarity studies of chemicals [54].

3.2.2. Matrix-based descriptors

Matrix-based descriptors are topological indices calculated in two steps. First, the information encoded in the H-depleted molecular graphs of chemicals was encoded into the graph-theoretical matrices. Then, quantitative indices were obtained by applying a set of basic algebraic operations to the graph-theoretical matrices [32]. All the calculations were performed by the software DRAGON [61].

The topological indices are molecular descriptors derived from the molecular graph. They numerically quantify the molecular topology independently from the vertex numbering or labeling. These indices are able to encode the structural features of the molecules such as shape, size, cyclicity, molecular branching and atom types [62,63]. One example of the most used topological indices is the connectivity indices. These latter ones are derived from the H-depleted where each vertex is weighted by the vertex degree [64].

The adjacency matrix (**A**), also called vertex adjacency matrix, is one of the fundamental graph-theoretical matrices. It encodes the connections between the adjacent pairs of atoms [65]. This matrix is an important source for molecular descriptors calculation since different other useful matrices, such as Laplacian (**L**), Barysz (**Dz**) and Burden (**B**), are derived from it [32]. The latter matrices are used to calculate the different 2D matrix-based descriptors considered in this study.

Laplace matrix **L** is given by the difference between a diagonal vertex degree matrix and the adjacency matrix **A**:

$$[\mathbf{L}]_{ij} = \begin{cases} -1 & \text{if } (i,j) \in \mathrm{E(G)} \\ \delta_i & \text{if } i = j \\ 0 & \text{if } (i,j) \notin \mathrm{E(G)} \end{cases}$$

where *δi* is the *i*-th vertex degree, that is, the number of vertices adjacent to vertex *i* and E(G) is the set of graph edges.

Burden matrices $\mathbf{B}(w)$ are augmented adjacency matrices defined to account for heteroatoms and bond multiplicity calculated as the following:

$$[\mathbf{B}(w)]_{ij} = \begin{cases} \sqrt{\pi_{ij}^*} & \text{if } (i,j) \in \mathrm{E(G)} \\ \dfrac{w_i}{w_\mathrm{C}} & \text{if } i = j \\ 0.001 & \text{if } (i,j) \notin \mathrm{E(G)} \end{cases}$$

The diagonal elements are atomic carbon-scaled properties such as the mass (m) and the polarizability (p). The off-diagonal elements corresponding to pairs of bonded atoms are the square roots of conventional bond orders π* (i.e., 1, 2 , 3, and 1.5 for single, double, triple and aromatic bonds, respectively). The remaining matrix elements are set at 0.001 by default.

Barysz matrices $\mathbf{Dz}(w)$ are weighted distance matrices obtained by generalizing the Barysz weighting scheme in terms of conventional bond orders π* and any atomic property [66]:

$$[\mathbf{Dz}(w)]_{ij} = \begin{cases} d_{ij}(w,\pi^*) & \text{if } i \neq j \\ 1 - \frac{w_\mathrm{C}}{w_i} & \text{if } i = j \end{cases} \qquad d_{ij}(w,\pi^*)$$

$$= \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \cdot \frac{w_\mathrm{C}^2}{w_{b(1)} \cdot w_{b(2)}} \right)$$

where w_C is any atomic property, such as Sanderson electronegativity (e), of the carbon atom and w_i the corresponding value of the i-th atom. $d_{ij}(w,\pi^*)$ is a weighted topological distance that is the sum of the edge weights over all bonds involved in the shortest path between vertices v_i and v_j. The subscripts $b(1)$ and $b(2)$ are representing the two vertices incident to the considered b-th edge.

The hyper-Wiener-type indices ($HyWi$) and the Balaban-like indices (J) are two examples of the topological indices that can be derived from the previously described matrices ($\mathbf{B}(w)$ and $\mathbf{Dz}(w)$) [67,68]. Variances of theses indices calculated using the mass (m) and electronegativity (e) as weighting schemes have shown interesting modeling properties [58].

The $HyWi$ indices, also called hyper-Wiener operator, are calculated by analogy to the hyper-Wiener index (WW) derived from the Wiener matrix by taking into consideration also the diagonal elements of the weighted matrix $\mathbf{M}(w)$ [32,69].

The general formula for calculating the hyper-Wiener-type index is the following [67]:

$$HyWi(\mathbf{M}; w) = \frac{1}{2} \cdot \sum_{i=1}^{A} \sum_{j=i}^{A} \left([\mathbf{M}(w)]_{ij}^2 + [\mathbf{M}(w)]_{ij} \right)$$

where A is the number of graph vertices and $\mathbf{M}(w)$ is a graph-theoretical matrix calculated using the weighting scheme w.

While the original Wiener index (*W)*, which is one of the first molecular descriptors, is obtained by summing the lengths of the shortest paths in the graph [70]. It was the first descriptor proposed for molecular branching [71].

The Balaban-like indices are similar to the Balaban distance connectivity index which is a graph invariant molecular descriptor independent from the molecular size or number of rings [72–74]. They are also calculated in a similar way. However, in the Balaban-like index the vertex distance degrees are substituted by the row sums of the considered graph-theoretical matrix [32].

The Balaban-like index general formula is given by [68]:

$$J(\mathbf{M};w) = \frac{\mathrm{B}}{\mathrm{C}+1} . \sum_{\mathrm{i}=1}^{A-1} \sum_{\mathrm{j}=\mathrm{i}+1}^{A} a_{ij} . \left[VS_i(\mathbf{M};w) . VS_j(\mathbf{M};w)\right]^{-1/2}$$

where A, B and C are the number of vertices, edges and rings, respectively. $\mathbf{M}$ is the graph-theoretical matrix calculated using the weighting scheme w. a_{ij} the elements of the adjacency matrix and VS is the vertex sum operator applied to the matrix $\mathbf{M}$.

3.2.3. Vectorial descriptors

The vectorial descriptors are a special class of molecular descriptors, initially developed to perform queries in big databases for similarity searching [75,76]. Recently, these bit-strings started to be used as descriptors for QSAR modeling [77–80]. Since they usually consist of fixed lengths of strings mostly varying from hundreds to thousands of bits to enclose the most of the needed information, the variable selection step is always skipped.

This class of descriptors can be categorized into two groups: structural keys and fingerprints. Starting from a set of predefined structural features, the structural keys can be binary vectors specifying the presence and absence by 1 and 0, respectively, or can be counts of the selected functional groups, augmented atoms, atom pairs, atom-type electro-topological states (E-states), pharmacophore points, etc [81,82]. Fingerprints, in the other hand, are Boolean vectors defining a set of patterns and generated, by means of hashing algorithms, in a way to capture the common chemical features present in a data set [83]. Whereas structural keys present a straightforward correspondence between bin and fragments, hashed fingerprints may encode several fragments into a single bin according to the used string hashing algorithm [10].

Following the general classification pattern for molecular descriptors, these string representations of chemical structures are categorized in 2D, 3D and 4D accordingly [79,80,84–87].

In this work, only structural keys have been tested for QSAR modeling towards the endpoints of interest for REACH. These fragmental bit-strings have been already used in the literature to model biodegradability of chemicals [77].

Several types of structural keys have been presented in the literature. Their string lengths can vary depending on the amount of information encoded. The predefined dictionary of fragments used in indexing the chemical structures usually consists of small groups of atoms, functional groups or rings.

Examples of commonly used 2D structural keys implemented in specific automated tools are MACCS and PubChem keys.

MACCS keys, the Molecular ACCess System descriptors, are created by Molecular Design Limited [88]. They are 2D substructural descriptor encoding atoms types, rings and bond information. Originally, it was generated in a 960 key-bits format and later a subset of 166 key-bits was extracted [89].

The PubChem binary substructure keys are developed to be used by PubChem database in order to perform the searching queries [5]. The length of this string is 881 bits, with a four-byte prefix, the size of this descriptor is therefore 115 bytes. The PubChem bit-string is divided in 7 sections of SMILES or SMARTS (SMiles ARbitrary Target Specification) notations [10]. These sections encode hierarchic atom-type counts, rings, atom pairs, atom nearest neighbours, atom connections, simple and complex SMARTS patterns [90].

3.3. Software for descriptor calculation

Several tools for descriptor calculation have been used along this thesis. Owing to the wide variety of packages available, only software used during this work are presented.

3.3.1. DRAGON

Thanks to its large number of descriptors, DRAGON software is one of the most widely used tools for molecular descriptors calculation [61]. It was the main tool of molecular descriptors calculations used in this work. It calculates almost 5000 molecular descriptors [91]. To facilitate the calculation task for users, the descriptors are categorized in 29 logical blocks of known groups such as constitutional indices, topological indices, geometrical descriptors, 2D and 3D atom pairs, functional groups and atom-type E-states. In addition, the calculation of several important molecular properties such as logP, topological polar surfaces, Van der Waals surfaces as well as some drug-like indices such as Lipinski's rule of 5 is also provided. These properties and many others are also available in the related application dProperties [92]. These two packages support all the commonly used molecular formats and perform a preliminary check for the structures, i.e., erroneous and disconnected structures are usually rejected. DRAGON calculations can be performed from its intuitive and user-friendly interface or in batch mode by command line. Recently, DRAGON can also be executed in batch mode from a KNIME workflow using its dedicated node. In addition to molecular descriptor calculation, this software allows performing a preliminary analysis of the calculated descriptors prior to the modeling stage. Pair-wise correlations, Principal Component Analysis (PCA), graphical analysis and import of external variables are other facilities provided by DRAGON.

3.3.2. SubMat

SubMat is a commercial software developed by the Chemometrics group of the Wien University of Technology [93]. It allows the generation of binary substructure descriptors from a user-provided list of predefined substructures checking for their presence/absence. The input files of both molecular structures and fragments dictionary must be in Molfile format [88]. The substructure searching method is based on the complete atom-atom and bond-bond matching [94,95]. The developers of the software have also provided a list of 1365 substructures covering a wide range of fragments based on mass-spectrometry fragmentation [96]. The maximum molecule size allowed is 127 atoms explicitly defined and 255 bonds per structure.

3.3.3. The Chemistry Development Kit

The Chemistry Development Kit (CDK) is an open-source Java library for structural Chemoinformatics and Bioinformatics [97]. It is available under the terms of the GNU Lesser General Public License (LGPL) [98]. Thus it is freely available for use and modification by academic and industrial institutions and may be integrated in proprietary packages [99]. Subsequently, its libraries started to be a basis for several software projects [97]. The development of the tool-kit is involving an international team of collaborators to maintain and update its packages providing a rich list of molecular modeling methods including structural rendering, searching, parsing and generation of chemical structures. In the recent versions of the software, the library became more Chemoinformatics oriented by adding packages for 2D and 3D molecular descriptor calculations as well as QSAR modeling tools [100].

A dedicated graphical user interface was designed for the molecular descriptor calculations [101]. The CDK Descriptor Calculator GUI is divided in two sections. One is providing a list of 6 blocks of descriptors such as the topological, constitutional and geometrical descriptors [102]. The second section is dedicated to the substructure keys including MACCS, PubChem and E-state keys, as well as a hashed fingerprint of 1024 bits based on the Daylight theory [10,97]. The CDK Cheminformatics tool-kit is also available as package of several nodes for KNIME.

3.3.4. PaDEL

PaDEL is a useful software for calculating molecular descriptors and fingerprints [103]. It provides 863 descriptors which are categorized in 729 1D-2D descriptors and 134 3D descriptors, in addition to 10 types of vectorial descriptors consisting of sub-structural keys and fingerprints. The software is mainly based on the CDK tool-kit, however, additional descriptors were implemented by the developers. These descriptors include E-state indices, logP, energy relation descriptors, ring descriptors as well as Laggner's and Klekota-Roth molecular substructures [104–106]. Developed in Java programming language, PaDEL has the possibility to be easily integrated into other software (e.g. for QSAR modeling), called by command line or used as a standalone application GUI. Nodes for KNIME are also developed and available for free download as well as the source classes of the software [107].

4. Variable selection techniques

Though only one tool of molecular descriptor calculation is used and not all available types of descriptors are considered, the initially calculated descriptors can reach several hundreds or thousands. Certainly, such a large pool of descriptors will enclose not only feature rich but also redundant and irrelevant information for the subsequent QSAR modeling. However, a good QSAR model should be parsimonious, that is, including a set of variables which is information rich but as small as possible in order to avoid overfitting and allow the model interpretation. Hence, it is important to reduce the initial number of calculated descriptors before the modeling step.

The first step of feature selection is usually a filtering step. It consists of the removal of highly correlated, constant and near constant descriptors. The methods that can be applied at this stage are unsupervised since the studied experimental response is not included in the analysis of variables.

In DRAGON, this step can be carried out before exporting the calculated descriptors. Pair-wise correlation coefficients are calculated for all the descriptors. If a pair of descriptors has a linear correlation coefficient larger than a defined threshold the descriptor showing the largest average correlation with all others is discarded.

Once the initial pool of descriptors has been reduced by means of initial filters, the suitable subset to build the QSAR model for the studied activity/property must be selected. Hence, feature selection methods coupled with the desired regression or classification algorithms can be applied. Several algorithms for variable selection have been proposed in literature. Most common examples are Genetic Algorithms (GAs) [108–110], stepwise forward/backward selection [111], particle swarms [112], simulated annealing and ant colony algorithms [113,114]. In this work, GAs and forward selection were considered.

4.1. Stepwise forward selection

Forward variable selection is one of the most simple and fast selection techniques. Starting from a first descriptor and adding the remaining descriptors one by one, it evaluates the performance of the model by optimizing a fitness function [111]. The fitness function is chosen according to the type of the modeled response that can be continuous for regression models or categorical in the case of classification models. Thus, it could be for example the error rate in classification or the sum of squared residuals in regression. The results of this method are highly depending on the first included variables and the information included in the initial pool of descriptors cannot be completely explored. Consequently, the final selected descriptors are not necessarily the best representative descriptors of the original set.

4.2. Genetic Algorithms (GAs)

Genetic Algorithms (GAs) are one of the nature-inspired evolutionary algorithms. It is based on the biological concept of evolution to optimize the searching methods [115]. GAs are widely used in the fields of Chemometrics and Chemoinformatics [110,116,117].

In QSAR modeling, these algorithms are applied on the multivariate descriptor space in order to find the optimal subsets of descriptors. The evolution process is carried out by maximizing the predictive ability of the models measured by a fitness function [108,109].

The used terminology is adopted from the field of biological evolution. Thus, a population is an ensemble of individuals consisting of a chromosome and its associated fitness value. A chromosome is defined as Boolean vector describing the presence/absence of genes that represent the subset of selected variables. Each chromosome corresponds to a model with a certain predictive ability.

The evolution process is performed in several steps. First, the initial population is randomly created. The number of initial chromosomes as well as their size are user defined, a priori. The models are, then, built and ordered according to their predicting ability. The fitness function depends on the nature of the endpoint being modeled. The different predictive and fitting measure methods are explained in Section II.6.

The following is the reproduction step aiming to create the child population. Starting from the parents that are pairs of individuals randomly selected, the son chromosome is generated using the same genes of the parents by applying the two-fold genetic operations. A newly created individual is evaluated and ranked if it is unique in the current population, otherwise, it is automatically rejected. If its rank is better than at least one of the existing, the created child is a new member of the population excluding the worst one to keep the size constant.

Crossover is a genetic operation that consists of swapping portions of the chromosomes of the parents. A variety of crossover ways have been described in the literature [108]. One of the possible implementations is to restrict the cutting operation to a single point. Then the two new chromosomes are created by exchanging the descriptors from one side of the split. The intent of the cross over is to generate better models than those in the initial population by preserving the best portions of the starting chromosomes.

The second operation is the mutation which is performed on a single chromosome. In order to mirror its low frequency in natural biological evolutions, mutation is restricted to a low user defined probability. It consists of randomly changing one of the descriptors of a given chromosome by another one from the pool aiming to explore the maximum of the descriptors space and to avoid “premature” convergence by getting stuck in a local solution and miss the optimal one.

These two operations are repeated creating generations of populations that are evaluated and ranked during the evolution process that takes a user defined number of cycles. At the end, the top ranked models are reported to the user who can decide about the best results based on different parameters and not only the used fitness criteria.

The GAs used to perform the variable selection operations in the current study were inspired by the approach of Leardi *et al.* and implemented in MATLAB environment [109,110,118].

5. Modeling methods in QSAR

QSAR and QSPR are based on the observations that a change in the physicochemical properties of molecules can be induced by varying the chemical structures. QSARs started to have their concrete beginning with the works of Hansch and Free-Wilson in the early sixties of the last century [119,120]. Since then, the arsenal of modeling methods applied to QSAR studies have been broadened by adding several multivariate chemometric methods which have been continuously refined during the last decades.

QSAR's general mathematical form is:

$$Activity = f(physicochemical\ and/or\ structural\ properties)$$

Thus, the development of a QSAR model requires three key components. The first two ones, described in the previous sections, are:

- experimental data acquisition and curing
- description of the physicochemical properties and/or chemical structures by a set of molecular descriptors.

The third one is the core of QSAR modeling and it consists of a theoretical function based on mathematical and statistical methods to find the required relationship linking the molecular properties to their structural descriptors.

A multitude of prominent chemometric methods are used in QSAR studies. Methods considered in this work were:

- exploratory data analysis methods such as Principal Component Analysis (PCA) and the Multi-Dimensional Scaling (MDS);
- regression methods including Multiple Linear Regression (MLR) and Partial Least Squares (PLS);
- classification methods such as k^{th} Nearest Neighbors (kNN), Support Vector Machines (SVM) and Partial Least Squares Discriminant Analysis (PLSDA) [121–129].

In this thesis, most of the used techniques were implemented and used within the MATLAB environment.

5.1. Unsupervised methods for exploratory data analysis

Unsupervised learning methods are used in descriptor data analysis for pattern recognition without making use of the experimental response.

5.1.1. Principal Component Analysis (PCA)

Most of the chemical applications require multivariate data analysis. Since descriptors hyperspace usually encodes redundant and noisy information, it requires a powerful chemometric method to deal with the collinearity. PCA is one of the widely used tools for reducing dimensionality [130–132]. It is an exploratory technique used to visually estimate the structure of the multivariate data, detect pattern in the data as well as the presence of potential outliers.

PCA adopts a compression technique of the correlated descriptors by projecting them into a new set of variables called Principal Components (PCs). These new orthogonal variables are linear combinations of the original descriptors. Since only few PCs are commonly retained, most of the dataset's variability is enclosed in a lower dimensional space of orthogonal PCs. The first PC defines the direction of the maximum data variance, while the subsequent PCs describe the maximum of the remaining variance in directions which are orthogonal to each others. The redundancy is, therefore, removed and most of the initial information is explained by the first few PCs.

5.1.2. Multi-Dimensional Scaling (MDS)

MDS is a useful method that reconstructs the distribution of the initial hyper-dimensional data into a much lower space on the basis of the distances between the samples [121,122]. Thus the aim of MDS is to let the user to visualize the distances between the samples in order to have an approximate idea about the degree of similarity in the analyzed data. The degree of approximation in the low-dimensional space is explained by the residuals between the original and the new distances separating the samples.

5.2. Supervised learning methods for modeling

Unlike previously mentioned data exploratory methods, supervised learning methods use the experimental response being modeled. Thus, care needs to be taken in order to avoid over-fitting.

The nature of the modeled response is a crucial factor in the choice of the method to be used. There are two types of methods:

- classification methods handling categorical responses such as active/non active, toxic/non toxic or biodegradable/non biodegradable;
- regression methods dealing with continuous responses such as logP and BCF. Nevertheless, some techniques are suitable both for classification and regression tasks.

5.2.1. Regression methods

5.2.1.1. The k Nearest Neighbors in regression

kNN is one of the simplest techniques for modeling. It makes use of the congenericity principle assuming that within a selected descriptors space, the closest compounds will have similar response.

The commonly used metric in kNN modeling is the Euclidean distance. Other metrics such as Manhattan distance and Mahalanobis distance can also be applied [133]. Several methods can be applied to obtain the predicted response for a test sample. In this work, the predictions were processed in two ways:

- by averaging the observed values of the k nearest neighbors
- by weighting the observed values according to the distances of the test sample to the k nearest neighbors.

In this work, k is optimized to get the best performance in cross-validation. The kNN approach often presents good results, however, its predictive ability in regression can be altered in the case of high-dimensional data [134].

5.2.1.2. Multiple linear regression

MLR is a mathematical method used to find a linear relationship between the observed response and a number of independent variables (descriptors) as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \qquad i = 1,2, \ldots , n$$

where y_i is the observed response, $x_{i1}, x_{i2}, \ldots, x_{ip}$ are the independent variables for the ith sample, p is the number of variables, n is the number of samples and ε_i is the error of prediction. By estimating the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ the equation of the linear model is:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip}$$

where $b_0, b_1, b_2, \ldots, b_p$ are the estimates of the previous parameters and $\hat{y}_i$ is the predicted value of the model.

MLR is based on the Orthogonal Least Square (OLS) algorithm that minimizes the sum of squares of the error between the predicted and the observed values $\sum(y - \hat{y})^2$.

The vector of predicted values $\hat{\boldsymbol{y}}$ is obtained as following:

$$\hat{\boldsymbol{y}} = \mathbf{bX}$$

where $\mathbf{b}$ is the vector of estimated parameters $b_0, b_1, b_2, \ldots, b_p$ calculated as:

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'y}$$

where $\mathbf{X}$ and $\mathbf{y}$ are the matrix of descriptors and the vector of experimental responses, respectively.

MLR modeling is based on the assumption that the errors are a normally distributed random variable with constant variance. The obtained model is optimal when the regression estimators are unbiased, efficient, and consistent with a bias and variance approaching zero when the number of samples tends to the infinity.

The disadvantage of this method is that collinearity between the descriptors highly affects the reliability of the regression coefficient estimates. Thus, reducing the number of included variables by removing those with insignificant coefficients can reduce the risk of multi-collinearity and contribute to enhance the reliability of predictions.

5.2.1.3. Partial Least Squares (PLS)

PLS is a powerful statistical method applied in Chemometrics and other fields of scientific research [124]. A major advantage of this method is its ability to overcome the problem of singularity of $(\mathbf{X}'\mathbf{X})$ in MLR due to the number of columns (variables) larger than the number of rows (samples) as well as to the collinearity of variables. This problem is solved by decomposing **X** into orthogonal scores **T** and loadings **P** as follows;

$$\mathbf{X} = \mathbf{TP}$$

Then, **y** is correlated to the first columns of the scores instead of the original variables of **X**. In this way, PLS includes information from both, **X** and

$\mathbf{y}$ in the calculation of the scores and loadings aiming to explain the maximum of variance in the original variables as well as the observed response.

The general decomposition formula of multivariate PLS is:

$$\mathbf{X} = \mathbf{TP}' + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{UQ}' + \mathbf{F}$$

where $\mathbf{T}$ and $\mathbf{U}$ are the matrices of scores of $\mathbf{X}$ and $\mathbf{Y}$, respectively. While $\mathbf{P}$ and $\mathbf{Q}$ are the loading matrices. $\mathbf{E}$ and $\mathbf{F}$ are the matrices of residuals. The aim of this decomposition is to maximize the covariance of $\mathbf{T}$ and $\mathbf{U}$.

There are several implementations of PLS algorithms in the literature giving similar results especially in the case of a single vector response but may differ slightly when dealing with multivariate responses [135,136].

In PLS regression, the components are called Latent Variables (LVs) and are, thereby, incorporating information from the descriptors, the experimental observation as well as the correlation between them. The LVs are calculated by SVD decomposing the cross-product of the variables $\mathbf{S} = \mathbf{X}'\mathbf{y}$.

5.2.2. Classification methods

5.2.2.1. The k Nearest Neighbors (*k*NN)

The kNN approach for classification operates similarly to regression. Assuming that the class probabilities are approximately uniform within its neighborhood, a new sample's class is predicted according to the majority class of its k neighbors. However, this assumption could become invalid in the case of high-dimensional datasets. Even though, kNN performs better in classification than in regression for with such high dimensionality [137].

After choosing the metric distance, the optimal number of neighbors can be determined by trying different values and comparing the errors in prediction.

5.2.2.2. Partial Least Squares Discriminant Analysis (PLSDA)

PLSDA takes advantage of both methods, PLS and Linear Discriminant Analysis [129,138]. It first performs a dimensional reduction of collinear and noisy data into orthogonal Latent Variables. Then, these PLS-type LVs are used to make a prediction for the new investigated sample as if the observed response was a continuous variable. The obtained value is then compared with a threshold in order to predict the class of the sample. The model interpretation can be carried out with respect to the original variables.

In PLS as well as in PLSDA, the choice of the optimal number of LVs to be selected is made using the measure of fit and validation techniques.

5.2.2.3. Support Vector Machines (SVM)

SVM are a relatively new and sophisticated nonlinear learning method originally developed by Vapnik et al. for binary classification purposes [139–141]. Basically, the idea is to find an hyper-plane able to separate a multidimensional data into two classes. The hyper-plane should be placed in a way to maximize the margin to the nearest data points from the two classes (Figure 2). However, real data is not usually linearly separable, thus, the notion of a kernel function was introduced. This feature enables casting the original data into a higher dimensional space where the data points can be separable. The optimal hyper-plane is determined by a number of Support Vectors (SVs). The commonly used kernel functions are linear, polynomial, sigmoid and radial basis functions (RBF).

Although, computational difficulties could rise from such operation in addition to the high risk of over-fitting. Being an intuitive and theoretically well-founded technique, SVM introduced several parameters to reduce these concerns. Hence, this method was also extended to solve regression problems. The linear model in the high-dimensional space is given by:

$$f(\mathbf{X}, \omega) = \sum_{j=1}^{p} \omega_j g_j(\mathbf{X}) + b$$

where $g_j(\mathbf{X}), j = 1, \dots, p$ represent a set of nonlinear transformations and b is the bias term.

In addition to the type of the kernel function, another important parameter is the constant C that optimizes the compromise between the model complexity and the degree of tolerance to deviations larger than the insensitive loss function ϵ, which is the trade-off between maximizing the margin and minimizing the error rate. The good performance of SVM depends on the suitable setting of these 3 parameters.

The parameter C is also important for the best fit of the model and at the same time to avoid over-fitting problems. It depends on the amount of noise in the training data and it usually varies between 1 and 10. If it's too small the algorithm will insufficiently fit the training data, on the contrary, if it's too large the method will tend to over-fit the data. The parameter ϵ, on the other hand, controls the number of SVs. The higher ϵ, the lower the number of selected SVs. These parameters can only be optimized by analyzing the data and applying proper measures of fit and validation techniques.

In this work, the SVM models were calculated using the LibSVM library written in C programming language and developed by Chih Chang and Chih-Jen Lin [142,143]. This library was implemented in MATLAB to be coupled with the GAs for the variable selection and modeling steps.

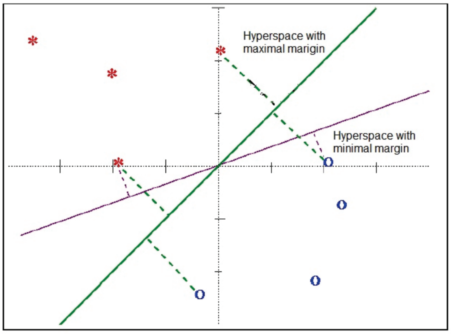

Hyperspace with maximal marigin
Hyperspace with minimal margin

Figure 2: *Choosing the hyperspace with the optimal margin*

6. Goodness of fit measures and validation methods

One of the most important features of a QSAR model is its predictive ability and validity. This condition is also foreseen by the fourth OECD principle for the use of QSARs in regulatory assessment of chemicals. According to the OECD guidance, the validation is defined as: "…*the process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose*" [144].

Care should be taken while assessing the validity of QSAR models in order to avoid the problem of over-fitting and provide predictive algorithms. The optimal model is the one showing the best balance between its complexity and the gain in performance without modeling the noise in the data [1,145,146]. The problem of over-fitting can be due to the bad choice of the modeling technique that doesn't properly fit the studied endpoint or the use of a high number of descriptors with few molecules. Another main reason could be the failure in selecting the suitable descriptors for a given response. The improperly included variables may be inter-correlated, by-chance correlated with the response or too many till capturing higher variance than necessary [147–150].

6.1. Validation methods

As a matter of fact, once a model has been developed, regardless of its type, it is crucial to investigate its predictive ability by means of proper validation methods.

One of the widely used approaches for this purpose is to split the original data into a training and a test set. The test set is usually consisting of 20 to 25% of the whole dataset. This set of molecules is exempted from model calibration process, and it is used to verify the predictive ability of the calibrated model. The model's true predictive ability is evaluated according to the statistics obtained from the external test set. Testing the model using an external validation set is strongly required if the model has shown a significant predictive performance during the modeling process.

Another method to evaluate the model predictive performances is Cross-Validation (CV). There are two varieties of this technique; the Leave-One-Out (LOO) and the Leave-Many-Out (LMO).

The LOO approach consists of leaving out one of the compounds in the training set, fitting the model with the remaining compounds and then predicting the left-out one using the built model. This procedure is repeated for all the compounds in the training set using the same selected descriptors. The statistics are later calculated using the predicted values.

Since LOO is omitting only one compound at a time, it provides over optimistic predictions [151]. This problem can be solved by applying the more robust LMO approach [152]. Albeit its robustness, this method is computationally expensive and irreproducible because it depends on the random selection of the left-out compounds. The k-fold cross-validation is a valid alternative, where k is the number of times one group is left out and predicted using the fitted model. The commonly considered values of k are 5 and 10 with portions equal to 20% and 10%, respectively. Usually, the k groups are divided using venetian blinds or contiguous blocks techniques:

- in venetian blinds method, the test set consists of selecting every k -th sample in the dataset, starting at the first sample.
- the contiguous blocks test set consists of selecting the n/k samples in the dataset, starting at the first sample.

6.2. Regression parameters

The quality of a model can be evaluated using two groups of statistical indices:

- the goodness of fit parameters measuring the fitting ability;
- the goodness of prediction parameters measuring the true predictive ability of a model; these are related to the reliability of prediction in the validation step.

Only the parameters used in this work are presented in this section. However, several indices have been proposed in literature [32].

6.2.1. Goodness of fit indices.

These indices are used to measure the degree to which the model is able to explain the variance contained in the training set. The coefficient of determination R^2 is one of the most used parameters. It is the square multiple correlation coefficient given by:

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

where $\hat{y}$ is the estimated response and $\bar{y}$ is the average observed response over the n training compounds.

R^2 ranges from 0 to 1. The higher this parameter is, the more fitted the model.

The second mainly used parameter is the Root Mean Square Error $(RMSE)$ calculated as following:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

6.2.2. Goodness of prediction indices.

These parameters are used in the validation step. The most important one is the predictive squared correlation coefficient Q^2. Different ways of calculating this parameter are available in the literature [153,154]. In this work, the following formula was considered:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{EXT}} (y_i - \hat{y}_i)^2 / n_{EXT}}{\sum_{i=1}^{n_{TR}} (y_i - \bar{y})^2 / n_{TR}}$$

where n_{EXT} is number of test compounds, n_{TR} is the number of training compounds.

The second parameter commonly used is the Root Mean Square Error in Prediction $(RMSEP)$ calculated as follows:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n_{EXT}}}$$

6.3. Classification parameters

The performance of classification models was evaluated using statistical indices proposed in literature [32,155]. These indices are calculated from the confusion matrix which collects the number of samples of the observed and predicted classes in the rows and columns, respectively (Table 1).

For a two-class dataset, the classification parameters are defined using the number of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

Table 1: The confusion matrix in classification

	Class A (predicted)	**Class B (predicted)**
Class A (observed)	TP	FN
Class B (observed)	FP	TN

The most important parameter that should be maximized during the modeling step is the Non-Error Rate (NER). It is usually expressed in percentage and given by:

$$NER\% = \frac{(Sn + Sp)}{2}$$

where Sn is the sensitivity and Sp is the specificity.

The Sensitivity (Sn), also called the True Positive Rate (TPR) or recall, determines the ability of a model to correctly predict the elements of a given class and calculated as:

$$Sn \equiv TPR = \frac{TP}{TP + FN}$$

The Specificity (Sp), also called the True Negative Rate (TNR), expresses the ability of the model to correctly reject the elements from a given class and defined as:

$$Sp \equiv TNR = \frac{TN}{TN + FP}$$

The Error Rate (ER) is also a significant parameter since it is the complementary value of NER. Thus, it is calculated as following: $ER = 100 - NER\%$

7. Applicability domain of models

The validity of a QSAR model is not sufficient to consider it as adequate for regulatory purposes. General considerations are given in the REACH guidance indicate that it is essential for a QSAR estimate to be valid and applicable to the chemical of interest in order to assess its acceptability [2].

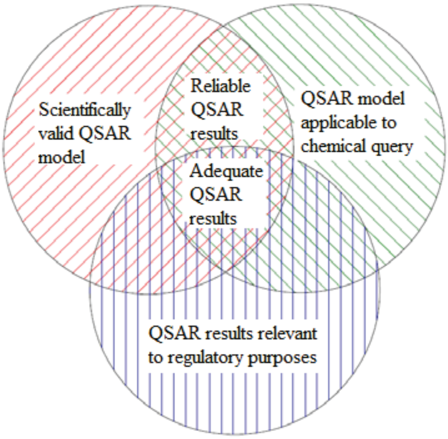

Scientifically valid QSAR model
Reliable QSAR results
QSAR model applicable to chemical query
Adequate QSAR results
QSAR results relevant to regulatory purposes

Figure 3: *The overlapping conditions for the adequacy of QSARs in regulatory purposes.*

This implies that several considerations should overlap in order to fulfill the adequacy condition of a QSAR model in regulatory assessing of chemicals. As shown in Figure 3, a QSAR model should be scientifically well founded and applied within its applicability domain to produce reliable predictions. If these results meet the regulatory field of interest, the model is adequate.

According to the third OECD principle, a QSAR model should be associated with a defined domain of applicability. This includes limitations in terms of types of chemical structures, physicochemical properties and mechanisms of action. When a model is applied within the boundaries of its limitations, it is expected to give reliable estimates. Conversely, using it outside of its applicability domain could affect the accuracy of the predicted results.

Since there is no unique mode of action to define the applicability domain, several methods have been proposed in the literature [156–158]. Depending on the used methodology for describing the descriptor based interpolation space, the suggested methods can be categorized in different groups. The range-based methods include the bounding box, PCA bounding box that define the AD in a univariate way by setting an interval for each variable. The geometric methods such as the convex hull set an external delimiter for the training set as the limit of the AD. Some of the commonly used centroid-based approaches make use of the Leverage, Euclidean, Mahalanobis and City Blok distances with a user defined threshold as a warning value for the AD.

Many other methods have been developed and used in QSAR studies: the kNN approach, the probability density distribution-based method, decision trees and the stepwise approach. Some of the above mentioned approaches have been discussed and a comparison study was conducted on different environmental datasets [159].

In this work, the mostly used approach to define the AD of the developed models was the Leverage approach. The leverages of a given descriptor matrix **X** are obtained from the Hat matrix **H** calculated as follows:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

The diagonal values of **H** are the leverages of the different samples from the centroid of the dataset. According to this approach, the AD of a QSAR model is delimited by a threshold value [156,157]. If a test compound has a leverage value higher than the cut-off it will be considered as outside the AD, thus, associated with low reliable prediction. The user-predefined threshold is generally $3 * p/n$ where p is the number of descriptors plus one and n is the number of samples in the training set.

8. Multi-criteria decision making in model selection

In addition to a thoroughly prepared experimental data, the quality of QSAR model depend on several parameters. As described in the previous sections, the number and type of the crucial parameters vary according to the selected modeling method. In order to build a model with a good compromise between the complexity and the predictive ability, these parameters should be optimized simultaneously during the variable selection step. However, feature selection techniques usually optimize only one parameter such as Q^2 in cross-validation for PLS regression. However, a reliable PLS model should also have a low number of LVs to avoid over-fitting problems. Moreover, a high number of outliers could affect the predictive ability of a model. Thus, ranking the models on the basis of only one parameter can be restrictive and could not give the best results.

Since several criteria can be important for any modeling methodology, suitable techniques for multivariate optimization are required. In the field of Chemometrics, MultiCriteria Decision Making methods (MCDM) have been developed to deal with such problems [160–163]. These methods are able to perform multivariate rankings on the basis of Desirability and Utility indices, and make the optimal choice among the different possibilities. The Utility is calculated as an arithmetic mean of the parameters while the Desirability is defined their geometric mean.

The Utility $\mathrm{U_i}$ of each *i*th alternative for the non-weighted and weighted cases are given by:

$$U_i = \frac{\sum_{j=1}^{p} t_{ij}}{p}, \qquad U_i = \sum_{j=1}^{p} w_j t_{ij}, \qquad 0 \leq U_i \leq 1$$

where p is the number of criteria t.

The Desirability $\mathrm{D_i}$ of each *i*th alternative for the non-weighted and weighted cases are given by:

$$D_i = \sqrt[p]{t_{i1} t_{i2} \dots t_{ip}}, \qquad D_i = t_{i1}^{w_1} t_{i2}^{w_2} \dots t_{ip}^{w_p}, \qquad 0 \leq D_i \leq 1$$

The weight constraint is:

$$\sum_{j=1}^{p} w_r = 1$$

The weights are calculated using the method of normalized weights for ranked criteria [164,165]:

$$w_j' = \frac{Q / r_j^k}{\sum_{j=1}^{p} Q / r_j^k}$$

where $\mathrm{r_j}$ is the *j*th criterion rank, k is a smoothing parameter and Q is defined as:

$$Q = \prod_{j=1}^{p} r_j^k = exp\left[\sum_{j=1}^{p} k \ln(r_j)\right]$$

A new approach for model ranking was developed during this study. It is based on the GAs for variable selection and exploiting the principle of MCDM methods by using the Utility and Desirability functions. The aim of this approach was to include all the relevant criteria in the variable selection process.

This approach was applied on PLS for regression. An algorithm was implemented in MATLAB for the purpose of the study. The variable selection process was performed in multiple double CV (dCV) in order to keep an evaluation set in each step [166]. Intuitively, the dCV is performed in two steps as explained in the algorithm. The included parameters for optimization were: Q^2, the number of variables, LVs, $R^2$for the double CV evaluation set and the number of outliers (nOutliers). This latter parameter is evaluated using the leverage approach as explained in Section II.7.

Each criterion is independently transformed into an Utility/Desirability index. This step is performed by an arbitrary function which transforms the actual value f_{ij} of each ith alternative for the jth criterion into a value between 0 and 1 [165].

The proposed algorithm is the following:

```
Repetition loop: GA runs: FOR r=1 to the total number nRUNS
    (1) Split all n objects randomly into SEGTEST segments (typ. 10).
    (2) Outer loop (dCV): FOR τ = 1 TO SEGTEST
        (a) Select nTEST molecules (1 segment) & nCALIB (the other segments)
        (b) Make GA on the nCALIB molecules (Inner loop: k-fold CV, typ. 5)
                - Select a set of descriptors (nVars) optimizing {D,U}=f($Q^2$, LVs)
        (c) Make PLS models on the nCALIB molecules, predict the nTEST and calculate $R^2$Test.
        (d) Rank chromosomes according to {D,U}= f($Q^2$, LVs, nVars, nOutliers, $R^2$Test).   → next τ dCV
    (3) Do Stepwise forward selection on the τ dCV according to the frequency of selection and rank models according to {D,U}= f($Q^2$, LVs, nVars).   → next r run
    (4) Do final Stepwise forward selection on the nRUNS according to the frequency of selection and rank models according to {D,U}= f($Q^2$, LVs, nVars).
```

After each GA run and in the final stepwise forward selection, the models were ranked using the Utility function because the Desirability appeared to be much restrictive. In fact, even if only one criterion is low, the overall desirability will be low as well. Also if the desirability of one criterion is equal to 0, the overall desirability will be 0.

References

1. Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How not to develop a quantitative structure–activity or structure–property relationship (QSAR/QSPR). *Sar Qsar Environ. Res.* **2009**, *20*, 241–266.

2. ECHA Guidance on Information Requirements and Chemical Safety Assessment. Chapter R6. 2008.

3. ChemSpider | The free chemical database http://www.chemspider.com/ (accessed Apr 26, 2013).

4. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities. In *Annual Reports in Computational Chemistry*; Ralph A. Wheeler and David C. Spellmeyer, Ed.; Elsevier, 2008; Vol. Volume 4, pp. 217–241.

5. The PubChem Project http://pubchem.ncbi.nlm.nih.gov/ (accessed Apr 26, 2013).

6. ChemExper - catalog of chemicals suppliers, physical characteristics and search engine http://www.chemexper.com/ (accessed Apr 26, 2013).

7. Freeland, R. G.; Funk, S. A.; O'Korn, L. J.; Wilson, G. A. The chemical abstract service chemical registry system. II. Augmented connectivity molecular formula. *J Chem Inf Comput Sci* **1979**, *19*, 94– 98.

8. CAS, Chemical Abstracts Service http://www.cas.org/ (accessed Apr 26, 2013).

9. PubMed - NCBI http://www.ncbi.nlm.nih.gov/pubmed (accessed Apr 26, 2013).

10. James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; Chemical Information Systems: Aliso Viejo, CA, USA, 2008.

11. Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - Design and description. *J. Comput. Aided Mol. Des.* **2005**, *19*, 453–463.

12. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.; Schwab, C.; Baskin, I. I.; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q.-Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J. Comput. Aided Mol. Des.* **2011**, *25*, 533–554.

13. The OpenTox project http://www.opentox.org/ (accessed Apr 26, 2013).

14. QSAR DataBank http://qsardb.org/ (accessed Apr 26, 2013).

15. SPARC http://ibmlc2.chem.uga.edu/sparc/test/login.cfm?CFID=264604&CFTOKEN=52651371 (accessed Apr 26, 2013).

16. PBT Profiler http://www.pbtprofiler.net/ (accessed Apr 26, 2013).

17. OECD *QSAR Toolbox*; Oasis, 2011.

18. ECOTOX | MED | US EPA http://cfpub.epa.gov/ecotox/data_download.cfm (accessed May 29, 2013).

19. EPI Suite Data http://esc.syrres.com/interkow/EpiSuiteData.htm (accessed Apr 26, 2013).

20. D, W.; A, W.; SMILES, W. J. L. Algorithm for Generation of Unique SMILES Notation. *J Chem Inf Comput Sci* **1989**, *29*, 97.

21. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

22. Bone R. G.; Firth M.; Sykes R. SMILES Extensions for Pattern Matching and Molecular Transformations: Applications in Chemoinformatics. *J Chem Inf Comput Sci* **1999**, *39*, 846.

23. *ChemBioFinder*; PerkinElmer Informatics Databases.

24. *ChemBioOffice Ultra 13.0 Suite*; Desktop Software; PerkinElmer Informatics.

25. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer, 2007.

26. GNU General Public License - GPL - Free Software Foundation (FSF) http://www.gnu.org/licenses/gpl.html (accessed May 1, 2013).

27. ChemSpider API Services http://www.chemspider.com/AboutServices.aspx (accessed Apr 28, 2013).

28. OCHEM web-services - user's manual http://docs.eadmet.com/display/MAN/Using+web-services (accessed Apr 28, 2013).

29. CIR Chemical Identifier Resolver NCI/CADD http://cactus.nci.nih.gov/chemical/structure (accessed May 16, 2013).

30. Talete, S. R. L. *CIR node for KNIME*; Talete srl, http://www.talete.mi.it.

31. *Guide to intelligent data analysis: how to intelligently make sense of real data*; Texts in computer science; Springer: London ; New York, 2010.

32. Todeschini, R.; Consonni, V. *Molecular descriptors for chemoinformatics*; Wiley-VCH, 2009.

33. Testa, B.; Kier, L. B. The concept of molecular structure in structure–activity relationship studies and drug design. *Med Res Rev* **1991**, *11*, 35– 48.

34. Jurs, P. C.; Dixon, J. S.; Egolf, L. M. Representations of molecules. In *Chemometrics Methods in Molecular Design*; VCH Publishers, New York,, 1995; Vol. 2.0, pp. 15– 38.

35. Thormann, M.; Vidal, D.; Almstetter, M.; Pons, M. Nomen Est Omen: Quantitative Prediction of Molecular Properties Directly from IUPAC Names. *Open Appl. Informatics J.* **2007**, *1*, 28–32.

36. Benigni, R.; Bosa, C. Structural alerts of mutagens and carcinogens. *Curr Comput -Aided Drug* **2006**, *2*, 169– 176.

37. Vidal, D.; Thormann, M.; Pons, M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J Chem Inf Model* **2005**, *45*, 386– 393.

38. Kubinyi, H. Comparative molecular field analysis (CoMFA). In *Handbook of Chemoinformatics*; Wiley- VCH Verlag GmbH, Weinheim, Germany,, 2003; Vol. 4.0, pp. 1555–1575.

39. Kim, K. H. Comparative Molecular Field Analysis (CoMFA). In; Chapman & Hall London, UK, 1995; pp. 291–331.

40. Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

41. Consonni, V.; Todeschini, R. New spectral indices for molecule description. *Match* **2008**, *60*, 3–14.

42. Ivanciuc, O.; Balaban, A. T. The graph description of chemical structures. *Topol. Indices Relat. Descriptors Qsar Qspr* **1999**, 59–167.

43. Hall, G. G. Eigenvalues of molecular graphs. *Bull Inst Math Appl* **1981**, *17*, 70–72.

44. Ivanciuc, O.; Ivanciuc, T. Matrices and structural descriptors computed from molecular graphs distances. In *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach Science Publishers, Amsterdam, The Netherlands, 1999; pp. 221– 277.

45. Ivanciuc, O. Design of topological indices. Part27. Szeged matrix for vertex- and edgeweighted molecular graphs as a source of structural descriptors for QSAR models. *Rev Roum Chim* **2002**, *47*, 479– 492.

46. Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. The complementary distance matrix, a new molecular graph metric. *Ach - Models Chem* **2000**, *137*, 57– 82.

47. Ivanciuc, O. QSAR and QSPR molecular descriptors computed from the resistance distance and electrical conductance matrices. *Ach - Models Chem* **2000**, *137*, 607– 631.

48. Gutman, I. Topological formulas for freevalency index. *Croat Chem Acta* **1978**, *51*, 29– 33.

49. Gutman, I. The energy of a graph: Old and new results. *Algebr. Comb. Appl.* **2001**, 196–211.

50. Gutman, I. Topology and stability of conjugated hidrocarbons. The dependence of total π-electron

energy on molecular topology. *J. Serbian Chem. Soc.* **2005**, *70*, 441–456.

51. Gutman, I.; Zhou, B. Laplacian energy of a graph. *Linear Algebra Its Appl.* **2006**, *414*, 29–37.

52. Zhou, B.; Gutman, I. On Laplacian energy of graphs. *Match* **2007**, *57*, 211–220.

53. Lovasz, L.; Pelikan, J. On the eigenvalue of trees. *Period Math Hung* **1973**, *3*, 175– 182.

54. Estrada, E. Spectral moments of the edge adjacency matrix of molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *J Chem Inf Comput Sci* **1996**, *36*, 844– 849.

55. Estrada, E. Spectral moments of the edgeadjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *J Chem Inf Comput Sci* **1997**, *37*, 320– 328.

56. Estrada, E. Spectral moments of the edge adjacency matrix in molecular graphs. 3. Molecules containing cycles. *J Chem Inf Comput Sci* **1998**, *38*, 23– 27.

57. Estrada, E.; Paltewicz, G.; Uriarte, E. From molecular graphs to drugs. A review on the use of topological indices in drug design and discovery. Indian. *J Chem* **2003**, *42*, 1315– 1329.

58. Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure-Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53*, 867–878.

59. Mansouri, K.; Consonni, V.; Durjava, M. K.; Kolar, B.; Öberg, T.; Todeschini, R. Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling. *Chemosphere* **2012**, *89*, 433–444.

60. Needham, D. E.; Wei, I. C.; Seybold, P. G. Molecular modeling of the physical properties of the alkanes. *J Am Chem Soc* **1988**, *110*, 4186– 4194.

61. *DRAGON (Ver. 6) (Software for Molecular Descriptor Calculations)*; Talete srl, http://www.talete.mi.it: Milano, Italy, 2012.

62. Bonchv, D. an R. *Chemical Graph Theory: Reactivity and Kinetics*; Gordon and Breach Science Publishers: New York, 1992.

63. Devillers, J. *Topological indices and related descriptors in QSAR and QSPR*; Gordon & Breach: Amsterdam, 1999.

64. Kier, L. B.; Hall, L. H. *Molecular connectivity in structure-activity analysis*; Research Studies Press ; Wiley: Letchworth, Hertfordshire, England; New York, 1986.

65. Trinajstić, N. *Chem. Graph Theory* **1992**, 225–273.

66. Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412–1422.

67. Ivanciuc, O.; Ivanciuc, T.; Diudea, M. V. Molecular Graph Matrices and Derived Structural Descriptors. *Sar Qsar Env. Res* **1997**, *7*, 63–87.

68. Balaban, A. T. Local versus global (i.e. atomic versus molecular) numerical modeling of molecular graphs. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 398–402.

69. Klein, D. J.; Lukovits, I.; Gutman, I. On the definition of the hyper-Wiener index for cyclecontaining structures. *J Chem Inf Comput Sci* **1995**, *35*, 50–52.

70. Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.

71. Bonchev, D.; Trinajstic, N. Information theory, distance matrix, and molecular branching. *J Chem Phys* **1977**, *67*, 4517.

72. Balaban, A. T. Enumeration of cyclic graphs. In *Chemical Applications of Graph Theory*; Academic Press, London, UK, 1976; pp. 63– 105.

73. Hanser, T.; Jauffret, P.; Kaufmann, G. A new algorithm for exhaustive ring perception in a molecular graph. *J Chem Inf Comput Sci* **1996**, *36*, 1146– 1152.

74. Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.

75. Hagadone, T. R. Molecular substructure similarity searching: efficient retrieval in twodimensional structure databases. *J Chem Inf Comput Sci* **1992**, *32*, 515– 521.

76. Barnard, J. M. Substructure searching methods: old and new. *J Chem Inf Comput Sci* **1993**, *33*, 532–538.

77. Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In Silico Assessment of Chemical Biodegradability. *J. Chem. Inf. Model.* **2012**, *52*, 655–669.

78. McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J Chem Inf Comput Sci* **1999**, *39*, 569– 574.

79. Senese, C. L.; Duca, J. S.; Pan, D.; Hopfinger, A. J.; Tseng, Y. J. 4D-fingerprints universal QSAR and QSPR descriptors. *J Chem Inf Comput Sci* **2004**, *44*, 1526– 1539.

80. Sciabola, S.; Morao, I.; De Groot, M. J. Pharmacophoric fingerprint method (TOPP) for 3D-QSAR modeling: Application to CYP2D6 metabolic stability. *J. Chem. Inf. Model.* **2007**, *47*, 76–84.

81. Crowe, J. E.; Lynch, M. F.; Town, W. G. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part 1. Non-cyclic fragments. *J Chem Soc C* **1970**, *23*, 990– 997.

82. Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part II. Atom-centred fragments. **1971**, 3702– 3706.

83. Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An algorithm to determine structural commonalities in diverse

datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.

84. Casañola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T. H.; Khan, S. B.; Torrens, F.; Pérez-Jiménez, F.; Rescigno, A.; Abad, C. Bond-based 2D Quadratic fingerprints in QSAR studies: Virtual and in vitro tyrosinase inhibitory activity elucidation. *Chem. Biol. Drug Des.* **2010**, *76*, 538–545.

85. Eckert, H.; Bajorath, J. Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J Chem Inf Model* **2006**, *46*, 2515– 2526.

86. Pozzan, A. 3D pharmacophoric hashed fingerprints. In *Rational Approaches to Drug Design*; 2001; pp. 224– 228.

87. Duca, J. S.; Hopfinger, A. J. Estimation of molecular similarity based on 4D-QSAR analysis: formalism and validation. *J Chem Inf Comput Sci* **2001**, *41*, 1367– 1387.

88. *MDL Information Systems, Inc.*; 14600 Catalina Street, San Leandro, CA 94577, 2004.

89. Durant, J. L.; Leland, B. A.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* **2002**, *42*, 1273– 1280.

90. PubChem Substructure Fingerprint ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt (accessed May 1, 2013).

91. Talete srl. Via V. Pisani, 13 - 20124 Milano - Italy http://www.talete.mi.it/index.htm (accessed May 1, 2013).

92. *dProperties (software for molecular property calculation)*; Talete srl., http://www.talete.mi.it/: Milano, Italy, 2012.

93. Scsibrany, H.; Varmuza, K. *Software SubMat (Generation of Binary Substructure Descriptors)*; Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology: Vienna, 2004.

94. Structure and Substructure Searching. In *Encyclopedia of computational chemistry*; New York : J. Wiley: Chichester, 1998; pp. 2764–2771.

95. Scibrany, H.; Varmuza, K. ToSiM: PC-software for the investigation of topological similarities in molecules. In *Software Development in Chemistry*; Jochum C., Ed.; Gesellschaft Deutscher Chemiker: Frankfurt am Main,, 1994; Vol. 8, pp. 235–249.

96. Varmuza, K.; Demuth, W.; Karlovits, M.; Scsibrany, H. Binary substructure descriptors for organic compounds. *Croat. Chem. Acta* **2005**, *78*, 141–149.

97. Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493–500.

98. GNU Lesser General Public License (LGPL) v3.0 - GNU Project - Free Software Foundation (FSF)

http://www.gnu.org/licenses/lgpl.html (accessed May 1, 2013).

99. The Chemistry Development Kit (CDK) SourceForge http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page (accessed May 1, 2013).

100. Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. L. Recent developments of the Chemistry Development Kit (CDK) - An open-source Java library for chemo- and bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2111–2120.

101. CDK Descriptor Calculator GUI http://rguha.net/code/java/cdkdesc.html (accessed May 1, 2013).

102. Guha, R. Using R to provide statistical functionality for QSAR modeling in CDK. *Cdk News* **2005**, *2*, 2–6.

103. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474.

104. Laggner, C. *SMARTS Patterns for Functional Group Classification*; Inte:Ligand Software-Entwicklungs und Consulting GmbH, 2005.

105. Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinforma. Oxf. Engl.* **2008**, *24*, 2518–2525.

106. Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J Chem Inf Comput Sci* **1995**, *35*, 1039– 1045.

107. PaDEL-Descriptor http://padel.nus.edu.sg/software/padeldescriptor/ (accessed May 2, 2013).

108. Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning*; Addison-Wesley: Reading, MA., 1988.

109. Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J Chemom* **1992**, *6*, 267–281.

110. Leardi, R. Genetic algorithms in chemometrics and chemistry: A review. *J. Chemom.* **2001**, *15*, 559–569.

111. Jennrich, R. I. Stepwise discriminant analysis. *Stat. Methods Digit. Comput.* **1977**, 76–95.

112. Agrafiotis, D. K.; Cedeño, W. Feature Selection for Structure–Activity Correlation Using Binary Particle Swarms. *J. Med. Chem.* **2002**, *45*, 1098–1107.

113. Shen, M.; LeTiran, A.; Xiao, Y.; Golbraikh, A.; Kohn, H.; Tropsha, A. Quantitative structure-activity relationship analysis of functionalized amino acid anticonvulsant agents using k nearest neighbor and simulated annealing PLS methods. *J. Med. Chem.* **2002**, *45*, 2811–2823.

114. Izrailev, S.; Agrafiotis, D. K. Variable selection for QSAR by artificial ant colony systems. *Sar Amp Qsar Env. Res* **2001**, *13*, 417– 423.

115. Forrest, S. Genetic algorithms: principles of natural selection

applied to computation. *Science* **1993**, *261*, 872–878.

116. Venkatraman, V.; Dalby, A. R.; Yang, Z. R. Evaluation of Mutual Information and Genetic Programming for Feature Selection in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1686–1692.

117. Lavine, B. K.; Moores, A. J. Genetic Algorithms in Analytical Chemistry. *Anal. Lett.* **1999**, *32*, 433–445.

118. MATLAB *Version 7.13.0.564*; MathWorks www.mathworks.com, 2011.

119. Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.

120. Free, S. M.; Wilson, J. W. A mathematical contribution to structure–activity studies. *J Med Chem* **1964**, *7*, 395– 399.

121. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 1–27.

122. Winsberg, S.; Carroll, J. D. A quasi-nonmetric method for multidimensional scaling VIA an extended euclidean model. *Psychometrika* **1989**, *54*, 217–229.

123. Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; Wiley, 1986.

124. Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185–196.

125. Geladi, P.; Kowalski, B. R. Partial least squares regression: a tutorial. *Anal Chim Acta* **1986**, *185*, 1– 17.

126. Kowalski, B. R.; Bender, C. F. The K-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal. Chem.* **1972**, *44*, 1405–1411.

127. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14.

128. Bradley, P. S.; Mangasarian, O. L. Feature selection via concave minimization and support vector machines. *Mach. Learn. Proc. Fifteenth Int. Conf. Icml98* **1998**, 82–90.

129. Louwerse, D. J.; Tates, A. A.; Smilde, A. K.; Koot, G. L. M.; Berndt, H. PLS discriminant analysis with contribution plots to determine differences between parallel batch reactors in the process industry. *Chemom. Intell. Lab. Syst.* **1999**, *46*, 197–206.

130. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

131. Anderson, T. W. Asymptotic theory for principal component analysis. *Ann Math Stat* **1963**, *34*, 122–148.

132. Jolliff, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.

133. Weinberger, K. Q.; Saul, L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J Mach Learn Res* **2009**, *10*, 207–244.

134. Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning data mining, inference, and prediction*; Springer: New York, 2009.

135. De Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.

136. Mevik, B. H.; Wehrens, R. The pls package: Principal component and partial least squares regression in R. *J. Stat. Softw.* **2007**, *18*, 1–24.

137. Cover, T.; Hart, P. Nearest neighbor pattern classification. *Ieee Trans. Inf. Theory* **1967**, *13*, 21–27.

138. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173.

139. Vapnik, V. N. *Statistical Learning Theory*; 1st ed.; Wiley-Interscience, 1998.

140. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; 1st ed.; Cambridge University Press, 2000.

141. Schlkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; 1st ed.; The MIT Press, 2001.

142. Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; National Taiwan University, Department of Computer Science: Taipei 106, Taiwan, 2001.

143. Hsu, C.-W.; Lin, C.-J. A comparison of methods for multiclass support vector machines. *Ieee Trans. Neural Networks* **2002**, *13*, 415–425.

144. OECD *Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment*; Series on Testing and Assessment Number 34.; Organisation for Economic Cooperation and Development: Paris, France., 2005.

145. Jouan-Rimbaud, D.; Massart, D. L.; De Noord, O. E. Random correlation in variable selection for multivariate calibration with a genetic algorithm. *Chemom. Intell. Lab. Syst.* **1996**, *35*, 213–220.

146. Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J Chem Inf Comput Sci* **2003**, *43*, 579– 586.

147. Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

148. Wold, S.; Dunn, W. J. I. Multivariate quantitative structure–activity relationships (QSAR): conditions for their applicability. *J Chem Inf Comput Sci* **1983**, *23*, 6– 13.

149. Clark, M.; Cramer, R. D. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Relationships* **1993**, *12*, 137–145.

150. Aptula, A. O.; Jeliazkova, N. G.; Schultz, T. W.; Cronin, M. T. D. The Better Predictive Model: High q2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set? *Qsar Comb. Sci.* **2005**, *24*, 385–396.

151. Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

152. Burden, F. R.; Brereton, R. G.; Walsh, P. T. Cross-validatory selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy. *Analyst* **1997**, *122*, 1015–1022.

153. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678.

154. Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* **2010**, *24*, 194–201.

155. Frank, I. E.; Todeschini, R. The data analysis handbook. *Data Handl. Sci. Technol.* **1994**, *14*, 366.

156. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; Van De Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *Atla Altern. Lab. Anim.* **2005**, *33*, 155–173.

157. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Atla Altern. Lab. Anim.* **2005**, *33*, 445–459.

158. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.

159. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.

160. Keller, H. R.; Massart, D. L.; Brans, J. P. Multicriteria decision making: A case study. *Chemom. Intell. Lab. Syst.* **1991**, *11*, 175–189.

161. Hendriks, M. M. W. B.; de Boer, J. H.; Smilde, A. K.; Doornbos, D. A. Multicriteria decision making. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 175–191.

162. Lewi, P. J.; Van Hoof, J.; Boey, P. Multicriteria decision making using Pareto optimality and PROMETHEE preference ranking. *Chemom. Intell. Lab. Syst.* **1992**, *16*, 139–144.

163. Pavan, M.; Mauri, A.; Todeschini, R. Total ranking models by the genetic algorithm variable subset selection (GA-VSS) approach for environmental priority settings. *Anal. Bioanal. Chem.* **2004**, *380*, 430–444.

164. Pavan, M.; Todeschini, R. Chapter 2 Total-Order Ranking Methods. *Data Handl. Sci. Technol.* **2008**, *27*, 51–72.

165. Pavan, M.; Todeschini, R. Multicriteria Decision Making Methods. In *Comprehensive Chemometrics*; Elsevier, 2009; Vol. 1, pp. 591 – 629.

166. Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated double cross validation. *J. Chemom.* **2009**, *23*, 160–171.

Part III: Results and Discussion

1. Introduction

According to the first OECD Principle, a QSAR model should be associated with a defined endpoint. In the regulatory context, “a defined endpoint” refers to any physicochemical property, biological activity or environmental effect that can be experimentally measured under specific conditions [1]. To ensure reliable predictions for the endpoint being modeled, the considered datasets should be self-consistent and generated by homogeneous experimental protocols. In addition, a QSAR model can be appropriately used for regulatory purposes when the test guidelines used to produce the modeled data are specified. However this is not always feasible, especially when different sources are combined or proprietary databases are used [1].

The transparency of the endpoint being predicted by a given QSAR model is an essential requirement in the assessment of the validity of the model, which is the intent of the first OECD Validation Principle. The predictions of a model can be considered as reliable if its endpoint is congruent with the regulatory endpoint under evaluation. Since the reproducibility of measurements is guaranteed by standardized guidelines, QSAR models based on harmonized test protocols are more likely to provide compliant estimations with the regulatory purposes requirements [1,2].

Table 2: REACH regulatory endpoints associated with the OECD test guidelines.

Category	Endpoint
Physicochemical Properties	Melting Point Boiling Point Vapor Pressure Octanol/Water Partition Coefficient (logP) Water Solubility
Environmental Fate	Biodegradation Hydrolysis Atmospheric Oxidation Bioaccumulation
Ecological Effects	Acute Fish Toxicity Acute Daphnid Toxicity Alga Toxicity Long-term Aquatic Toxicity Terrestrial Effects
Human Health Effects	Acute Oral Toxicity Acute Inhalation Toxicity Acute Dermal Toxicity Skin Irritation /Corrosion Eye Irritation/Corrosion Skin Sensitization Repeated Dose Genotoxicity Reproductive Toxicity Developmental Toxicity Carcinogenicity Organ Toxicity

For regulatory assessment of chemicals within REACH, QSAR models are categorized according to their defined endpoints. The endpoints of interest to this regulation are collected in Table 2, where also the OECD test guideline is specified [1].

In this work, Octanol/Water Partition Coefficient (logP) and two environmental fate endpoints (Biodegradation and Bioaccumulation) were considered. Experimental data for these endpoints were collected from reliable sources and therefore assumed to be produced by means of comparable protocols. The models were developed, validated and interpreted taking in consideration the five OECD principles according to the REACH regulatory requirements.

2. Octanol/Water Partition Coefficient

The chemical interactions of a substance with its surroundings is a key feature for its environmental impact assessment, hence, it is one of the requirements of REACH regulation [3]. The behavior and fate of a chemical substance are mostly depending on its physicochemical properties [4]. In absence of reliable experimental data, non-testing methods such as QSPR estimations can be used to provide such required information about chemicals [3].

The octanol/water partition coefficient (k_{ow}), usually expressed in log values ($logk_{ow}$ or logP) is a key parameter in environmental assessment of chemicals since it is related to lipophilicity/hydrophobicity [5–9]. It is used as the basic predictor in many estimation models for water solubility, bioavailability, bioaccumulation, toxicity/ecotoxicity and PBT assessment/screening [5,10–14]. In REACH regulation, providing a logP value is required for all tonnage bands of chemicals [2,3].

LogP is defined as the ratio of the concentrations of a dissolved chemical in two immiscible phases, octanol and water, at the equilibrium [15]. Since temperature can affect the results, the measurements are typically carried out at 25 °C.

Owing to the large number of available experimental values, robust QSPR models can be developed for this property. When used within their domain of applicability, validated QSPR estimations for logP can be considered in regulatory purposes as more reliable than a single test [1].

Several QSPR models using different methods have been developed and published in the literature [11,16–20]. These models and their results have been compared in several reviews [21,22].

A comparison study of different methods for predicting logP was published by Mannhold and Dross [22]. Later, an exhaustive overview of different methods for estimation of octanol/water partition coefficient as well as other physical properties was published by Katritzky *et al.* [21].

There are two OECD test protocols for logP, OECD Guideline 107 and OECD Guideline 117 [1]. These protocols consider the neutral, undissociated form of a chemical. However, the dissociation of ionisable substances in an environmentally relevant pH could affect their physicochemical properties and, subsequently, their environmental fate. As a result, the partition coefficient of the dissociated form is a different physicochemical property, referred to as logD, and could differ from its neutral form by a factor of 4 to 5 orders of magnitude [23].

In this work, two datasets, with a significant number of molecules, were considered for QSAR modeling. Each dataset was processed separately using appropriate tools and following different modeling strategies.

2.1. Case study 1: the logP-1000 contest

The aim of this study was to participate in a challenge that aimed to develop a predictive model for logP. The logP-1000 contest started with a first dataset of 1000 compounds selected from the ZINC database [24–26]. This initial set was later extended to 1000 clusters of about 5 compounds each. The total of 5200 compounds with unknown logP values will be predicted by the models of the participating groups. In addition to this contest dataset, the organizing group provided also a dataset to be used for fitting the models. The provided dataset consisted of 17233 compounds downloaded from the OCHEM online database [27].

2.1.1. Data set up and curing

The information provided for the compounds of the dataset included the CAS-RN, the chemical name, the SMILES code, the logP experimental value and the internal identifier of the OCHEM database. The dataset was initially analyzed in order to check the presence of erroneous structures.

The first analysis was carried out by means of ChemBio-Office (CambridgeSoft) and revealed 454 molecules associated with wrong structures. In particular, 204 compounds had wrong covalent bonds and 363 compounds had exceeding valence for Nitrogen. The dataset contained also 1648 duplicates and 1727 tautomers.

Using DRAGON software, the unusual covalent bonds of the previously detected 204 compounds were disconnected by converting covalent bonds between Nitrogen and halogens (X) into the disconnected ionic form N^+ X^-. Also covalent bonds between Sodium and Oxygen as well as Potassium and Oxygen were changed into the ionic forms Na^+ O^- and K^+ O^- respectively.

Then, the 454 wrong entries were checked using the following online databases: Pubmed Substance, Chemspider and ChemIDPlus-Advanced. First, the CAS-RN was used, if nonexistent or invalid then the name of the molecule was checked for full match. 219 structures were corrected and 235 were deleted. The final dataset consisted of 16998 compounds.

2.1.2. Molecular descriptor calculation and selection

An initial set of 3130 molecular descriptors was calculated using DRAGON (version 6) [28]. The considered descriptors were related to 9 DRAGON descriptor blocks: atom pairs, atom centered, atom type, CATS, topological, constitutional, functional groups, molecular properties and Muriguchi parameters.

Constant, near constant and highly correlated descriptors were processed as explained in Section II.4.

Then, a univariate correlation analysis with the response (logP) was carried and descriptors with absolute value of correlation coefficient lower than 0.1 were removed. A final set of 1062 descriptors was considered for the modeling step.

The screened dataset was randomly divided into training (12482) and test (4493) sets, representing 74% and 26% of the whole dataset, respectively.

The Genetic Algorithms (GAs) and Stepwise Forward Selection (FS) were used to select the appropriate molecular descriptors for the studied response. The regression models were developed by means of PLS and kNN for regression. The number of Latent Variables (LVs) for PLS and the number of nearest neighbors for kNN were selected maximizing the model's predictive ability Q^2. Cross-validation was performed with 5 cancellation groups divided using the venetian blinds method (details in Section II.6.1).

2.1.3. Results and discussion

Before the proper QSAR modeling, the relationship between molecular weight and logP was analyzed. Most of the compounds demonstrated molecular weights ranging from 150 to 350 g/mol and logP values from 0 to 4. The distribution of molecular weights and logP values can be divided in three intervals:

- 319 compounds with molecular weights ranging from 0 to 100 g/mol related to the lowest logP values;
- 8396 compounds with molecular weights of 100 to 300 g/mol associated with logP values ranging from 0 to 4;
- 3767 compounds with molecular weights higher than 300 g/mol associated with the highest logP values.

The observed correlation between the logP values and the molecular weights (Figure 4) was exploited in order to build a local model using the mentioned molecular weight ranges. Thus, a PLS model was built using the molecules contained in each of the three intervals.

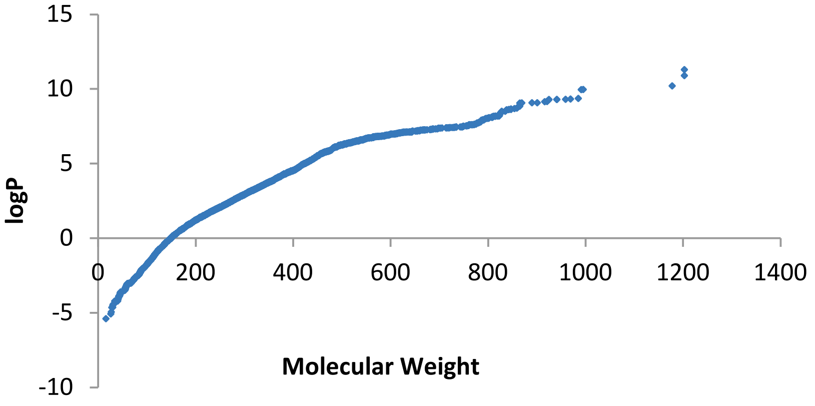


Figure 4: *The correlation between logP and the molecular weights.*

Molecular descriptors were selected by means of GAs and the calibrated models were then validated using the test set. The best results of the different modeling methods (in fitting, cross-validation and test) as well as the number of selected molecular descriptors are collected in the Table 3.

Table 3: QSPR models for logP using different modeling methods.

Method	**No. Desc.**	**LVs/ *k***	**R^2**	***Q*^{2}CV**	***Q*^{2}test**	**RMSEC**	**RMSEP CV**	**RMSEP**
GA_PLS_1	255	20	0.85	0.84	0.86	0.75	0.78	0.75
GA_PLS_2	156	20	0.85	0.84	0.85	0.77	0.80	0.78
FS_PLS	65	15	0.84	0.84	0.85	0.78	0.78	0.77
PLS_MW	65	15	0.86	-	0.86	0.74	-	0.74
*k*NN_1	255	5	-	0.86	0.88	-	0.74	0.68
*k*NN_2	30	5	-	0.84	0.85	-	0.80	0.75
*k*NN_3	65	5	-	0.86	0.87	-	0.74	0.71

No. Desc.: number of descriptors.
GA_PLS: GA coupled with PLS.
FS_PLS: stepwise forward variable selection coupled with PLS.
PLS_MW: GA coupled with PLS using the 3 intervals of molecular weights.

The overall performance of the calibrated models was generally satisfactory, and overfitting was likely limited, if present, since performance in fitting, cross-validation and on the external test set was comparable. The relatively high number of descriptors in these models can be due to the fact that such a big dataset may cover a wide range of structurally diverse chemicals. Thus, a high number of descriptors and LVs for PLS were required to explain most of the variance.

Two of the commonly used QSPR models for predicting logP were developed by Muriguchi (MlogP) and Ghose-Crippen (AlogP) [29,30]. These models were calculated using DRAGON software and used to benchmark the predictive ability of the new proposed models towards the logP-1000 contest dataset of 5200 chemicals.

Table 4: Statistics of MlogP and AlogP for the training and test sets.

Model	$\boldsymbol{R^2}$	**RMSEC**	$\boldsymbol{Q^2}$**test**	**RMSEP**
MlogP	0.68	1.10	0.68	1.10
AlogP	0.80	0.86	0.81	0.86

The performance of AlogP and MlogP models are collected in Table 4. It is clear that AlogP performed better than MlogP for both training and test sets. However, the predictive ability of the new proposed models, summarized in Table 3, is higher than these two models from the literature. The correlation between the predictions obtained from AlogP and MlogP for the whole dataset (training and test set) is 0.88, while their correlation on the logP-1000 contest dataset decreased to 0.69. The difference between these two correlation values was unexpected and could indicate structural difference between the dataset used for fitting the models and that to be predicted by them.

Three of the developed models (GA_PLS_2, FS_PLS and kNN_3) were selected to predict logP for the logP-1000 contest dataset, taking into consideration the compromise between their performance and complexity (number of selected molecular descriptors). These models were benchmarked by calculating the correlations coefficients between their respective predictions on the test set and the contest dataset and those predictions obtained from AlogP and MlogP models. The obtained results are summarized in the Table 5.

Table 5: Benchmarking the predictions of the selected models.

Models	GA_PLS_2		FS_PLS		*k*NN3	
	Test set	Contest data	Test set	Contest data	Test set	Contest data
MlogP	0.83	0.59	0.88	0.81	0.81	0.52
AlogP	0.89	0.62	0.94	0.88	0.87	0.60

According to Table 4, the predictions of the selected models showed higher correlation with AlogP than MlogP. This fact can be considered as proof of the reliability of the selected models since AlogP was considered to be more reliable according to Table 3.

2.1.3. Conclusion

The developed logP models showed similar results. In general, the three final selected models demonstrated better predicting ability than the two classical logP models (AlogP and MlogP), which were used for benchmarking the predictions on the logP-1000 contest dataset.

The *k*NN model showed the best statistics for the training and test sets. The comparison study on the contest data, based on the correlation with AlogP and MlogP indicated better results with the PLS models. In particular, FS_PLS model showed the highest correlation with AlogP which is considered to be better than MlogP. However, it was noticed that the benchmarking models showed low correlation considering the predictions for the contest dataset. This could be due to the fact that the logP-1000 dataset includes several chemicals that are structurally different from those used to fit and validate the models. Consequently, considering both AlogP and MlogP in the evaluation of the predictions on the contest dataset, the FS_PLS could be selected as the best predictive model.

2.2. Case study 2: modeling PHYSPROP dataset for logP

Unlike case study 1 where the data source was constrained, this second study on logP focused more on the dataset preparation in order to have a curated dataset for modeling. Moreover, the previously introduced MCDM variable selection algorithm (Section II.8) was applied to select the best models. Since most of the datasets available in the literature may contain wrong entries, attention was paid to data screening and curation. Then, the modeling step was carried out in order to propose a QSAR model with a good compromise between the predictive ability and complexity.

2.2.1. Data set up and curing

The dataset was downloaded from the US-EPA (Environmental Protection Agency) website [31,32]. This dataset was originated from the PHYSPROP database [33,34]. The same dataset was used for the development of KOWWIN, the EpiSuite's model for estimating logP [19].

The original dataset consisted of 13'445 compounds. For each compound, the CAS-RN, the SMILES structure, the chemical name and the experimental value are provided with the corresponding bibliographic reference. However, not all compounds were associated with a valid CAS-RN since 1872 compounds were associated with a generic internal identifier that has the same number of digits as a CAS-RN.

The data curation was performed using different tools in order to prepare a good quality dataset for modeling purposes. The software dProperties was used to carry out the first check [35]. Since this tool revealed 187 erroneous SMILES structures, further investigations were needed. The data-mining environment, KNIME was used to set-up a workflow which allowed different automatic checks of the dataset entries [36]. The developed workflow (Figure 5) was used to run a series of queries through the web-services of the online databases ChemSpider and CIR [37,38].

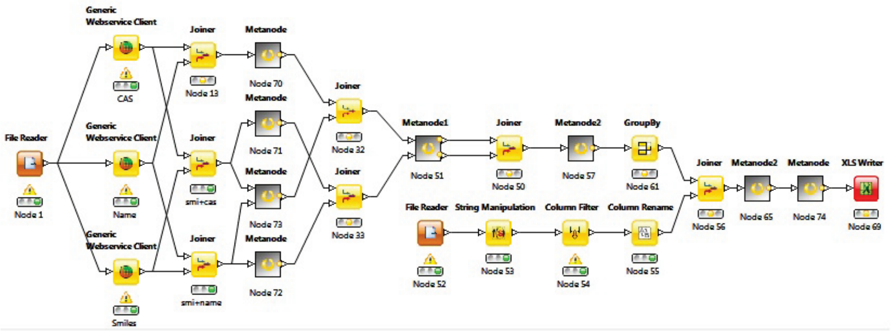


Figure 5: The *KNIME workflow used to prepare the dataset.*

The available identifiers for each compound were used in a combined way. The performed queries are listed from the most to the less restrictive:

- 5524 compounds were found to match the CAS-RN, SMILES and chemical names.
- 6178 compounds were found to match the CAS-RN and the chemical names. This list overlaps with the previous one and adds 662 compounds satisfying only the criteria of the second query.
- 6893 compounds were found to match the CAS-RN and SMILES. This list overlaps with the previous one and adds 1210 compounds.
- 6566 compounds were found to match the SMILES and chemical names. This list overlaps with the previous one and adds 941 compounds.
- 4168 compounds found to match the SMILES and chemical formula were added to the previous list.

The resulting dataset consisted of 12505 molecules with checked molecular structures. The obtained SMILES were used to retrieve the missing CASRNs from ChemSpider database and 407 valid identifiers were found.

79 disconnected structures were removed from the dataset, thus, 11’426 compounds remained for molecular descriptor calculation and modeling.

2.2.2. Molecular descriptors calculation

DRAGON software was used to calculate 2469 molecular descriptors [28]. In order to build easily interpretable models, only 2D descriptors were considered. The calculated descriptors belong to different DRAGON blocks: Constitutional indices, Ring descriptors, Topological indices (except E-state indices sub-block), Walk and path counts, Connectivity indices, Information indices, ETA indices, Functional group counts, Atom Centered fragments, Atom-type E-state indices, CATS 2D and 2D Atom Pairs.

Then, the number of descriptors was reduced by screening the descriptors on the basis of constant, near constant and highly correlated values as explained in Section II.4. The remaining 1167 descriptors were saved for the variable selection and modeling step.

2.2.3. Results and discussion

A test set of 3110 compounds corresponding to 25% of the whole dataset was selected using the venetian blinds technique. The remaining 9316 compounds were considered as training set on which the variable selection step was performed.

The previously described MCDM variable selection based on the GA coupled with PLS was performed on the training set. In each run, 10 double Cross-Validations (dCV) of 5 cancellation groups were performed while the 10% of the training set was left out as a validation set for the best model of the dCV.

During the GA evolutions, 5 parameters were optimized, the inner Q^2 5-fold Cross Validation (5-f CV) and outer Q^2 Cross Validation (dCV) were maximized while the number of variables, the number of LVs and the number of outliers were minimized as explained in Section II.8. The rankings of these 5 criteria and their corresponding weights are listed in Table 6.

Table 6: Ranks and weights of the considered parameters.

	Q^2 5-f CV	**Q^2 dCV**	**Number of descriptors**	**LVs**	**Number of outliers**
Ranking	1	2.5	3.5	3.5	4.5
Weight	0.683	0.171	0.076	0.043	0.027

During the stepwise forward selection performed after each run and at the end of the procedure, 3 parameters were optimized: Q^2 5-fold CV, the number of variables and the number of LVs. The corresponding rankings of the 3 criteria used for calculating their weights were 1, 2.5 and 2.5, respectively. The models were ranked on the basis of the score calculated by the Utility function (*U*). The Desirability function (*D*) was also reported. The smoothing parameter k for calculating the weights was equal to 2.

The maximum number of descriptors and LVs was fixed to 60 and 10, respectively. During the inner 5 fold CV, all the allowed LVs were tested and the model showing the best compromise between the used LVs and Q^2 according to the *U* score was retained.

Since the calculations were computationally expensive due to the big training set and the high number of descriptors, the variable selection procedure was performed in 3 steps. The algorithm was first executed for 20 runs in order to reduce the list of descriptors. In the second step, 331 retained descriptors were subject to 20 runs to select the most pertinent subset. Finally, the 150 descriptors which were the most frequently selected during the second step were included in the last selection step of 20 runs.

Table 7 summarizes the optimized models obtained during the 10 dCVs performed in the first run of the GA and their corresponding parameters used to calculate the *U* score. The descriptors which were selected at least twice in the 10 models were included in the stepwise forward selection according to their frequency of selection. The obtained models from this first run are summarized in Table 8.

Table 7: The 10 dCV performed during the first GA run of the third step.

dCV	*U*	Q^2 5-f CV	Q^2 dCV	No. descs.	LVs	No. outliers
dCV1	0.78	0.81	0.78	39	4	25
dCV2	0.79	0.82	0.81	41	4	48
dCV3	0.78	0.79	0.78	36	3	34
dCV4	0.76	0.76	0.77	33	3	33
dCV5	0.79	0.83	0.84	46	5	30
dCV6	0.79	0.81	0.78	34	4	46
dCV7	0.78	0.80	0.79	38	3	50
dCV8	0.79	0.82	0.84	40	5	27
dCV9	0.78	0.81	0.82	39	4	28
dCV10	0.78	0.80	0.80	34	4	34

Model M6 had the highest *U* score and was, therefore, retained as the best model of the first run. Table 8 reports also the Desirability (*D*) score that showed the highest value for the same model as *U*. Since the maximum of the descriptors to be included in the models was set to 60, models with descriptors exceeding this number had a *D* score equal to 0.

Table 8: Nine models obtained by means of stepwise forward selection performed after the 10 dCVs of the first GA run.

Parameter	M1	M2	M3	M4	M5	M6	M7	M8	M9
Descriptors	1	2	3	4	7	15	33	64	111
Q^2	0.16	0.28	0.43	0.45	0.48	0.79	0.81	0.83	0.84
Selection	10	9	8	7	6	5	4	3	2
LVs	1	1	1	2	2	2	4	4	4
U	0.36	0.45	0.56	0.57	0.58	0.81	0.77	0.73	0.73
D	0.25	0.38	0.52	0.54	0.56	0.81	0.76	0	0

The same procedure was repeated for 20 runs and the best models were saved. Figure 6a showed the frequency of selection of descriptors, while Figure 6b showed the Q^2 CV and the corresponding U score of the 20 obtained models. From these descriptors, those having a frequency of selection of at least 2 over 20 were included in the last stepwise forward selection.

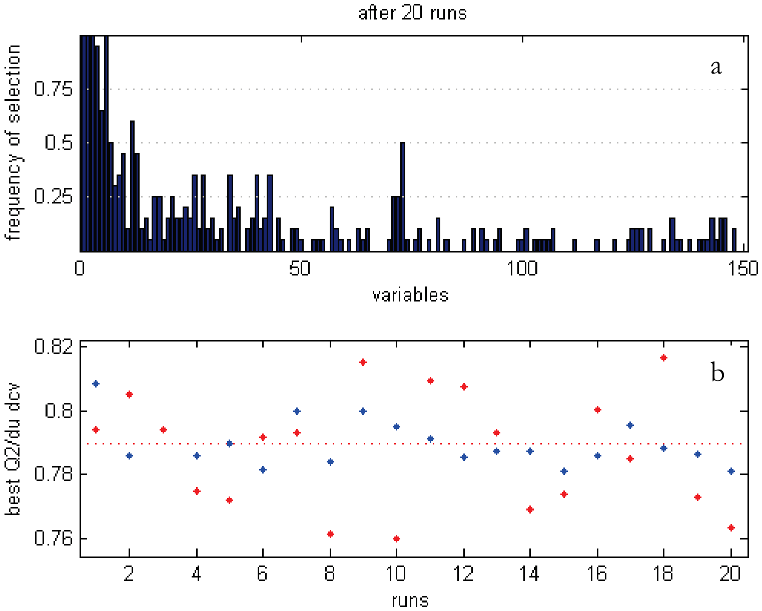


Figure 6: *The frequency of descriptors' selection during 20 runs (a) and the obtained models (b) and their parameters* $\boldsymbol{Q^2}$ *(red points) and U scores (blue points).*

According to Table 9 and Figure 7, the best model resulting from the last stepwise forward selection is model M16 that is associated with the highest U score. It represents the best compromise between performance and complexity since it included 17 descriptors and only 2 LVs for a Q^2 CV equal to 0.8.

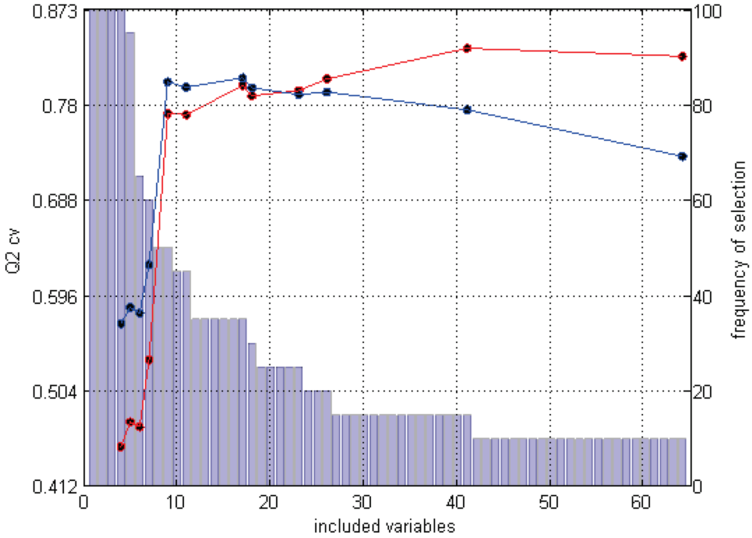


Figure 7: *The evolution of* $\boldsymbol{Q^2}$ *(red line) and U (blue line) during the final Stepwise forward selection. The histogram represents the frequency of selection of the descriptors in percentage over the number of total runs.*

All the descriptors of M16 were included at least 7 times in the best models of the 20 GA runs (Table 9).

Table 9: Evaluation of models resulting from the stepwise forward selection.

Parameter	**M 10**	**M 11**	**M 12**	**M 13**	**M 14**	**M 15**	**M 16**	**M 17**	**M 18**	**M 19**	**M 20**
Desc.	4	5	6	7	9	11	17	18	23	26	41
Q^2	0.45	0.47	0.47	0.53	0.77	0.77	0.80	0.79	0.79	0.80	0.83
Nb. Select.	20	19	13	12	10	9	7	6	5	4	3
LVs	2	2	2	2	2	2	2	2	2	2	3
U	0.57	0.58	0.58	0.63	0.80	0.80	0.81	0.80	0.79	0.79	0.77
D	0.54	0.56	0.55	0.61	0.80	0.80	0.80	0.79	0.79	0.79	0.75

Desc.: the number of descriptors.
Nb. Select.: the number of selection of the added descriptors in the 20 runs.

The selected descriptors, listed in Table 10, are simple 2D descriptors encoding information about the size of molecules, functional groups and fragments which can be related to the lipophilicity of chemicals.

Table 10: The molecular descriptors included in the model M16.

Symbol	Description	Block
B07[C-X]	Presence/absence of C - X at topological distance 7	2D Atom Pairs
B05[C-X]	Presence/absence of C - X at topological distance 5	2D Atom Pairs
H-046	H attached to C0(sp3) no X attached to next C	Atom-centered fragments
O-058	=O	Atom-centered fragments
C-006	CH2RX	Atom-centered fragments
C-001	CH3R /CH4	Atom-centered fragments
O-056	alcohol	Atom-centered fragments
CATS2D_01_LL	CATS2D Lipophilic-Lipophilic at lag 01	CATS 2D
nX	number of halogen atoms	Constitutional indices
nHM	number of heavy atoms	Constitutional indices
RBN	number of rotatable bonds	Constitutional indices
nHDon	number of donor atoms for H-bonds (N and O)	Functional group counts
nHAcc	number of acceptor atoms for H-bonds (N,O,F)	Functional group counts
nCbH	number of unsubstituted benzene C(sp2)	Functional group counts
nCb-	number of substituted benzene C(sp2)	Functional group counts
nBnz	number of benzene-like rings	Ring descriptors
PCD	difference between multiple path count and path count	Walk and path counts

The selected best model was finally validated by means of the external test set that was not used during the modeling step. The regression performance of this model in fitting, CV and prediction on test set are summarized in Table 11.

Table 11: Statistics of model M16.

Fitting		5-fold CV		Test	
R^2	RMSEC	Q^2	RMSECV	Q^2	RMSEP
0.80	0.82	0.80	0.82	0.81	0.80

On the basis of the results shown in Table 11, model M16 can be considered to be robust since the statistics in fitting, cross validation and test are comparable.

The applicability domain of the model was investigated by means of the leverage approach. The number of outliers detected in the test set was 86. These compounds did not affect the statistics of the model. This low number of molecules outside the AD could be a result of the optimization of the number of outliers during the modeling step. Consequently, it can be concluded that the selected descriptors are an optimal subset to cover a wide range of the chemical space of the training set.

2.2.4. Conclusion

The developed MCDM-GA algorithm was able to select the best subset of descriptors by optimizing all the important parameters of the PLS method in a weighting scheme. The performed procedure leaded to a QSPR model with good compromise between the performance and the complexity. The utility of the final stepwise selection according to the frequency of the descriptors is to include all the gathered information from the different GA runs.

Although the size of the dataset and the high variance it contained, the statistics of the built model were satisfactory for a global model. In comparison with the first study in the framework of the logP-1000 contest (see Section III.2.1), the selected final model required a much lower number of descriptors and latent variables for a small difference in the predictive ability.

3. Bioaccumulation

The bioaccumulation of a chemical substance in aquatic organisms is a crucial information for understanding its environmental behavior. The increase of concentration of a chemical in the tissues due to its accumulation over long term exposure may cause toxic effects and transfer through the food web leading to biomagnification.

Consequently, for REACH it is a relevant information at all supply levels and it is a requirement for substances manufactured or imported in quantities of 100 ton/year or more. This information is also used in chemical safety assessment and food chain exposure as well as PBT classification [39]. For these reasons, REACH encourages the establishment of bioaccumulation data although below the requirement tonnage and the use of prediction techniques such as QSARs as alternatives to animal testing.

3.1. Definitions

In the literature, there are several valid definitions describing the accumulation of chemicals in biota. In common terms, it is the result of the 4 phases a substance goes through in an organism: absorption (uptake), distribution, metabolism and excretion (ADME). The elimination of chemicals in aquatic organisms is processed by diffusive transfer across intestinal walls and gill surfaces or biotransformation to more easily excreted metabolites [40,41].

Bioconcentration is a term referring to the accumulation of a substance in an aquatic organism. The BioConcentration Factor (BCF) of a chemical is the ratio of its concentration in the tissues of an organism over its concentration in water at the steady state as following: $\mathrm{BCF} = \mathrm{C_o}/\mathrm{C_w}$

where BCF is the bioconcentration factor (L/kg), $\mathrm{C_o}$is the chemical concentration in the whole organism (mg/kg, wet weight) and $\mathrm{C_w}$is the chemical concentration in water (mg/L).

The BioAccumulation Factor (BAF) is expressed as the ratio of the concentrations of a chemical in the organism tissues and the surrounding medium at equilibrium. It considers the uptake from all the environmental sources including water, food and sediments.

The BioMagnification Factor (BMF) measures the accumulation of chemical substances via the food chain. It is expressed by the ratio of the concentrations of the substance in the predator and the prey: $\mathrm{BMF} = \mathrm{C_o}/\mathrm{C_d}$

where BMF is dimensionless, $\mathrm{C_o}$ is the steady-state chemical concentration in the organism (mg/kg), $\mathrm{C_d}$ is the steady-state chemical concentration in the diet (mg/kg).

The concentrations should be expressed on a wet weight basis. They may also be normalized on the basis of the lipid content [39].

3.2. Assessing bioaccumulation by QSARs

QSAR modeling is one of the most pertinent non testing methods accepted within REACH. Validated models for assessing bioaccumulation could provide relevant and reliable predictions on the chemicals of interest for the regulatory purposes.

Different approaches for modeling the bioaccumulation factors have been proposed and reviewed in the literature [41–43].

The most important approaches can be divided in 2 categories according to the used descriptors: models based on experimental descriptors and models based on theoretical descriptors.

In all cases, attention should be paid when merging datasets obtained from different experimental conditions because it can affect the model's predictions [44].

3.2.1. QSAR models based on experimental descriptors

LogP is commonly used as a simple estimator for bioaccumulation exploiting the correlation between BCF and the hydrophobicity of chemicals. The mechanistic interpretation of such relationship can be the analogy of the partition process between the lipid tissues and water as a passive diffusion through gill membranes in the aquatic organisms to its simulation in the logP experiments [39].

Several logBCF/logP relationships have been proposed for specific chemical classes, such as polycyclic hydrocarbons, while many others were developed for diverse classes of chemicals [45–51]. Some of these models have already been used in regulatory applications of a number of chemicals [39].

Linear models based on logP provide acceptable estimations of the BCF for non ionic and slowly metabolized chemicals. However, since the range of logP values may be too large, this correlation is valid only for logP values varying from 1 to 6 and breaks down for more hydrophobic compounds [52]. The BCF values of such compounds are lower than the predictable limit of the correlation hypothesis and this is due to several reasons including the low aqueous solubility leading to low bioavailability, failure in reaching the steady state in the case of large molecules in addition to metabolism and degradation processes [44,52].

More advanced approaches have been proposed to overcome this problem. Bilinear models and polynomial relationships have been developed for logP values ranging from 1.12 to 8.6 [49,53]. Another logP based approach was developed for the EpiSuite's model BCFWIN. It suggested the use of different fragments for each group of chemicals in multi-logP ranges models with correction factors to improve the accuracy of the global model [7].

However, the logP based predictions for high hydrophobic compounds remain uncertain for regulatory use [39].

Another experimental descriptor correlated with BCF is the aqueous solubility (S) which is highly, negatively, correlated with the previous descriptor. Although it is less extensively used than logP, several models for estimating BCF were based on this physicochemical property [54–57]. As for the previous experimental descriptor, BCF models based on S may have accuracy problems for specific chemical groups [57].

3.2.2. QSAR models based on theoretical molecular descriptors.

The experimental descriptors, such as logP and S, were selected prior to the modeling procedure in order to fit a predefined mechanistic interpretation of the mode of action of the training set compounds. In addition to the explained drawback of such hypothesis that could not be valid for some groups of chemicals, these approaches are facing another problem which is the lack of experimental input data for the structures to be predicted.

To overcome these limitations, the use of theoretical molecular descriptors which can be calculated for any chemical structures was proposed in the literature. Using statistical methods, different classes of molecular descriptors were correlated with the bioaccumulative potential of chemicals including molecular connectivity indices, solvation energy, molecular fragments and quantum chemical descriptors [58–62].

Theoretical descriptors avoid the problem of variability encountered with experimental descriptors. However, the models proposed in literature for mixed groups of chemicals are not always associated with a defined applicability domain which is a requirement for the regulatory applications [1].

3.3. Case study: QSARs for assessing bioaccumulation

In order to comply with the regulatory requirements for the assessment of the environmental behavior of chemicals, cautious approaches are needed. The lack of input data can be avoided by the use of theoretical descriptors, which are independent of any experimental testing.

The aim of this study was to develop theoretical descriptors-based QSAR models for the assessment of bioaccumulation. The models were specifically built for the chemical group of interest to avoid any extrapolation of the applicability domain.

3.3.1. Polybrominated diphenyl ethers (PBDEs)

During the last decades, Polybrominated diphenyl ethers (PBDEs) were the most commonly used group of brominated flame retardants (BFRs). These chemicals were used in textile and electrical equipment industries as additives to polymers and resins [63,64]. Since they are not bonded to plastics, these pollutants are easily released to the environment during the manufacture phase, while the consumers are using the products and continue to leak out of the wastes that constitute the major diffuse source of pollution [64].

PBDEs are known for their long range atmospheric transport, in fact, they are usually detected in different geographical regions distant from their original sources [65]. Because of their toxicity, persistence and potential for bioaccumulation these pollutants were included in the OSPAR list of chemicals for priority action and some of them were added to the list of Stockholm convention for POPs [64,66].

Depending on the number and positions of the bromine atoms on the two phenyl groups, there are 209 possible congeners. In a similar way as for Chlorobiphenyls (CBs), the PBDE congeners are numbered according to the International Union of Pure and Applied Chemistry (IUAPAC) nomenclature. Similar toxic properties have also been notices between CBs and PBDEs [67–

69]. However, the second group of chemicals are more lipophilic than their corresponding chlorinated compounds [70].

3.3.2. Results of PBDEs bioaccumulation models

The aim of this study was to assess the bioaccumulation of PBDEs by means of QSAR modeling [71]. However, bioaccumulation is a complex biological and environmental procedure involving a multitude of factors. Hence, modeling such an endpoint can be compromised by the possible biotransformation of these compounds. In this work, attention was paid to the metabolism of some BDE congeners by debromination which can affect the reliability of the predictions.

The modeling procedure of this study was achieved in 3 steps corresponding to the 3 factors (BCF, BAF and BMF), which are usually used to assess bioaccumulation. Different regression methods were applied and several models were compared. For each one of the 3 factors, the model presenting the best compromise between performance and simplicity was selected. Since the aim of the study was to propose reliable models for a maximum number of BDEs, much attention was paid to the applicability domain of the developed models. The complete study can be found in the published article provided in the attached Annex I.

4. Biodegradability

The transformation of a chemical substance in the environment by degradation is an important process influencing the long term exposure to pollutants. The degraded chemical can give stable and/or toxic products. Hence, understanding this process leads to better risk assessment of adverse effects on biota. Degradation is abiotic or non-biological when it involves only physicochemical reactions. While biotic degradation is a biological process known as biodegradation and can occur in aerobic or anaerobic conditions depending on the presence/absence of oxygen.

Information on biodegradability of chemicals may also be used in classification and labeling within the persistency assessment (PBT/vPvB). In the literature, there are several experimental datasets for degradation rates of chemicals. The most applicable experimental conditions for regulatory purposes are based on the standardized OECD guidelines such as OECD 301, OECD 303, OECD 111, OECD 308 and OECD 309.

Within the context of REACH, biodegradability is an endpoint of high interest for the regulation of chemicals [72]. Starting from a volume of production of 1 ton/year, the registration dossier should include information on the ready biodegradability of the substance since the exposure potential increases with the volume [72]. However, independently from the tonnage trigger, all sources of information can be considered for the risk characterization including non-testing predictive methods such as QSARs [72].

4.1. QSARs for assessing biodegradability of chemicals.

Biodegradability can be computationally assessed in a quantitative or a qualitative way. Several models have been proposed in the literature for both types. A comprehensive review of biodegradability models was published in the literature [73]. Most of these models were derived from a dataset consisting of 894 compounds assessed by the Japanese Ministry of International Trade and Industry (MITI).

The EpiSuite's probability program BIOWIN is one of the commonly used tools that provide estimations of the biodegradability under aerobic conditions with mixed cultures of microorganisms [74].

CATALOGIC is a less known quantitative model for assessing biodegradability based on a mechanistic approach. It predicts the Biological Oxygen Demand (BOD) and the microbial biodegradation CO_2 production. It provides also an attempt to the metabolic pathways and the plausible biodegradation products that may arise [75].

TOPCAT, which is a commercial suite for toxicology predictions, also includes a module for quantitative assessment of aerobic biodegradability. It consists of 4 models applicable on specific classes of chemicals [76].

The list can be extended to several other models such as the commercial software MULTICASE for ecotoxicity and TOXTREE which is a free decision tree based tool [77,78]. Both of these models are based on molecular fragments and structural alerts.

4.2. Summary of the published study on biodegradability

The aim of this work was to apply advanced modeling methods in order to build QSAR models with high predictive ability to contribute to the implementation of REACH regulation. The used classification methods were: *k*NN, PLSDA and SVM as well as consensus modeling. Attention was paid to the screening and preparation of the dataset for the modeling steps. The study was extended by an analysis of the used molecular descriptors and their relationship with the modeled endpoint, based on information retrieved from the literature. In particular, the newly used molecular descriptors for modeling biodegradability, such as the matrix-based descriptors, were further explained by means of simple MLR models involving classical interpretable descriptors encoding information such as molecular branching and size [79].

More details can be found in the published article of the study provided in the attached Annex II.

4.3. Substructural keys for predicting biodegradability

This study aimed to evaluate the ability of some substructural descriptors to predict the biodegradability. More details on the used dataset for this purpose can be found in the published article provided in the Annex II.

This QSAR study used only binary descriptors based on several structural keys calculated by PADEL and SubMat (Table 12). For this purpose, a kNN routine using binary descriptors was implemented in MATLAB. The similarity indices Jaccard-Tanimoto (JT) and Consonni-Todeschini (CT4) were used for calculating the binary distances ($1 - Similarity$) [80]. The best QSAR models obtained in this first step are summarized in Table 10. All models were validated with 5 cancellation groups and then using the test set. The classification performance of the models was evaluated by means of error rate, class specificity (Sp, correctly predicted ready biodegradable) and sensitivity (Sn, correctly predicted non ready biodegradable). The statistics of the best obtained models were comparable in cross-validation (5f-CV) and different for the test set. However, the 166 MACCS keys calculated by PADEL seemed to have more accurate predictions on the test set with the lowest ER equal to 15.2%. Despite the amount of information encoded into the 4860 structural keys, Klekota showed average performance on CV and test set.

The published models in the previously mentioned study based on the DRAGON descriptors performed better than the different used substructural keys [79].

Table 12: The selected ***k***NN models using different combinations of structural keys and distance measures.

Structural keys (number)	**Distance**	***k***	**5f-CV**			**Test**		
			ER CV	**Sp**	**Sn**	**ER test**	**Sp**	**Sn**
Submat (1365)	JT	10	0.196	0.754	0.854	0.184	0.708	0.925
MACCS (166)	JT	8	0.198	0.718	0.886	0.152	0.806	0.890
Padel-E_State (79)	CT4	2	0.201	0.771	0.826	0.256	0.667	0.822
Klekota (4860)	JT	4	0.205	0.775	0.816	0.179	0.806	0.836
Pubchem (881)	CT4	10	0.208	0.754	0.830	0.204	0.750	0.842

4.4. Predicting biodegradability from the BOD values

This modeling approach aimed to make a biodegradability classification based on the BOD values. First regression models were built in order to predict the BODs, then the compounds were categorized using the threshold of 60%. Compounds with BODs lower than 60% are considered as NRB while those exceeding this threshold were considered as RB. The kNN in regression was used in both weighted and non-weighted versions as explained in Section II.5.2.1. The used metric distances were the Manhattan, Minkowski and Euclidean.

Several blocks of DRAGON descriptors were calculated, then GA was applied in order to select the most appropriate subsets. The parameter k was optimized, from 1 to 10, in order to get the best Q^2 in 5-fold CV. The models with the best Q^2 CV were selected. Their statistics were calculated also for the test set and summarized in Table 13.

For this dataset, the Euclidean distance showed the best results. Thus, only the models using this distance were reported in Table 13.

Table 13: Statistics of weighted and non-weighted $\boldsymbol{k}$NN regression models.

Model	Descs.	k	CV				Test			
			non-weighted		weighted		non-weighted		weighted	
			Q^2	RMSEC	Q^2	RMSEC	Q^2	RMSEP	Q^2	RMSEP
1	24	8	55.9	32.55	58.3	31.67	45.6	37.39	45.6	37.39
2	38	10	54.6	33.04	57.2	32.08	46.0	37.24	47.3	36.78
3	42	10	53.8	33.33	56.6	32.29	46.4	37.08	47.9	36.57
4	49	8	53.8	33.32	56.4	32.36	46.2	37.18	46.9	36.94
5	15	6	52.9	33.64	55.1	32.84	47.9	36.60	49.5	36.02

Desc.: the number of included descriptors.

The statistics of the 5 models were not very high compared to usual regression models. However, when the predictions of the 1st model, which showed the best Q^2, were plotted against the experimental BOD values (Figure 8), the majority of the compounds seemed to be assigned to their correct classes.

Figure 8 is, indeed, divided into 4 sections by the BOD threshold of 60%. The upper left square contains the NRBs predicted as RBs, the dots in lower left section represent the correctly predicted NRBs while the correctly predicted RBs are in the upper right section leaving the wrongly assigned RBs to the lower right side. It is clear that the ER in the compounds assigned as RBs is higher than NRBs.

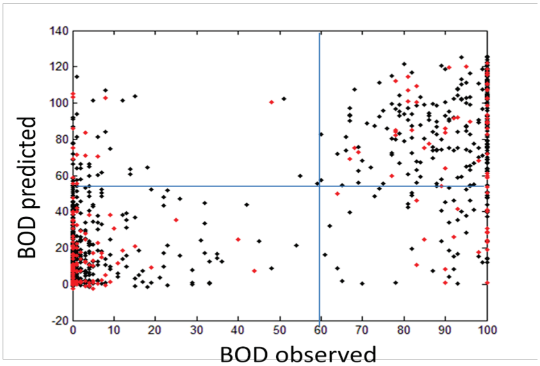


Figure 8: *Predicted versus observed BOD values of the training set (black points) and test set (red points).*

The predicted BOD values were after that used to make a classification of the training and test compounds using the predefined threshold. The results of the classification procedure are summarized in Table 14.

Table 14: Statistics of weighted and non-weighted ***k***NN classification models.

Model	**CV**						**Test**					
	non-weighted			**weighted**			**non-weighted**			**weighted**		
	ER	**Sp**	**Sn**	**ER**	**Sp**	**Sn**	**ER**	**Sp**	**Sn**	**ER**	**Sp**	**Sn**
1	0.160	0.761	0.919	0.145	0.789	0.922	0.156	0.750	0.938	0.159	0.750	0.932
2	0.155	0.778	0.911	0.138	0.806	0.917	0.163	0.764	0.911	0.176	0.736	0.911
3	0.153	0.785	0.910	0.143	0.799	0.915	0.159	0.750	0.932	0.159	0.750	0.932
4	0.158	0.768	0.917	0.149	0.785	0.917	0.156	0.764	0.925	0.149	0.778	0.925
5	0.165	0.746	0.924	0.159	0.757	0.926	0.149	0.778	0.925	0.152	0.778	0.918

Albeit the average statistics in regression, the classification performance was acceptable compared to the previously developed models using the structural keys (Table 12). In particular, Model 1 and Model 5 showed interesting performances in addition to low numbers of descriptors.

The ER for the weighted predictions showed a better performance in CV but it did not follow the same behavior for the test set. Hence, it can't be concluded which method is performing better. It can also be noted that the sensitivity and specificity are not balanced as it is supposed to be for a good model that accurately predicts both classes. As noticed in Figure 8, all 5 models confirmed that the NRB compounds of this dataset are easier to predict than RBs.

5. Applicability domain of QSARs

Defining the applicability domain of QSAR models is the third OECD principle and is one of the requirements for the predicted results to be used for regulatory purposes.

The AD is defined by the chemical space covered by the training set of the model. This is equivalent to the descriptor space that describes the structures of the used compounds. Thus, the applicability of a model is limited to the structurally similar compounds to the training set. The model's estimate is considered reliable when the chemical in query is interpolated within the AD. Any extrapolation of that defined space is associated with lower reliability in prediction.

Different AD approaches have been proposed in the literature [81–83]. Depending on the adopted methodology in characterizing the interpolation descriptor space, the approaches discussed in this study can be categorized into range-based and geometric methods, distance based methods and probability density distributions.

5.1. Different approaches for defining the AD

These approaches differ by the way the delimiters of the training set's descriptor space is defined [81].

The simplest method is called the Bounding Box and is based on the range of individual descriptors. It considers that a compound is inside the AD only if its descriptors values are falling between the minimum and the maximum values of the corresponding descriptors of the training set. Another variety of the same approach considers the ranges of the principle components of a PCA instead of the original descriptors.

Convex Hull is a geometric approach aiming to define the AD by the smallest convex space that can enclose the whole training set. This approach is similar to the range based since it defines only the external delimiters of the chemical space independently from the data distribution [81].

The most commonly used approaches are distance based. The concept of these methods is similar to that of the previously defined leverage approach. It consists of measuring the distance separating a query data-point to the center of the training set, then compares it with a predefined threshold distance. If the test compound is less distant than the cut-off it can be considered inside the AD of the model. These approaches are considering that the further the test compound is from the center of the training set the less reliable the prediction is. The most usual distance measures employed for this purpose are Mahalanobis, Manhattan and the Euclidean distances.

Another approach tested in this work was the probability density distribution method. It consists of estimating the probability density and identifying the highest density region of the dataset. The created potential is at its highest value at each compound of the training set and decreases with the distance [81].

Each approach has its advantages and drawbacks. Even though, the behavior of an AD approach depends on the used model and the dataset it was applied on. The number of the detected compounds outside the AD is also a result of the predefined parameters. Consequently, it is up to the model developer and user to define the most appropriate approach to use for the specific model under evaluation.

5.2. Summary of the published study on the AD approaches

The aim of this study was to provide a comparison between different approaches for defining the applicability domain. In this work, some of the previously introduced approaches, in addition to few other ones, were defined and their adopted algorithms explained. Then the selected approaches for the study were evaluated and compared varying their thresholds [84].

The complete study is published and the article is provided in the attached Annex III.

6. Structure-activity landscapes

According to the congenericity principle, structurally similar compounds are assumed to be associated with similar activities. However, the activity landscape of QSAR datasets is not always as smooth as thought. Similar molecules may have different activities leading to uneven landscape with Activity Cliffs (ACs). The presence of ACs in a given dataset can raise several problems for QSARs. The difference between the SAR landscapes was compared by Maggiora (2006) to the difference between "*the gently rolling hills found on the Kansas prairie*" and "*the rugged landscapes of Utah's Bryce Canyon*" [85].

6.1. The Structure-Activity Landscape Index (SALI)

The first index for assessing the activity cliffs in a dataset was proposed by Maggiora (2006) and named the Structure-Activity Landscape Index (SALI) [85]. Later several different studies using the SALI index and graphical methods for characterizing the activity landscapes have been published [86–94].

According to Maggiora (2006), ACs are expressed by the ratio of the difference in activity of two compounds over their "distance" in the chemical space [85]. Activity cliffs are described in terms of the Structure-Activity Landscape Index (**SALI**) as follows:

$$SALI_{ij} = \frac{|A_i - A_j|}{1.01 - sim(i,j)}$$

where $A_i$and A_j are the activities of the ith and the jth molecules, and $sim(i, j)$ is the similarity coefficient between the two molecules.

Figure 9 shows an example of the activity landscape according to the SALI index using the Euclidean distance. The dataset used for the plot consisted of 49 points obtained from two simulated variables (**X**) and a simulated activity (**Y**). The points placed in the 2D space and the responses were chosen in a way to create activity and similarity cliffs. The 3D symmetric plot vaguely emphasized the (bright) regions in the dataset associated with high activity cliffs. In particular, two regions can be noticed: one is corresponding to the points 30-40 with the points 1-15 while the second one is corresponding to the points 40-49 with the points 20-40.

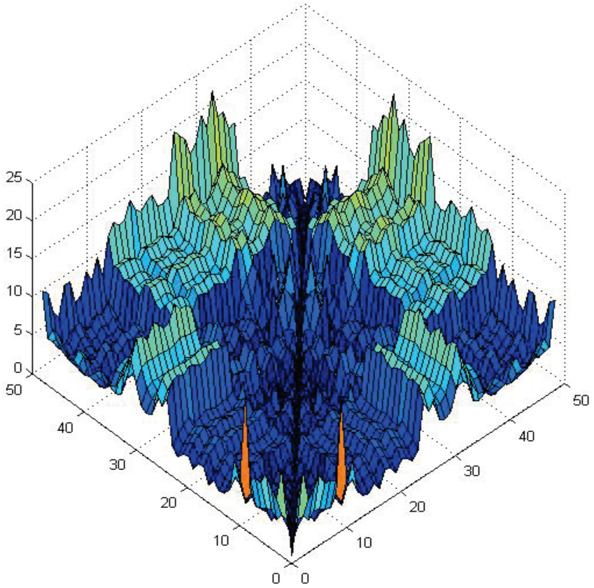


Figure 9: *The activity cliffs of the simulated dataset using SALI. The* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *axis represent the number of samples while the* $\mathbf{z}$*-axis represents the difference in activity.*

6.2. Graphical methods for characterizing SAR landscapes

6.2.1. The Structure-Activity Similarity (SAS) map

One of the widely used methods to graphically explore the activity landscape is the Structure-Activity Similarity (SAS) map where activity similarity and structural similarity for each pair of compounds are plotted [95–97]. An example of the SAS map applied on the previously mentioned simulated dataset is given in Figure 10.

The SAS map can be divided in four main regions (Figure 10). Pairs located in region I are characterized by low activity similarity and low structural similarity. Pairs with low activity similarity and high structural similarity are located in region II and therefore pairs of compounds in this region have a discontinuous SAR (activity cliffs). Data points located in region III are associated with low structural similarity and high similar activity; therefore this region is affected by structural cliffs. Finally, region IV identifies pairs of compounds with high structural similarity and high activity similarity and therefore correspond to continuous SAR.

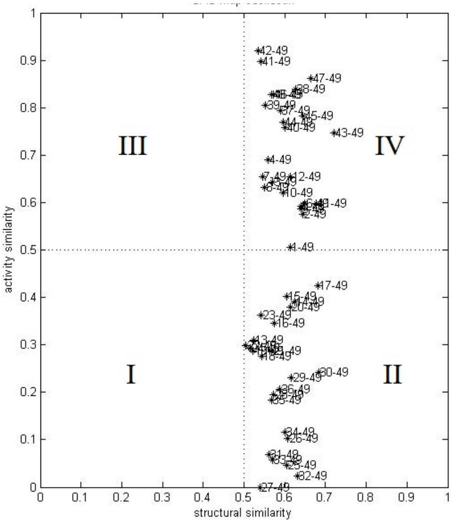


Figure 10: *SAS map applied on the simulated dataset using the Euclidean distance.*

6.2.2. The Patterson plot

The Patterson plot is a method to graphically investigate the structurally similar compounds and their relative activity similarity [98]. As in the SAS map, the points in the Patterson plot represent the pairs of molecules in the dataset. The absolute differences in activities of the pairs of molecules are plotted in function of the distances between them in the descriptor space. For binary descriptors, such as substructural keys, the used similarity measure is converted to a distance measure for the abscise axis as $1 - Similarity$.

If the dataset obey to the congenericity principle, the pairs of molecules will appear in the lower triangle of the plot [99]. Thus in comparison to the SAS map, the structural cliffs and activity cliffs regions will switch places.

To measure the degree to which the congenericity principle is respected, the "Patterson ratio" can be calculated. It is the ratio of the average absolute difference in activity for all the pairs of the dataset to the average absolute difference for the molecules with a similarity higher than a user defined threshold usually 0.7 (or 0.3 for $1 - Similarity$ distance). The higher the ratio value the lower activity cliffs present in the data.

6.3. Metric distances for investigating SAR landscapes

6.3.1. The used metric distances.

The metric distances employed in this work (in progress) in order to explore the SAR landscapes were the Euclidean, Manhattan and the Soergel distances.

The Euclidean distance between two samples s and t in a p dimensional space is calculated as follows:

$$d_{st} = \sqrt{\sum_{j=1}^{p} (x_{sj} - x_{tj})^2}$$

The Manhattan distance between the two samples s and t in the same p dimensional space is given by:

$$d_{st} = \sum_{j=1}^{p} \left| x_{sj} - x_{tj} \right|$$

These two distances vary between 0 and ∞. Thus, a prior scaling of the data or a conversion of the distance to a similarity measure between 0 and 1 is often needed. The most simple way to calculate the similarity from the distance is:

$$Sim = \frac{1}{1 + d_{st}} \qquad 0 \leq Sim \leq 1$$

The Soergel distance between the two samples s and t is calculated as follows:

$$d_{st} = 1 - \frac{\sum_{j=1}^{p} min\{x_{sj}, x_{tj}\}}{\sum_{j=1}^{p} max\{x_{sj}, x_{tj}\}} = \frac{\sum_{j=1}^{p} |x_{sj} - x_{tj}|}{\sum_{j=1}^{p} max\{x_{sj}, x_{tj}\}} \qquad 0 \leq d_{st} \leq 1$$

where p is the number of variables.

For binary data, the Soergel distance is the complement of the Jaccard-Tanimoto [100,101]. Thus, it could be possible to use the Soergel distance not only for real numbers but also for binary variables and mixed-type data without the necessity to any weighting scheme.

Since the Soergel distance varies between 0 and 1 and it was noticed that it is less sensitive to the scaling compared to the previous two metric distances. consequently, there is no need to scale the data before using the Soergel distance which is the case of many other distance measures.

6.3.2. Comparison of the distances using the Patterson plot.

A subset of 430 molecules was randomly extracted from the previously described logP dataset consisting of 12505 molecules (see Section III.2.2). Using DRAGON software, the substructural descriptors of the block Atom-centered fragments were calculated. The total number of retained descriptors was 105.

The Soergel, Euclidean and the Manhattan distances were used to make the Patterson plots. The different ratios were calculated using a threshold of 0.3. The red lines on the plots (Figure 11, Figure 12 and Figure 13) indicate the values used to calculate the Patterson ratio as explained in Section III.6.2.2. The average value and the 95 percentile of the SALI index are also calculated for each plot. The scaling is performed by dividing by maximum value of each descriptor.

All pairs of molecules with both Euclidean and Manhattan distances, without scaling are shown to be far from each other (Figure 11a and Figure 12a). In these two plots, the Patterson ratio reached 7 which is a relatively high value for an heterogeneous dataset. This high value do not indicate an optimal SAR landscape for QSAR modeling since the interval of distances between 0 and 0.6 is not populated. While with the scaled data, the plots showed a Gaussian pattern with a maximum value of distance between the pairs not exceeding 0.9 (Figure 11b and Figure 12b). In these two cases, the Patterson ratio is lower than the previous two plots which may indicate the presence of activity cliffs in the dataset. This is confirmed by the higher average and 95 percentile values for the SALI index.

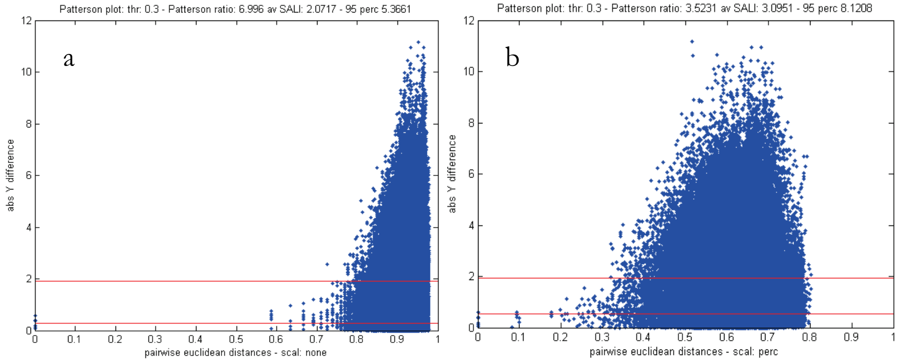


Figure 11: *The pairwise Euclidean distance without scaling (a) and scaled (b). thr: the used threshold for calculating the Patterson ratio; av SALI: the average value of the SALI index on all pairs; 95 perc: the 95 percentile of the SALI index on all pairs.*

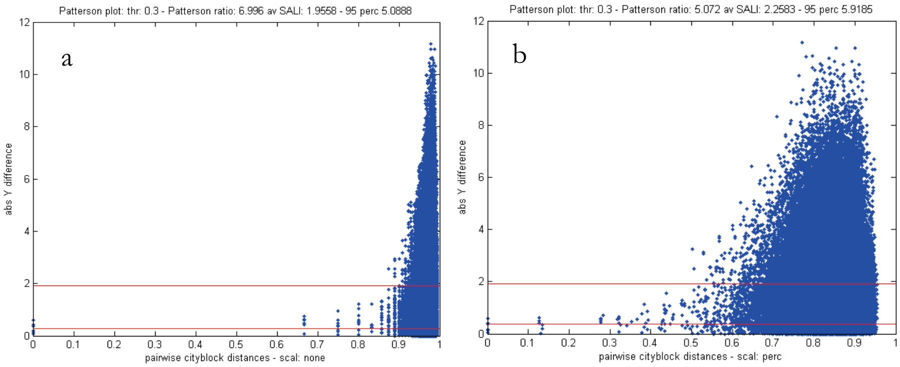


Figure 12: *The pairwise Manhattan distance without scaling (a) and scaled (b). thr: the used threshold for calculating the Patterson ratio; av SALI: the average value of the SALI index on all pairs; 95 perc: the 95 percentile of the SALI index on all pairs.*

In all the mentioned figures, the pairs of molecules seem to have similar distances between them since all of them are located in a narrow interval of the x-axis. This is not the usual distribution of randomly selected datasets of such a number of molecules. This means that, probably, the Euclidean and the Manhattan distances in both scaled and non-scaled cases did not show the real distribution of the molecules in the descriptor space of the dataset.

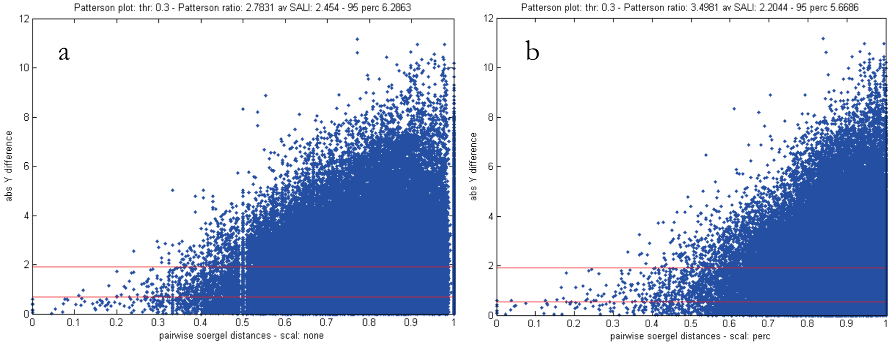


Figure 13: *The pairwise Soergel distance on non-scaled (a) and scaled data (b). thr: the used threshold for calculating the Patterson ratio; av SALI: the average value of the SALI index on all pairs; 95 perc: the 95 percentile of the SALI index on all pairs.*

Unlike the two previous distances, the Soergel distance showed similar patterns with the scaled and non-scaled data (Figure 13). This confirms the fact that Soergel distance is independent from the scaling. Also, the Patterson ratios, the average and 95 percentile of SALI index have similar values in both plots indicating a similar SAR landscape.

The pairs of molecules are more distributed on the x-axis to occupy most of the distance interval between 0 and 1 which is the expected behavior for such number of different molecules.

The general pattern of these last two plots, showing an increasing difference in the activity with the increase in the distance between the pairs of molecules, indicates a relatively smooth landscape for this dataset. Hence, this dataset obeys to the congenericity principle which makes it adapted for QSAR modeling.

The Soergel distance showed interesting properties making it more suitable for the investigation of the SAR landscapes compared to the Euclidean and the Manhattan distance measures because it is much less dependent on the scaling and does not require the calculation of the similarity, being already normalized between 0 and 1. As further work, the Soergel distance could be tested for SAR landscape exploring in the case of datasets with real and mixed-type values.

7. Conclusion

The manufactured chemical substances provide a large range of services and tools supporting the modern lifestyle and economies. Nevertheless, the increased quantities of chemicals in the environment may endanger human health and the environment. Hence, there is a need to improve the scientific understanding of the effects of the chemicals that can find their way to the environment and end-up in the living organisms.

In order to find the right balance between the benefits of chemicals and their side effects, their risk assessment is required by REACH. Since there is a need to waive animal testing and reduce the risk assessment costs, REACH promotes the use of alternative methods such as QSAR/QSPR models.

In this thesis, the conceptual basics of QSAR modeling were explained. After that, the different steps to be taken during the analysis study, the technical details of the applied methodologies as well as newly tested molecular descriptors have been introduced. In addition, the validation and the reliability assessment techniques were described, with reference to the REACH requirements.

The mentioned steps for QSAR modeling have been followed in the applications section of this work. Three endpoints with interest to REACH legislation including the physicochemical property logP the bioaccumulation and the biodegradation, have been modeled.

LogP is known to be an important parameter for a multitude of biological activities and environmental fate of chemicals. This property have been subject for two case studies in this work. The first was aiming to predict the logP values for a set of chemicals with unknown experimental responses within the log-1000 challenge involving several research groups. A number of QSPR models have been developed for the purpose and the best three models were selected and submitted. These models showed good and robust statistics in fitting, cross-validation and predictive ability on the test set. In addition the predictions for the contest dataset were benchmarked with commonly used models from the literature (MlogP and AlogP). The second case study was intended to test a new approach for variable selection coupling the GAs with the MCDM methods on the Syracuse database for logP. The developed algorithm applied on PLS resulted in a model with reasonable compromise between the predictive ability and the complexity of the model parameters usually required for such big datasets. Hence, the Utility function used to score the models demonstrated its usefulness in selecting the best models when several parameters have to be optimized simultaneously. In this study, the quality of the data, which is an important factor in QSAR modeling, was a result of the use of the automated KNIME workflow.

Since the bioaccumulation is one of the REACH most required endpoints for environmental fate assessment, this endpoint was modeled for a specific group of chemicals. Being a list of widely used POPs during the last decades, PBDEs are the centre of a number of studies involving toxicity and environmental side effects of these chemicals. The three commonly used factors for assessing the bioaccumulation of chemicals, (BCF, BAF and BMF) have been modeled using different data sources. Then, the values of the three factors have been predicted for the whole set of 209 BDE congeners [71]. The developed models showed good predictive ability and their applicability domain demonstrated a maximal coverage of the 209 BDE congeners. Especially for BCF, which is the most important factor between the three mentioned, the proposed model in this study presented better results on PBDEs in comparison with global models from the literature.

The last modeled endpoint of interest to REACH regulation was the biodegradability. In this study, a special interest was given to the preparation of the dataset before the modeling step. Then three models and their consensus have been proposed using different classification methods: PLSDA, kNN and SVM. The developed models were validated in three steps using cross-validation, a test set left out from the same dataset and an external validation set gathered from different sources. The models showed a good predictive ability in comparison with previous published studies in the literature [79]. The thorough data screening contributed in a significant way to good results of the models. Moreover, the consensus modeling also improved the predictive ability of the developed models by considering the three classification methods at the same time.

In addition to the modeling results, methodological aspects of QSARs have been discussed. Theory and applications of applicability domain approaches were explained in a comparison study [84].

In addition, the SALI index for the assessment of the structure-activity landscapes have been introduced. Then, it was used to compare the usefulness of three metric distances (Euclidean, Manhattan, Soergel) for the characterization of activity cliffs in QSAR data. The Soergel distances showed interesting features that will be further investigated for the purpose.

Even though, the biological activity is a complex process involving multiple parameters, the developed QSAR models showed good estimation of the predicted endpoints especially when the data is well curated and the appropriate tools applied. Thus QSAR/QSPR modeling is a useful technique for filling the gap of knowledge about chemicals, thus it is useful for regulatory purposes.

This work was an attempt to contribute to the implementation of the European regulation on chemicals REACH. The studies were conducted within the European project ECO-ChemOinformatics (http://www.eco-itn.eu/), in collaboration with different partner groups participating to the same project as well as other related, ongoing and finished, European projects.

References

1. OECD *Guidance Document on the Validation of (Quantitative) Structure Activity Relationship (Q)SAR Models.*; OECD Environment Health and Safety Publications. Series on Testing and Assessment No. 69.; Organisation for Economic Cooperation and Development: Paris, France., 2007.

2. REACH - Environment - European Commission http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed Apr 21, 2013).

3. European Chemicals Agency *Guidance on information requirements and chemical safety assessment. Chapter R.7a: Endpoint specific guidance*; European Chemicals Agency: Annankatu 18, Helsinki, Finland, 2012.

4. Mackay, D.; Di Guardo, A.; Paterson, S.; Cowan, C. E. Evaluating the environmental fate of a variety of types of chemicals using the EQC model. *Environ. Toxicol. Chem.* **1996**, *15*, 1627–1637.

5. Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handb. Chem. Prop. Estim. Methods* **1990**.

6. Fisk, A. T.; Norstrom, R. J.; Cymbalisty, C. D.; Muir, D. G. G. Dietary accumulation and depuration of hydrophobic organochlorines: Bioaccumulation parameters and their relationship with the octanol/water partition coefficient. *Environ. Toxicol. Chem.* **1998**, *17*, 951–961.

7. Meylan, W. M.; Howard, P. H.; Boethling, R. S.; Aronson, D.; Printup, H.; Gouchie, S. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environ. Toxicol. Chem.* **1999**, *18*, 664–672.

8. Xia, Z.; Yang, J.; Li, L.; Yang, F.; Jiang, X. Determination of Octanol-Water Partition Coefficients by MEEKC Based on Peak-Shift Assay. *Chroma* **2010**, *72*, 495–501.

9. Lombardo, A.; Roncaglioni, A.; Boriani, E.; Milan, C.; Benfenati, E. Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. *Chem. Cent. J.* **2010**, *4*.

10. Sabljic, A.; Guesten, H.; Hermens, J.; Opperhuizen, A. Modeling octanol/water partition coefficients by molecular topology: chlorinated benzenes and biphenyls. *Environ. Sci. Technol.* **1993**, *27*, 1394–1402.

11. Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley & Sons, Inc.: New York, 1979.

12. Yang, G.; Cao, W.; Zhu, T.; Bai, L.; Zhao, Y. The QRAR model study of β-lactam antibiotics by capillary coated with cell membrane. *J. Chromatogr. B* **2008**, *873*, 1–7.

13. Detroyer, A.; Vander Heyden, Y.; Carda-Broch, S.; García-Alvarez-Coque, M. .; Massart, D. . Quantitative structure-retention and retention-

activity relationships of β-blocking agents by micellar liquid chromatography. *J. Chromatogr. A* **2001**, *912*, 211–221.

14. Oszwałdowski, S.; Timerbaev, A. R. Development of quantitative structure–activity relationships for interpretation of the migration behavior of neutral platinum(II) complexes in microemulsion electrokinetic chromatography. *J. Chromatogr. A* **2007**, *1146*, 258–263.

15. Weber Jr, W. J.; Chin, Y.-P.; Rice, C. P. Determination of partition coefficients and aqueous solubilities by reverse phase chromatography—I: Theory and background. *Water Res.* **1986**, *20*, 1433–1442.

16. Leo, A. J. Calculating log Poct from structures. *Chem Rev* **1993**, *93*, 1281–1306.

17. Rekker, R. F. *Hydrophobic Fragm. Constant* **1977**.

18. Klopman, G.; Li, J.-Y.; Wang, S.; Dimayuga, M. Computer automated log P calculations based on an extended group contribution approach. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 752–781.

19. Meylan, W. M.; Howard, P. H. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J. Pharm. Sci.* **1995**, *84*, 83–92.

20. Petrauskas, A.; Kolovanov, E. A. ACD/log P method description. *Persp Drug Disc* **2000**, *19*, 99– 116.

21. Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally diverse quantitative structure–property relationship correlations of technologically relevant physical properties. *J Chem Inf Comput Sci* **2000**, *40*, 1– 18.

22. Mannhold, R.; Dross, K. Calculation Procedures for Molecular Lipophilicity: a Comparative Study. *Quant. Struct.-Act. Relationships* **1996**, *15*, 403–409.

23. Tetko, I. V.; Bruneau, P. Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database. *J. Pharm. Sci.* **2004**, *93*, 3103–3110.

24. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.

25. Irwin, J. J.; Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

26. ZINC- A database of commercially-available compounds http://zinc.docking.org/ (accessed Jun 5, 2013).

27. Novotarskyi, S.; Sushko, I.; Körner, R.; Kumar, A.; Rupp, M.; Prokopenko, V.; Tetko, I. OCHEM - On-line CHEmical database & modeling environment. *J Cheminf* **2010**, *2*, 5.

28. *DRAGON (Ver. 6) (Software for Molecular Descriptor Calculations)*; Talete srl, http://www.talete.mi.it: Milano, Italy, 2012.

29. Moriguchi, I.; Hirono, S.; Liu, Q.; Nakagome, I.; Matsushita, Y. Simple method of calculating octanol/water

partition coefficient. *Chem. Pharm. Bull. (Tokyo)* **1992**, *40*, 127–130.

30. Ghose, A. K.; Crippen, G. M. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships. I. Partition coefficients as a measure of hydrophobicity. *J Comput Chem* **1986**, *7*, 565–577.

31. US Environmental Protection Agency http://www.epa.gov/ (accessed May 16, 2013).

32. EPI Suite Data http://esc.syrres.com/interkow/EpiSuiteData.htm (accessed Apr 26, 2013).

33. Tetko, I. V.; Tanchuk, V. Y.; Villa, A. E. P. Prediction of n-octanol/water partition coefficients from PHYSPROP database using artificial neural networks and E-state indices. *J Chem Inf Comput Sci* **2001**, *41*, 1407– 1421.

34. Rosenberg, S. A.; Hueber, A. E.; Aronson, D.; Gouchie, S.; Howard, P. H.; Meylan, W. M.; Tunkel, J. L. Syracuse research corporation's chemical information databases: Extraction and compilation of data related to environmental fate and exposure. *Sci. Technol. Libr.* **2002**, *23*, 73–87.

35. *dProperties (software for molecular property calculation)*; Talete srl., http://www.talete.mi.it/: Milano, Italy, 2012.

36. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer, 2007.

37. CIR Chemical Identifier Resolver NCI/CADD http://cactus.nci.nih.gov/chemical/structure (accessed May 16, 2013).

38. ChemSpider Royal Society of Chemistry, Thomas Graham House (290), Science Park, Milton Road, Cambridge CB4 0WF http://www.chemspider.com/ (accessed Oct 29, 2012).

39. European Chemicals Agency *Guidance on information requirements and chemical safety assessment. Chapter R.7c: Endpoint specific guidance*; European Chemicals Agency: Annankatu 18, Helsinki, Finland, 2012.

40. Ecetoc Persistence of chemicals in the environment. *Persistence Chem. Environ.* **2003**.

41. Boethling, R. S.; Mackay, D. *Handb. Prop. Estim. Methods Chem. Environ. Heal. Sci.* **2000**.

42. Dearden, J. QSAR Modeling of Bioaccumulation. In *Predicting Chemical Toxicity and Fate*; CRC Press, 2004.

43. Pavan, M.; Worth, A. P.; Netzeva, T. I. *Review of QSAR Models for Bioconcentration*; EUROPEAN COMMISSION JOINT RESEARCH CENTRE: Institute for Health and Consumer Protection Toxicology and Chemical Substances Unit European Chemicals Bureau I-21020 Ispra (VA) Italy, 2006.

44. Nendza, M. *Structure Activity Relationships in Environmental Sciences*; Springer, 1998.

45. Schüürmann, G.; Klein, W. Advances in bioconcentration prediction. *Chemosphere* **1988**, *17*, 1551–1574.

46. Neely, W. B.; Branson, D. R.; Blau, G. E. Partition coefficient to measure bioconcentration potential of organic chemicals in fish. *Environ. Sci. Technol.* **1974**, *8*, 1113–1115.

47. Veith, G. D.; DeFoe, D. L.; Bergstedt, B. V. Measuring and estimating the bioconcentration factor of chemicals in fish. *J Fish Res Board Can.* **1979**, *36*, 1040–1048.

48. Mackay, D. Correlation of Bioconcentration Factors. *Es T Contents* **1982**, *16*, 274–278.

49. Connell, D. W.; Hawker, D. W. Use of polynomial expressions to describe the bioconcentration of hydrophobic chemicals by fish. *Ecotoxicol. Environ. Saf.* **1988**, *16*, 242–257.

50. Gobas, F. A. P. C. A model for predicting the bioaccumulation of hydrophobic organic chemicals in aquatic food-webs: application to Lake Ontario. *Ecol. Model.* **1993**, *69*, 1–17.

51. Dimitrov, S. D.; Dimitrova, N. C.; Walker, J. D.; Veith, G. D.; Mekenyan, O. G. Predicting bioconcentration factors of highly hydrophobic chemicals. Effects of molecular size. *Pure Appl. Chem.* **2002**, *74*, 1823–1830.

52. Gobas, F. A. P. C.; Shiu, W. Y.; Mackay, D. Factors Determining Partitioning of Hydrophobic Organic Chemicals in Aquatic Organisms. In *QSAR in Environmental Toxicology - II*; Kaiser, K. L. E., Ed.; Springer Netherlands: Dordrecht, 1987; pp. 107–123.

53. Devillers, J.; Lipnick, R. L. Practical applications of regression analysis in environmental QSAR studies. In *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*; Springer: Kluwer, Dordrecht, The Netherlands, 1990; pp. 129–144.

54. Chiou, C. T.; Freed, V. H.; Schmedding, D. W.; Kohnert, R. L. Partition coefficient and bioaccumulation of selected organic chemicals. *Environ. Sci. Technol.* **1977**, *11*, 475–478.

55. Kenaga, E. E.; Goring, C. A. I. Relationship between water solubility and soil sorption, octanol-water partitioning and bioconcentration of chemicals in biota. In *Aquatic Toxicology*; American Society for Testing and Materials: Philadelphia, PA, USA, 1980; pp. 78–115.

56. Jorgensen, S. E.; Halling-Sorensen, B.; Mahler, H. *Handb. Estim. Methods Ecotoxicol. Environ. Chem.* **1998**.

57. Isnard, P.; Lambert, S. Estimating bioconcentration factors from octanol-water partition coefficient and aqueous solubility. *Chemosphere* **1988**, *17*, 21–34.

58. Sabljic, A. The prediction of fish bioconcentration factors of organic pollutants from the molecular connectivity model. *Z. Gesamte Hyg.* **1987**, *33*, 493–496.

59. Park, J. H.; Lee, H. J. Estimation of bioconcentration factor in fish, adsorption coefficient for soils and

sediments and interfacial tension with water for organic nonelectrolytes based on the linear solvation energy relationships. *Chemosphere* **1993**, *26*, 1905–1916.

60. Tao, S.; Hu, H.; Lu, X.; Dawson, R. W.; Xu, F. Fragment constant method for prediction of fish bioconcentration factors of non-polar chemicals. *Chemosphere* **2000**, *41*, 1563–1568.

61. Tao, S.; Hu, H.; Xu, F.; Dawson, R.; Li, B.; Cao, J. QSAR modeling of bioconcentration factors in fish based on fragment constants and structural correction factors. *J. Environ. Sci. Health B* **2001**, *36*, 631–649.

62. Wei, D.; Zhang, A.; Wu, C.; Han, S.; Wang, L.-S. Progressive study and robustness test of QSAR model based on quantum chemical parameters for predicting BCF of selected polychlorinated organic compounds (PCOCs). *Chemosphere* **2001**, *44*, 1421– 1428.

63. OSPAR Commission. *Certain Brominated Flame Retardants-Polybrominated diphenyl ethers, polybrominated biphenyls, hexabromo cyclododecane*; OSPAR Priority Substances; OSPAR Commission: London, 2001.

64. OSPAR Commission. *Tetrabromobisphenol-A-Update*; OSPAR Priority Substances; OSPAR Commission: London, 2005.

65. Mikula, P.; Svobodová, Z. Brominated flame retardants in the environment: Their sources and effects (a review). *Acta Vet. Brno* **2006**, *75*, 587–599.

66. UNEP Recommendations of the Persistent Organic Pollutants Review Committee of the Stockholm Convention to amend Annexes A, B or C of the Convention 2009.

67. De Wit, C. A. An overview of brominated flame retardants in the environment. *Chemosphere* **2002**, *46*, 583–624.

68. McDonald, T. A. A perspective on the potential health risks of PBDEs. *Chemosphere* **2002**, *46*, 745–755.

69. Pijnenburg, A. M.; Everts, J. W.; de Boer, J.; Boon, J. P. Polybrominated biphenyl and diphenylether flame retardants: analysis, toxicity, and environmental occurrence. *Rev. Environ. Contam. Toxicol.* **1995**, *141*, 1–26.

70. Webster, L.; Russel, M.; Walsham, P.; Moffat, C. F. *A REVIEW OF BROMINATED FLAME RETARDANTS (BFRS) IN THE AQUATIC ENVIRONMENT AND THE DEVELOPMENT OF AN ANALYTICAL TECHNIQUE FOR THEIR ANALYSIS IN ENVIRONMENTAL SAMPLES*; Fisheries Research Services Internal Report; Fisheries Research Services Marine Laboratory: Victoria Road Aberdeen AB11 9DB, 2006.

71. Mansouri, K.; Consonni, V.; Durjava, M. K.; Kolar, B.; Öberg, T.; Todeschini, R. Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling. *Chemosphere* **2012**, *89*, 433–444.

72. European Chemicals Agency *Guidance on information requirements and chemical safety assessment. Chapter R.7b: Endpoint specific*

guidance; European Chemicals Agency: Annankatu 18, Helsinki, Finland, 2012.

73. Pavan, M.; Worth, A. P. Review of estimation models for biodegradation. *Qsar Comb. Sci.* **2008**, *27*, 32–40.

74. EPISuite v. 4.0. **2010**.

75. L *CATALOGIC*; Laboratory of Mathematical Chemistry: Burgas, Bulgaria.

76. Accelrys *TOPKAT toxicity suite*; Accelrys.

77. Toxtree — Institute for Health and Consumer Protection – (JRC-IHCP), European Commission http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/toxtree (accessed Apr 26, 2013).

78. Multicase Ecotoxicity http://www.multicase.com/products/prod098.htm (accessed May 22, 2013).

79. Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure-Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53*, 867–878.

80. Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52*, 2884–2901.

81. Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T. D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; Myatt, G.; Nikolova-Jeliazkova, N.; Patlewicz, G. Y.; Perkins, R.; Roberts, D. W.; Schultz, T. W.; Stanton, D. T.; Van De Sandt, J. J. M.; Tong, W.; Veith, G.; Yang, C. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *Atla Altern. Lab. Anim.* **2005**, *33*, 155–173.

82. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Atla Altern. Lab. Anim.* **2005**, *33*, 445–459.

83. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.

84. Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.

85. Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

86. Amić, D.; Lučić, B.; Kovačević, G.; Trinajstić, N. Bond dissociation enthalpies calculated by the PM3 method confirm activity cliffs in radical scavenging of flavonoids. *Mol. Divers.* **2009**, *13*, 27–36.

87. Guha, R.; Van Drie, J. H. Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

88. Hu, Y.; Bajorath, J. Molecular scaffolds with high propensity to form multi-target activity cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 500–510.

89. Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477–491.

90. Namasivayam, V.; Bajorath, J. Searching for Coordinated Activity Cliffs Using Particle Swarm Optimization. *J. Chem. Inf. Model.* **2012**, *52*, 927–934.

91. Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 1021–1033.

92. Vogt, M.; Huang, Y.; Bajorath, J. From activity cliffs to activity ridges: informative data structures for SAR analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1848–1856.

93. Wassermann, A. M.; Bajorath, J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J. Chem. Inf. Model.* **2010**, *50*, 1248–1256.

94. Iyer, P.; Stumpfe, D.; Vogt, M.; Bajorath, J.; Maggiora, G. M. Activity Landscapes, Information Theory, and Structure - Activity Relationships. *Mol. Informatics* **2013**, n/a–n/a.

95. Shanmugasundaram, V.; Maggiora, G. M. Characterizing property and activity landscapes using an information-theoretic approach. In; Cinf-032; American Chemical Society: Washington, DC: Chicago, IL, United States, 2001.

96. Medina-Franco, J. L. Scanning Structure-Activity Relationships with Structure-Activity Similarity and Related Maps: From Consensus Activity Cliffs to Selectivity Switches. *J. Chem. Inf. Model.* **2012**, *52*, 2485–2493.

97. Méndez-Lucio, O.; Pérez-Villanueva, J.; Castillo, R.; Medina-Franco, J. L. Identifying Activity Cliff Generators of PPAR Ligands Using SAS Maps. *Mol. Informatics* **2012**, *31*, 837–846.

98. Patterson, D. E.; Cramer, R. D.; III, F.; Clark, R. D.; Weinberger, L. E. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J Med Chem* **1996**, *39*, 3049– 3059.

99. Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.

100. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

101. Ashton, M.; Barnard, J.; Casset, F.; Charlton, M.; Downs, G.; Gorse, D.; Holliday, J.; Lahana, R.; Willett, P. Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions. *Quant. Struct.-Act. Relationships* **2002**, *21*, 598–604.

List of papers published during the PhD period

Dates: 07/2010 – 06/2013

1) Sahigara, F.; **Mansouri, K.**; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. *Molecules* **2012**, *17*, 4791–4810.

2) **Mansouri, K.**; Consonni, V.; Durjava, M. K.; Kolar, B.; Öberg, T.; Todeschini, R. Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling. *Chemosphere* **2012**, *89*, 433–444.

3) **Mansouri, K.**; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure-Activity Relationship Models for Ready Biodegradability of Chemicals. *J. Chem. Inf. Model.* **2013**, *53*, 867–878.

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n. 238701 of Marie Curie ITN Environmental Chemoinformatics (ECO) project.

Annex I: Bioaccumulation of PBDEs

Mansouri, K.; Consonni, V.; Durjava, M. K.; Kolar, B.; Öberg, T.; Todeschini, R. Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling. *Chemosphere* **2012**, *89*, 433–444.

Chemosphere 89 (2012) 433–444

Contents lists available at SciVerse ScienceDirect

Chemosphere

journal homepage: www.elsevier.com/locate/chemosphere

Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling

Kamel Mansouri [a], Viviana Consonni [a], Mojca Kos Durjava [b], Boris Kolar [b], Tomas Öberg [c], Roberto Todeschini [a,*]

[a] Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.zza della Scienza 1, 20126 Milano, Italy
[b] Public Health Institute Maribor, Center for Risk Assessment of Chemicals with Laboratory, Maribor, Prvomajska 1, Slovenia
[c] School of Natural Sciences, Linnaeus University, 391 82 Kalmar, Sweden

HIGHLIGHTS

- ▶ We collected experimental data on bioaccumulation of polybrominated diphenyl ethers.
- ▶ We analyzed the datasets and checked for outliers.
- ▶ We model Bioconcentration factor for oligochaetes, bioaccumulation factor and biomagnification factor for fish.

ARTICLE INFO

Article history:
Received 10 February 2012
Received in revised form 13 April 2012
Accepted 19 May 2012
Available online 16 June 2012

Keywords:
PBDEs
Bioaccumulation
BCF
BAF
BMF
QSAR

ABSTRACT

Polybrominated diphenyl ethers (PBDEs) are used as flame retardants in textiles, foams and plastics. Highly bioaccumulative with toxic effects including developmental neurotoxicity estrogen and thyroid hormones disruption, they are considered as persistent organic pollutants (POPs) and have been found in human tissues, wildlife and biota worldwide. But only some of them are banned from EU market.

For the environmental fate studies of these compounds the bioconcentration factor (BCF) is one of the most important endpoints to start with. We applied quantitative structure–activity relationships techniques to overcome the limited experimental data and avoid more animal testing.

The aim of this work was to assess the bioaccumulation of PBDEs by means of QSAR. First, a BCF dataset of specifically conducted experiments was modeled. Then the study was extended by predicting the bioaccumulation and biomagnification factors using some experimental values from the literature. Molecular descriptors were calculated using DRAGON 6. The most relevant ones were selected and resulting models were compared paying attention to the applicability domain.

1. Introduction

Polybrominated diphenyl ethers (PBDEs) is a class of brominated flame retardants (BFRs). Over the last 30 years, for their interesting characteristic of delaying ignition and inhibition of fire, PBDEs have been used in textiles, polyurethane foams, upholstery stuffing for furniture and car seats, electronic components and plastics (de Wit, 2002). Due to the fact that these compounds are not chemically bound to the products, there is high risk of leaching into the environment (Hutzinger et al., 1976). The widespread use of these compounds increased their worldwide presence in various biotic and abiotic matrices (Mikula and Svobodova, 2006). Thus the need for toxicity investigations to determine potential adverse health and environmental effects associated with these compounds.

Studies have been conducted for a better understanding of the potential health risks of PBDEs. Some investigations revealed high concerns about liver toxicity, thyroid toxicity, developmental and reproductive toxicity (ATSDR, 2004). PBDEs are sharing some toxicological properties with other structurally similar polyhalogenated aromatic compounds, particularly PCBs; however their toxicity still varies by congeners due to their different chemical structures (Pijnenburg et al., 1995; Eriksson et al., 1998; McDonald, 2002; Palm et al., 2002). In particular, the presence of bromines in ortho position of the oxygen bridge presents a barrier to rotation that would prevent the two aromatic rings from assuming a fully coplanar configuration; this has implications on dioxin-type toxicities which are mediated by the aryl hydrocarbon receptor (AhR) pathway (ATSDR, 2004). In addition to the direct effects of PBDEs, another way of possible exposure as a secondary poisoning is the formation of brominated dibenzo-p-dioxins and dibenzofurans during incomplete combustion or other high temperature

* Corresponding author. Tel.: +39 02 64482820; fax: +39 02 64482890.
E-mail address: roberto.todeschini@unimib.it (R. Todeschini).

0045-6535/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chemosphere.2012.05.081

processes and recycling activities (Watanabe and Tatsukawa, 1987; Weber and Kuch, 2003). As a prevention of these risks, four BDE groups (tetra, penta, hexa and hepta) were added to the list of POPs in Appendix A of the Stockholm convention (UNEP, 2009).

The environmental fate of these xenobiotic compounds is governed by their physical and chemical properties and their propensity for biotic and abiotic transformation. In general, BDE congeners are relatively hydrophobic and lipophilic compounds that have low water solubility and low vapor pressures. Thus, air and water are primary transport media for PBDEs to end up in soils and sediments that are the most polluted. PBDEs are strongly adsorbed to soil, and adsorption increases with the bromination of PBDEs and organic carbon content of soil and sediment (ATSDR, 2004). The hydrophobic and lipophilic properties of PBDEs cause their bioaccumulation in aquatic organisms by exposure within their food web. PBDE congeners are also found in a wide variety of avian species, insects, and terrestrial mammals arriving to humans as it was detected in blood and breast milk (Noren and Meironyte, 2000; Patterson et al., 2000; Sharp and Lunder, 2004; Jakobsson et al., 2005).

In the environment (atmosphere, water, and biota), levels of PBDEs tend to be dominated by lower brominated congeners that are highly toxic, bioaccumulative and persistent (ATSDR, 2004). This is possibly due to the debromination of higher brominated BDEs (Env. Canada, 2004). The low brominated BDEs (1–5) are likely to be carcinogenic and endocrine disrupter toxicants (Fernlof et al., 1997; Darnerud et al., 1998; Eriksson et al., 2001). Higher brominated congeners (e.g., decaBDE) are typically detected only near point sources (Wania and Dugani, 2002).

PBDEs are mainly discharged into the air from production, use, and recycling of PBDE-treated plastics, electronics, textiles, and polyurethane foams. They are discharged into surface waters and soil from industrial activity and sewage treatment plants. The land disposal of sewage and industrial sludge also contribute to environmental loadings (USEPA, 2010).

Certain biotic and abiotic processes can transform PBDEs in the environment, the most important processes include biodegradation, biotransformation, and photolysis. Biodegradation is breaking the chemical structures by aerobic microbes, via oxidative dehalogenation, and anaerobic microorganisms (He et al., 2006; Kim et al., 2007). Biotransformation is the conversion of the congeners through metabolic pathways by removal of bromine atoms (Stapleton et al., 2002; Stapleton et al., 2004a,b,c; Tomy et al., 2004). Photolysis involves the breakdown of PBDEs by the action and the energy of sunlight (Watanabe and Tatsukawa, 1987; Fang et al., 2008; Stapleton and Dodder, 2008).

In addition to physico-chemical properties, the bioaccumulation is an important endpoint to assess the environmental fate of PBDEs. Bioconcentration factor (BCF), bioaccumulation factor (BAF) and biomagnification factor (BMF) are distinct measurements to study bioaccumulation of POPs (Arnot and Gobas, 2006). BAF describes a process whereby an organism acquires a body burden of a chemical in relation to contact through all possible pathways of exposure (i.e., dietary absorption, transport across the respiratory surface, dermal absorption, and inhalation (Gobas and Morrison, 2000). Unlike bioaccumulation, bioconcentration is the process by which a chemical substance is absorbed by an organism from the ambient environment only through its respiratory and dermal surfaces. While for BCF dietary exposure is not included; the biomagnification is expressing the ratio between the chemical concentration in the diet and in the organism. For chemicals that are known to enter ecological magnification in food webs, field BAFs tend to be greater than the BCFs from laboratory experiments that do not include dietary exposure (Arnot and Gobas, 2006). Both BCF and BAF are calculated by the ratio of the chemical concentration in the organism and its concentration in the water at steady state. However, BCF can only be measured under controlled laboratory conditions in which dietary intake of the chemical are deliberately not included.

BCF is consistently used for assessing bioaccumulation potential by regulatory agencies as a part of Persistent, Bioaccumulative, and Toxic (PBT) assessment programs (Arnot and Gobas, 2006). This factor and other bioaccumulation criteria (Supplementary information, Table SI.1) were also used for the development of environmental standards and guidelines (Walker and Gobas, 1999; USEPA, 2000).

One of the problems faced in assessing bioaccumulation of POPs, in addition to uncertainty with empirical values, is the lack of experimental data. A case study of organic chemicals on the Canadian Domestic Substances List indicates that empirical data are available for less than 4% of the chemicals that require evaluation. 76% of these chemicals have less than three acceptable quality BCF or BAF values (Arnot and Gobas, 2006). For PBDEs, only few values for some commercial mixtures are available in the literature. To fill the data gaps, we used quantitative structure–activity relationships (QSARs) that are becoming more and more recognized by scientific community and legislators as an alternative to animal testing and high costs of experiments.

In this study we aimed to predict BCF of PBDEs for aquatic organisms in the framework of the partnership between the two EU projects related to REACH (Registration, Evaluation, Authorization and Restriction of Chemicals) Environmental Chemoinformatics (ECO) and Case Studies on the Development and Application of in-Silico Techniques for Environmental hazard and Risk assessment (CADASTER). We first modeled BCF for oligochaetes from experimental data collected for this task, and then the study was extended by investigating and modeling BAF and BMF for aquatic organisms from the literature.

2. Materials and methods

Here we used the same numbering system for PCDEs and PBDEs as assigned to the PCB congeners by Ballschmiter and the International Union of Pure and Applied Chemistry (IUPAC) (Ballschmiter et al., 1992). PBDEs are compounds with a common structure of a diphenyl ether molecule having from 1 to 10 bromine atoms attached. Depending on the location and number of bromine atoms, there are 209 possible PBDE congeners. This similarity of congeners is making the QSAR study possible as the QSAR models are valid for similar compounds.

Experimental values from different sources were collected and evaluated to select the most suitable dataset for this study. Principal component analysis (PCA) was used to explore the data and regression methods, ordinary least squares (OLSs) and partial least squares (PLSs), were applied to obtain predictive QSAR models. Unlike OLS, PLS-regression can model data with collinear and noisy variables. It finds fundamental relations between the matrix of descriptors (*X*) and the response (*Y*) using new fewer variables called latent variables (LVs) that are orthogonal and explaining the maximum variance in the *Y* space (Wold et al., 2001).

In the modeling step, 167 descriptors based on the molecular topology were calculated by using the software DRAGON 6 (Talete, srl). After automatically removing constant and near constant variables, using the same software, descriptor pairwise correlation was checked with a fixed threshold of 95% to avoid the problem of multicollinearity (Slinker and Glantz, 1985; Miles and Shelvin, 2001). 30 descriptors were left for the next step of variable selection to catch the most predictive descriptors. This operation was performed by using genetic algorithms (GAs) (Leardi and Lupianez, 1998; Ballabio et al., 2011). The used GAs start from an initial random population of chromosomes. Each chromosome is a binary string (genes) that represents the presence/absence of the variable

Table 1
BCF for the 21 detected PBDEs in *Tubifex tubifex*.

Congeners	C_{org}[a] (μg/kg$_{ww}$)	C_{wat}[b] (ng L^{-1})	BCF (L kg^{-1})	logBCF[c]
BDE-28	1.2	0.094	12 860	5.83
BDE-47	77	5.9	13 060	5.84
BDE-51	0.93	0.065	14 290	5.88
BDE-66	4.12	0.195	21 080	6.05
BDE-77	60.3	1.69	36 760	6.29
BDE-99	87.9	4.55	19 320	6.01
BDE-100	46	1.04	44 430	6.36
BDE-119	0.384	0.014	28 100	6.29
BDE-126	45.3	0.89	50 740	6.43
BDE-153	9.73–15	0.394–1.03	24 680–14 627	6.13
BDE-154	13.85–3.36	0.405–0.29	34 160–11 582	6.15
BDE-180	0.21	0.14	1457	4.88
BDE-183	0.02–7.13	0.157–4.74	128–1505	4.36
BDE-197	7.93	2.23	3558	5.30
BDE-198	19.4	2.14	9080	5.68
BDE-203	2.68	0.38	7108	5.93
BDE-204	12.5	2.6	4800	5.75
BDE-206	0.73	0.21	3500	5.44
BDE-207	1.88–0.89	0.75–0.23	2496–3860	5.75
BDE-208	0.56	0.08	7130	5.57
BDE-209	16.9	5.8	2910	5.19

[a] Concentration in *Tubifex tubifex* on day 28 of uptake phase, based on wet weight.

[b] Concentration in water on day 28 of uptake phase, based on wet weight.

[c] Lipid content based logBCF. The fraction of the lipid weight was 1.9%. When multiple tests were available the mean is calculated on the log values.

in a model by maximizing a defined fitness function. Afterwards, an evolution process is simulated and new chromosomes are obtained by coupling the chromosomes of the initial population with genetic operations (crossover and mutation). The evolution process is repeated 30 times and the whole process is repeated for 100 runs. The fitness function used was Q^2 calculated in validation with 5-fold groups (Consonni et al., 2009, 2010).

The GAs provided a set of optimal models that were subject to a comparison in order to minimize the number of used descriptors, the number of latent variables for PLS models and the number of congeners outside the applicability domain (AD). The model with the best compromise between these criteria was selected. The evaluation of the AD was investigated using the leverage approach with a threshold of three times the average of leverages from the matrix of the used descriptor values of the training set. The calculations were performed in MATLAB 7 (MathWorks, Inc.).

2.1. Uncertainty

When dealing with experimental data we have to be aware of its uncertainty. For BCF a data quality assessment found that 45% of values are subject to at least one major source of uncertainty (Arnot and Gobas, 2006). There are two measurements to take: the concentration in water and concentration in the organism. Uncertainty related to water concentration comes from assumption if not measured, fluctuations in the water concentration

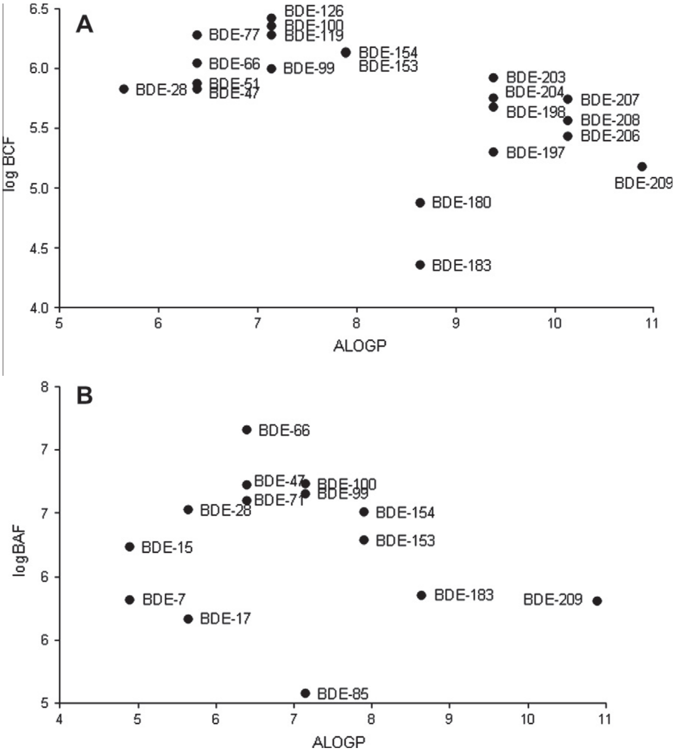


Fig. 1. logBCF versus ALOGP for selected PBDE congeners (A). Average logBAF values from Table 3 versus ALOGP (B).

during exposure (Gobas and Zhang, 1992), adsorption of chemicals with low aqueous solubility by surfaces of testing equipment or partition into the air if Henry's Law constant is high. The use of radio-labeled test chemicals to quantify concentrations and transformation of the chemical in the water phase may also contribute to errors in calculating the BCF if the parent compound is transformed and the metabolite with the radio-label is not eliminated from the organism (Goodrich et al., 1991; Toshima et al., 1992). Sources of uncertainty of chemical's concentration in the organism are the exposure period in relation to the steady state, the mass of the organism and its lipid content (Nichols et al., 1990), the temperature of exposure and other sources (Arnot and Gobas, 2006). Applying standard testing guidelines such as OECD (1984-315, 1996-305, and 2008-315) can contribute in reducing uncertainty in the measured data and to provide confidence in the quality of the data used to model the bioaccumulation assessments.

2.2. Bioconcentration factor

In order to assess bioaccumulation of PBDEs targeting aquatic ecosystems, first we considered a dataset of experiments that were conducted in the framework of the European project CADASTER, to calculate BCF for tubificid oligochaetes *Tubifex tubifex* (OECD 315). These aquatic organisms are relevant for our study as it is food for fish and are living in surficial sediments where most BFRs tend to concentrate. Three commercial mixtures from Wellington Laboratories were used: TBDE-71 for penta-BDE, TBDE-79 for octa-BDE and TBDE-83R for deca-BDE. Then some individual congeners from the same supplier were added for testing (PBDE-002, PBDE-077, PBDE-126, PBDE-198 and PBDE-204). During the experiments it was analytically difficult to separate the two tetrabrominated congeners BDE-42 and BDE-66. Konstantinov et al. (2008) were able to identify all BDEs in the mixture TBDE-71 by means of 1H NMR and GC/MS and they reported that the weight percentage of BDE-66 is 51 times higher than that of BDE-42; thus, it was assumed the absence of BDE-42. Table 1 collects the obtained results obtained at steady state. If the steady state was not reached during the uptake phase, the BCF was calculated for day 28.

Before considering this dataset to model BCF, we first had to check it for self consistency. It is known since early studies on BCF that it is correlated with octanol–water partition coefficient (Kow) (Mackay, 1982; Isnard and Lambert, 1988) and compounds with a log Kow above 4.5 are considered as bioaccumulating in a regulatory PBT assessment (ATSDR, 2004; USEPA, 2010). For PCBs and PBDEs Kow is increasing with the number of halogens (Hawker and Connell, 1988; Braekevelt et al., 2003). Moreover, this behavior is also observed for other groups of halogenated aromatic compounds (Kuramochi et al., 2007). All these compounds are lipophilic, and octanol is generally considered to be a reasonable surrogate phase for lipids in biological organisms, thus, BCF should correlate well with Kow (Fisk et al., 1998; Dimitrov et al., 2002). Higher bioaccumulation potential for PBDEs is generally noticed compared to PCBs with similar Kow (Stapleton et al., 2004c; Wang et al., 2007). But there is a general bilinear correlation pattern for the BCF with log Kow reported in several publications (Kannan et al., 1998; Meylan et al., 1999; Wania and Dugani 2002; Wang

Table 2
Lipid based logBAF of different aquatic species in log (L kg^{-1}).

Congeners	Japanese sea bass[a]	Ureogenic goby[a]	Green crab[a]	Grapsid crab[a]	Giant Pacific oyster[a]	Asian green mussel[a]	Black-striped mussel[a]	Std. dev.
BDE-7	–	–	–	6.18	–	–	5.46	0.51
BDE-15	6.32	6.15	6.29	6.32	6.25	6.20	6.14	0.08
BDE-17	6.33	5.76	5.52	4.59	5.90	5.92	5.72	0.54
BDE-28	6.79	6.37	6.29	6.65	6.59	6.55	6.48	0.17
BDE-47	6.90	6.86	7.04	6.70	6.57	6.57	6.47	0.21
BDE-66	7.37	–	6.74	7.22	7.32	7.19	7.14	0.22
BDE-71	6.99	6.36	6.44	6.15	6.88	6.77	6.65	0.30
BDE-85	–	–	–	5.16	5.22	5.12	4.83	0.17
BDE-99	6.64	–	7.22	6.61	6.56	6.68	6.24	0.32
BDE-100	6.87	7.01	6.65	6.75	6.73	6.70	6.44	0.18
BDE-153	6.52	–	6.81	6.63	5.70	5.93	6.15	0.43
BDE-154	6.71	6.81	6.59	6.49	6.42	6.07	6.50	0.24
BDE-183	–	–	5.90	–	–	–	5.82	0.06
BDE-209	–	5.78	5.12	–	5.50	6.05	6.64	0.58

Congeners	NorthernPikeminnow[b]	Cutthroat[b]	Smallmouth bass[b]	Peamouth[b]	Common carp[b]	Rainbowtrout[b]	Largescale sucker[b]	Mountain whitefish[b]	Lake Trout[b]	Std. dev
BDE-47	5.91	5.86	5.75	5.54	5.98	5.79	6.11	6.18	7.3	0.51
BDE-49	5.30	–	5.20	5.54	–	5.59	5.38	5.90	–	0.25
BDE-66	–	–	–	–	–	–	–	–	7.3	–
BDE-99	–	5.44	4.98	–	–	6.00	–	6.43	6.7	0.71
BDE-100	5.68	5.56	5.28	–	6.08	6.00	6.11	6.45	7.5	0.68
BDE-153	5.11	–	–	5.34	–	5.91	5.57	6.26	–	0.45
BDE-154	5.33	5.32	5.00	5.11	5.15	5.71	5.70	6.04	–	0.36

Congeners	Water snake[c]	Northern snakehead[c]	Mud carp[c]	Crucian carp[c]	Prawn[c]	Chinese mysterysnail[c]	Std. dev
BDE-28	3.10	3.90	4.17	3.62	3.49	2.94	0.47
BDE-47	4.68	4.46	4.57	4.30	4.05	3.27	0.51
BDE-99	4.57	1.91	3.23	2.61	3.65	3.72	0.93
BDE-100	5.11	4.89	4.67	4.36	4.42	3.67	0.50
BDE-138	4.10	4.00	4.03	4.46	4.24	–	0.19
BDE-153	5.51	3.86	5.18	4.26	4.44	4.14	0.64
BDE-154	5.06	4.85	4.91	4.37	4.33	3.79	0.48
BDE-209	4.50	3.32	5.53	4.06	4.56	–	0.80

[a] Aquatic species from Tokyo Bay, Japan (Mizukawa et al., 2009).

[b] Aquatic species from North American lakes and rivers (Johnson et al., 2006; Streets et al., 2006).

[c] Aquatic species samples from an e-waste recycling site, China (Jiang-Ping et al., 2008).

et al., 2007; Jiang-Ping et al., 2008; Mizukawa et al., 2009). BCF and Kow in the range up to log Kow of 7 are positively correlated: for low brominated congeners (up to penta-BDE) BCF is increasing from ~10^5 to ~10^7. This increase in BCF with bromination number can be explained by equilibrium partitioning due to hydrophobicity, because higher-brominated congeners are more hydrophobic (Meylan et al., 1999). Conversely, at log Kow above 7, BCFs of both PBDEs and PCBs decrease as Kow increase further. This can be partly explained by the steric hindrance that restricts permeation of the large and very hydrophobic organic compounds through the cell membrane (Shaw and Connel, 1982). Also the low water solubility and binding to surfaces would cause these compounds to bind to dissolved organics and particles (Moermond et al., 2005).

As there are no available experimental values of log Kow covering all selected congeners, we plot logBCF against ALOGP which is an estimation of log Kow (Ghose and Crippen, 1986). Fig. 1A is showing similar trend for oligochaetes as described in the literature, also the same threshold of about 7 over which BCF decreases as log Kow increases further.

Nevertheless, the two hepta-brominated BDEs (BDE-180 and BDE-183) disagree with the general pattern and showed low BCF values. This behavior can be probably due to the cited sources of uncertainty related to experiments. However, the biological processes in organisms, such as different metabolic rates of these chemicals, could influence this general correlation trend predicted by Kow. Chemical-specific, species-specific and site-specific bioaccumulative potential and their correlations with Kow were found in field studies, largely due to different metabolic rates of chemicals which were dependent on both the chemical structure and the metabolic capacity of the organism (Streets et al., 2006; Wang et al., 2007). The difference of the BCF values in the two congeners can be related to the structural dissimilarity especially the number of bromines in the ortho-position. But, for BDE-183, if we consider only the BCF value (4.9) from the second test then the two congeners will have the same BCF. Knowing that they had different concentrations in the water and in the lipid content of the organisms, this substantially diminishes the probability of an experimental error. Different studies have indicated that BDEs are metabolized in fish. Stapleton et al. (2004c) dedicated an experimental study to biotransformation of BDE 99 and 183 and suggested that this last one is debrominated to form two different hexabrominated congeners. Jiang-Ping et al. (2008) also attributed many data points that diverged from the general trend predicted by Kow in some aquatic species to metabolism; those showing low bioaccumulative rate by being metabolized and highly bioaccumulative by being the metabolic products of higher brominated congeners. BDE-99 and BDE-197 were also a bit below the general trend with lower BCF than congeners with similar Kow.

Whether it was due to experimental uncertainty or to biological metabolism, the situation of BDE-180 and BDE-183 is specific to this dataset of 21 BCF values for oligochaetes. So including them will probably affect the final QSAR model as our target is a generalized idea about PBDEs' BCF for aquatic biota. Also, from a regulatory perspective, it is not interesting to include biotransformation in the model. That's why we decided to remove them from our training set as outliers.

2.3. Bioaccumulation factor

Several studies sampled and measured PBDEs' concentrations in some aquatic species and water. In field conditions, the organisms are exposed to all possible contamination pathways including diet so the concentrations ratio is considered as BAF. Field BAF will also be different from laboratory BCF by routes and duration of uptake. Experimental values of BAFs for PBDEs were collected from different literature sources, where information about the lipid content concentrations was available.

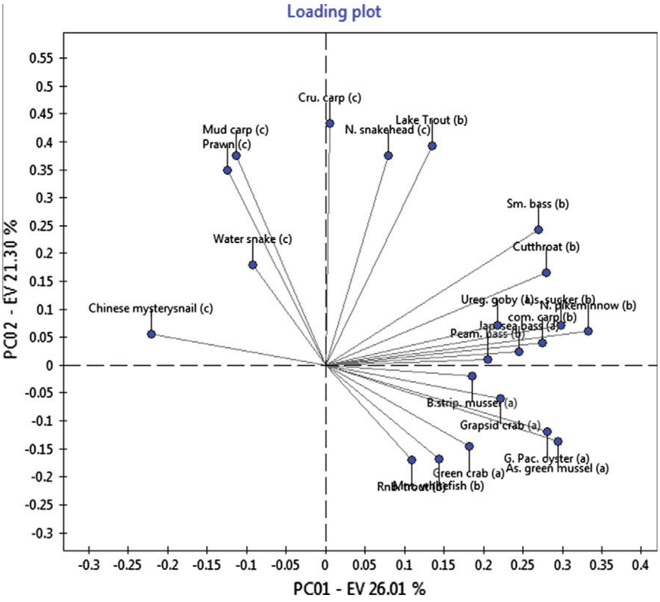


Fig. 2. Loadings plot of the studied species' PCA. Data from; Table 2a, Table 2b and Table 2c.

Table 3
Selected models of logBCF, logBAF and BMF.

Model	Endpoint	Descriptors	LVs	R^2%	RMSE	Q^2%(5f) cv	RMSE (5f)	No. outside AD
MLR-1	BCF	3	–	76.1	0.168	73.3	0.177	17
MLR-2	BCF	4	–	92.3	0.108	86.1	0.146	11
MLR-3	BCF	3	–	86.8	0.142	81.1	0.170	6
MLR-4	BCF	2	–	85.5	0.149	80.0	0.175	4
MLR-5	BCF	2	–	86.7	0.142	81.6	0.167	6
MLR-6	BAF	3	–	91.9	0.122	84.5	0.169	62
MLR-7	BAF	4	–	95.0	0.095	87.0	0.154	50
MLR-8	BMF	4	–	91.5	2.690	87.4	3.268	134
MLR-9	BMF	4	–	90.4	2.855	83.0	3.806	96
MLR-10	BMF	4	–	91.0	2.760	80.4	4.083	36
PLS-1	BAF	5	3	94.3	0.103	91.3	0.127	38
PLS-2	BAF	6	3	95.6	0.090	93.4	0.110	67

Values in Table 2a were obtained from concentrations in coastal organisms and water collected from the northwest head of Tokyo Bay, Japan in September 2005 (Mizukawa et al., 2009). Table 2b summarized logBAF values calculated from the fish fillet and water data collected in 2005–2006 from different North American lakes and rivers, these being the Spokane River, Lower Columbia River, Yakima River, Lake Washington and all five Great Lakes (Johnson et al., 2006; Streets et al., 2006). Data from an electronic waste recycling site located in Longtang Town, Qingyuan City, in South China sampled in 2006 was used to calculate logBAF of Table 2c (Jiang-Ping et al., 2008).

To check if these data are useful to model the BAF for aquatic species and whether it is possible to merge it or not, a principal component analysis of the present species using the software DRAGON 6 (Talete, srl) was performed. The analyzed matrix is constituted by 16 samples (PBDE congeners) and 22 variables (species). Fig. 2 is the loading plot of the first two principal components (PCs) where the relationships between the different aquatic species can be highlighted. The first two PCs, explaining a total variance of 47.3%, are showing that the data from the different sources gave different information and only the Japanese and the North American data can be studied together. The missing values are filled with the mean value of the corresponding variable. This divergence could be due to the different metabolism capacity between species and the different natural conditions studied.

Concerning the third dataset, the major reason that made it distant from the others is the low BAF values, possibly due to the higher water concentrations. The sampling site was surrounded by several e-waste recycling workshops where more than 80000 workers had been involved in the business of e-waste dismantling and recycling. Approximately 1.7 million tons of e-waste were annually dismantled causing high water pollution (Jiang-Ping et al., 2008). For the studied congeners, the mean water concentration of PBDEs in the Chinese site was 2.77 ng L^{-1} compared to 8.14 pg L^{-1} from the Japanese site also the standard deviation between species is relatively higher.

In highly polluted water, the PBDE concentrations in biota do not seem proportional with the water concentration. In addition, there was no correlation between the average values of the experimental data and ALOGP. So we decided to merge only the first two datasets to model aquatic species' BAF. The correlation profile of the average values from Table 2a with ALOGP (Fig. 1B) shows a decline of BDE-85 from the general pattern that can be related to metabolisation. In the literature, there was an indication about the debromination of this congener in fish (Tomy et al., 2004; Wolkers et al., 2004). Therefore, because of these uncertainties, this compound was removed from the dataset.

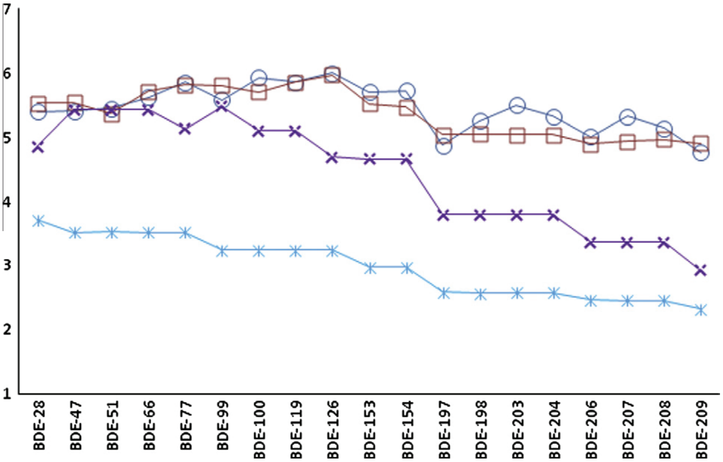


Fig. 3. Comparison of the predictions for the 19 PBDEs. Observed logBCF (○), predicted logBCF (□, this work), EpiSuite prediction (×), CAESAR prediction (∗).

2.4. Biomagnification factor

One possible way to avoid problems related to water concentrations is to study BMF. Tomy et al. (2004) studied bioaccumulation and biotransformation of PBDEs and calculated BMF for 13 different BDEs ranging from 3 to 10 brominated congeners. Table SI.2 is summarizing the low and high treatment food and the average values that were considered for the model. The studied aquatic specie was the lake trout (*Salvelinus namaycush*) that was exposed to spiked food for 56 d of uptake followed by 112 d for elimination. In this data, there were no clear trends in assimilation efficiencies and BMF with either bromine number or log Kow. A possible explanation is the long duration of the experiments that highly affected the bioaccumulation rates by biotransformation via debromination. Lower brominated congeners were created and there are no analytical methods to distinguish between the BDE that was bioaccumulated from BDE formed via debromination (Tomy et al., 2004). However, Burreau et al. (1997) showed that assimilation efficiencies of PBDEs had negative relationship with the number of bromines but not with log Kow.

BMFs were calculated using the equation: $BMF = \frac{\alpha \times F}{K_d}$ where α is the assimilation efficiency during uptake, F is the feeding rate and K_d the depuration rate constant.

3. Results and discussion

3.1. Bioconcentration factor

The CADASTER dataset generated for 19 PBDE congeners was used to build the models. As the congeners differ by the number of the bromines and their positions on the two rings, a multiple linear regression model of three variables was first built. It can also be considered as a non-linear model because it is using as descriptors the number bromine atoms (nBr), the number of bromines in ortho-position (F03[O–Br]) and the squared value of bromines (nBr^2) to deal with the observed bilinear correlation of logBCF and log Kow (MLR-1 of Table 3). It considered the number of bromines as it is correlated with log Kow and the bromines in ortho position as an homology with chlorines in ortho position and its

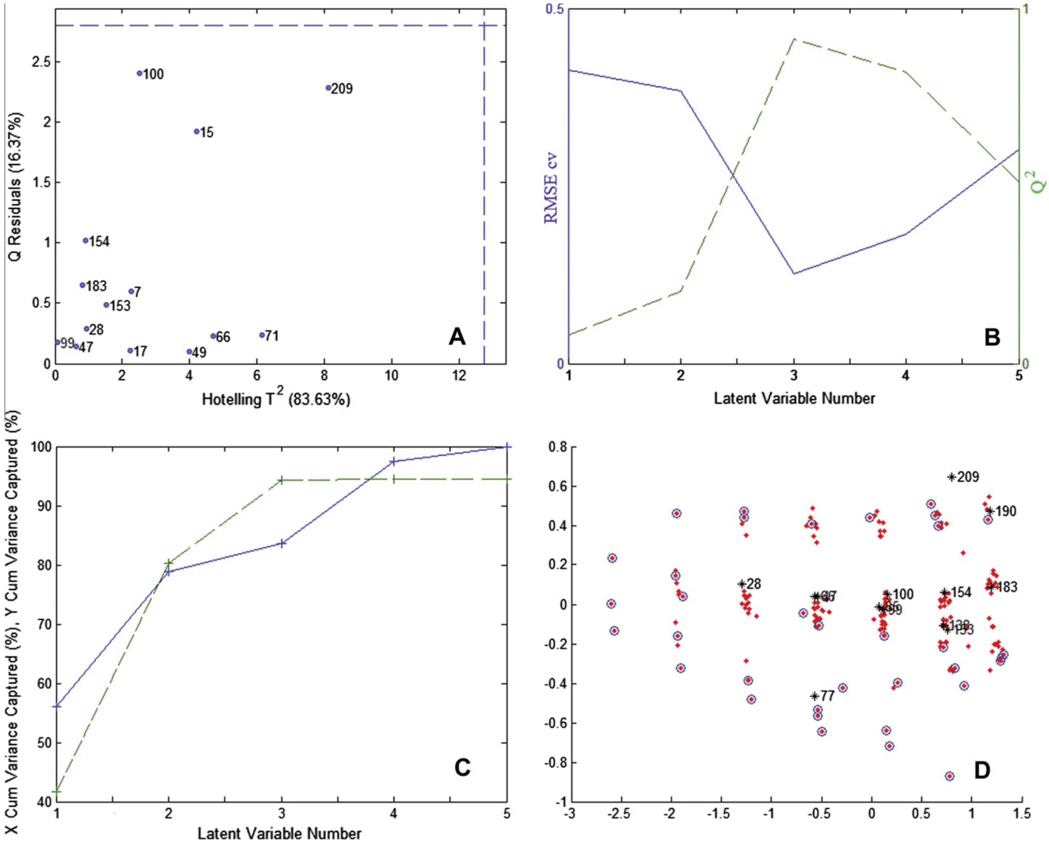


Fig. 4. Q Residuals versus Hotelling T^2 of the model PLS-2 (A). Latent variables selection of the model PLS-2: cross-validated RMSE (——) and Q^2 (— —) versus latent variable number (B). Latent variable selection of the model PLS-2: cumulated variance on X (——) and on Y (— —) versus latent variable number (C). Two dimensional MDS plot from the model MLR-10, training set with congeners numbers (∗), test set (■) and congeners outside the AD (○) (D).

Table A1
The predicted values for logBCF, logBAF and BMF. NR refers to not reliable predictions; *refer to congeners detected as outside the applicability domain.

Congeners	logBCF	logBAF	BMF	Congeners	logBCF	logBAF	BMF
BDE-1	5.69	5.22	8.77*	BDE-36	6.06	5.91	77.14*
BDE-2	5.86	5.25	45.21*	BDE-37	6.09	6.75	20.18
BDE-3	5.91	5.58	23.96*	BDE-38	6.07	6.26	36.89
BDE-4	5.47	5.10	NR*	BDE-39	6.07	6.25	60.46*
BDE-5	5.77	5.56	7.75	BDE-40	5.96	6.15	NR
BDE-6	5.77	5.62	16.66	BDE-41	5.97	6.25	NR
BDE-7	5.80	5.75	14.21	BDE-42	5.97	6.27	6.29
BDE-8	5.80	5.90	1.25*	BDE-43	5.96	5.17	18.51
BDE-9	5.77	4.87	20.71	BDE-44	5.96	5.35	7.91
BDE-10	5.61*	5.33*	NR*	BDE-45	5.79	4.63*	4.02
BDE-11	5.93	5.63	55.46*	BDE-46	5.79	5.54	NR
BDE-12	5.96	5.97	24.18	BDE-47	5.97	6.40	15.27
BDE-13	5.96	5.96	35.96*	BDE-48	5.97	5.47	6.29
BDE-14	5.93	5.48	67.24*	BDE-49	5.97	5.48	16.89
BDE-15	5.99	6.30	17.61	BDE-50	5.80	5.44	23.60
BDE-16	5.69	5.58	NR	BDE-51	5.80	5.60	13.00
BDE-17	5.71	5.70	7.13	BDE-52	5.96	4.55*	18.51
BDE-18	5.69	4.84	9.17	BDE-53	5.79	4.75	10.50
BDE-19	5.46	4.90	5.06	BDE-54	5.55	4.47*	25.57*
BDE-20	5.92	6.03	17.40	BDE-55	6.14	6.69	14.23
BDE-21	5.94	6.14	3.63	BDE-56	6.14	6.88	3.63
BDE-22	5.94	6.31	3.63	BDE-57	6.13	5.54	41.17
BDE-23	5.92	5.06	32.05	BDE-58	6.13	6.44	37.05
BDE-24	5.79	4.94*	NR	BDE-59	6.03	5.48	3.68
BDE-25	5.94	6.22	21.89	BDE-60	6.14	6.97	2.05
BDE-26	5.92	5.29	28.01	BDE-61	6.14	5.66	18.35
BDE-27	5.79	5.84	2.40	BDE-62	6.03	5.36	6.18
BDE-28	5.96	6.49	9.26	BDE-63	6.14	5.82	28.95
BDE-29	5.94	5.42	15.41	BDE-64	6.03	5.69	NR
BDE-30	5.80	5.74	15.09*	BDE-65	6.03	4.32*	11.84
BDE-31	5.94	5.57	15.41	BDE-66	6.14	7.08	8.60
BDE-32	5.80	6.05	NR*	BDE-67	6.14	5.92	24.83
BDE-33	5.94	6.38	NR	BDE-68	6.14	6.64	41.99*
BDE-34	5.92	5.95	35.67	BDE-69	6.03	6.37	19.22
BDE-35	6.07	6.40	36.89	BDE-70	6.14	6.08	14.23
BDE-71	6.03	6.68	NR*	BDE-107	6.34	6.42	28.39
BDE-72	6.13	5.64	47.65*	BDE-108	6.34	7.24	34.95
BDE-73	6.03	6.32	16.72	BDE-109	6.29	6.02	10.39
BDE-74	6.14	6.19	12.65	BDE-110	6.29	6.35	NR
BDE-75	6.04	6.57	11.04	BDE-111	6.34	5.98	59.05*
BDE-76	6.14	6.79	10.18	BDE-112	6.29	4.88*	15.69
BDE-77	6.24	7.28	24.82	BDE-113	6.29	5.99	18.12
BDE-78	6.23	6.76	51.76*	BDE-114	6.34	6.53	17.77
BDE-79	6.23	6.77	62.37*	BDE-115	6.29	6.23	3.81
BDE-80	6.23	6.25	98.62*	BDE-116	6.29	4.76*	9.13
BDE-81	6.24	7.11	35.42	BDE-117	6.29	5.08	9.13
BDE-82	6.23	6.93	NR	BDE-118	6.34	6.80	13.72
BDE-83	6.23	5.76	17.64	BDE-119	6.29	7.36	5.07
BDE-84	6.13	5.30	0.38	BDE-120	6.34	6.38	44.38
BDE-85	6.23	7.07	5.77	BDE-121	6.29	6.99	31.68*
BDE-86	6.23	5.88	7.01	BDE-122	6.34	7.41	14.92
BDE-87	6.23	6.05	7.01	BDE-123	6.34	7.62	19.03
BDE-88	6.13	5.18	12.64	BDE-124	6.34	6.52	24.35
BDE-89	6.13	6.33	NR	BDE-125	6.29	7.28	0.00
BDE-90	6.23	5.89	25.86	BDE-126	6.40	7.75*	41.15*
BDE-91	6.13	5.35	12.64	BDE-127	6.40	7.23	75.92*
BDE-92	6.23	4.87*	27.07	BDE-128	5.95	7.93*	NR
BDE-93	6.13	4.12*	13.85	BDE-129	5.95	6.63	8.06
BDE-94	6.13	5.24	15.11	BDE-130	5.95	6.62	17.48
BDE-95	6.13	4.41*	9.80	BDE-131	5.90	6.00	8.20
BDE-96	5.97*	4.30*	21.55	BDE-132	5.90	6.14	NR
BDE-97	6.23	6.06	7.01	BDE-133	5.95	5.32	35.94
BDE-98	6.13	6.21	18.02	BDE-134	5.90	4.82	8.99
BDE-99	6.23	6.19	15.19	BDE-135	5.90	4.93	14.37
BDE-100	6.13	6.27	30.25	BDE-136	5.81	4.05*	16.53
BDE-101	6.23	5.17	16.44	BDE-137	5.95	6.77	16.65
BDE-102	6.13	5.46	8.60	BDE-138	5.95	6.92	7.23
BDE-103	6.13	5.33	27.45	BDE-139	5.90	6.06	20.91
BDE-104	5.97*	5.12*	43.20*	BDE-140	5.90	7.19	16.79
BDE-105	6.34	7.68	4.30	BDE-141	5.95	5.61	17.48
BDE-106	6.34	6.26	28.39	BDE-142	5.90	4.70	12.32
BDE-143	5.90	6.11	4.15	BDE-178	5.68	4.36*	20.01
BDE-144	5.90	5.00	17.62	BDE-179	5.64	3.56*	20.60
BDE-145	5.81	4.99	29.16	BDE-180	5.65	6.61	17.00

Table A1 (*continued*)

Congeners	logBCF	logBAF	BMF	Congeners	logBCF	logBAF	BMF
BDE-146	5.95	5.63	25.73	BDE-181	5.67	5.66	19.42
BDE-147	5.90	4.88	21.74	BDE-182	5.67	7.14	19.43
BDE-148	5.90	5.98	31.25	BDE-183	5.67	5.95	15.30
BDE-149	5.90	5.14	8.20	BDE-184	5.63	5.97*	44.68*
BDE-150	5.81	5.01	38.59*	BDE-185	5.67	4.41*	15.59
BDE-151	5.90	3.82*	18.41	BDE-186	5.64	4.67*	24.36
BDE-152	5.81	3.91*	25.95	BDE-187	5.67	4.57*	15.59
BDE-153	5.95	5.94	16.65	BDE-188	5.64	4.71*	40.85*
BDE-154	5.90	6.20	25.04	BDE-189	5.64*	8.18*	31.06
BDE-155	5.81	5.97*	59.59*	BDE-190	5.71	6.60	2.34
BDE-156	5.99	7.30	20.56	BDE-191	5.71	8.07*	2.42
BDE-157	5.99	8.41*	16.44	BDE-192	5.71	6.26	24.50*
BDE-158	5.99	7.08	NR	BDE-193	5.71	6.56	6.83
BDE-159	5.99	6.86	48.48*	BDE-194	5.39*	7.67*	15.91
BDE-160	5.99	5.47	13.10	BDE-195	5.46	6.95	2.89
BDE-161	5.99	6.73	23.79*	BDE-196	5.46	7.12	7.02
BDE-162	5.99	7.01	39.06	BDE-197	5.47	5.99	28.77
BDE-163	5.99	5.79	4.86	BDE-198	5.48	5.26	14.81
BDE-164	5.99	7.03	NR	BDE-199	5.48	5.41	6.56
BDE-165	5.99	5.44	28.62*	BDE-200	5.49	4.40*	15.99
BDE-166	5.99	5.67	8.11	BDE-201	5.49	4.42*	24.24
BDE-167	5.99	7.43	24.68	BDE-202	5.50	2.85*	19.78
BDE-168	5.99	8.18*	9.45*	BDE-203	5.46	5.48	11.14
BDE-169	5.99	8.40*	58.80*	BDE-204	5.47	5.84	37.02*
BDE-170	5.65	7.78*	8.75	BDE-205	5.46*	7.81*	1.64
BDE-171	5.67	7.14	7.06	BDE-206	5.32	6.97	NR
BDE-172	5.66	6.28	25.54	BDE-207	5.36	5.87	17.15
BDE-173	5.67	5.61	7.35	BDE-208	5.39	3.75*	11.25
BDE-174	5.67	5.88	3.23	BDE-209	5.33	5.74	0.09
BDE-175	5.67	5.75	19.72				
BDE-176	5.64	4.83	24.36				
BDE-177	5.67	5.76	7.35				

relation to toxicity and bioaccumulation in the case of PCBs (Safe, 1990; Leonards et al., 1998). Since the number of possible bromines is 10, considering these variables this first model predicted 34 different values for the 209 congeners. This model presented good statistics and was used as a benchmark for the genetic algorithm results.

After that we applied the genetic algorithm to select the models with the best statistics and in the same time minimizing the number of variables and the congeners outside the applicability domain. Table 3 lists some of the models with best statistics and good coverage for the 209 PBDEs. 5-fold cross validation was used to validate the models.

The model with the best statistics is MLR-2 (Q^2 = 86.1%), but it used a higher number of variables than MLR-1 that is considered as a benchmark. The number of the compounds outside the AD varied with the models because using different variables is associated with different distribution of the data and thus different coverage of the training set to the chemical space. The model with the best compromise between the prediction ability, the number of variables and the number of compounds outside the AD is MLR-5. It is using two variables: the centered Moreau-Broto autocorrelation weighted by polarizability (ATSC6p) (Moreau and Broto, 1980a,b) and the frequency of Br–Br bonds at topological distance 6 (F06[Br–Br]). The equation of the model to predict logBCF with the lipid content correction is:

$$logBCF = -0.3882 \times ATSC6p - 0.0486 \times \mathrm{F06[Br{-}Br]} + 7 \qquad (1)$$

In order to make a comparison of the predictions obtained by the selected MLR model with existing models for BCF, EpiSuite's model (BCFBAF), widely used for risk assessment, and CAESAR BCF model were selected (Meylan et al., 1999; Lombardo et al., 2010). As both models from the literature are predicting BCF for fish and not stating that their used data were lipid content normalized we proceeded to a correction of 5% lipid content as suggested by OECD guideline 305 for testing BCF in fish. To make the comparison possible we plotted the experimental and predicted values for the oligochaetes in a 5% lipid content basis instead of 1.9%. Fig. 3 shows that both models were underestimating logBCF values for PBDEs even after the lipid content correction. This can be explained by the fact that both models had not any BDEs included in their training sets. In the prediction report CAESAR declared that its predictions might be associated with low reliability in the case of PBDEs while EpiSuite did not mention anything about the applicability domain.

The higher experimental logBCF values for BDE-203 and BDE-207 compared to the predicted values can be explained by the debromination of higher brominated congeners such as BDE-209 209 since the specie specific metabolisation is not included in the model. The predictions for all PBDEs were presented in Appendix A and congeners outside the AD are marked by asterisks.

3.2. Bioaccumulation factor

To predict BAF the genetic algorithm was applied on the dataset of 14 congeners and the initial 30 descriptors with the same approach as in the case of BCF. The best models were listed in the Table 3.

The last model (PLS-2) has the best statistics. However, it is associated with the highest number of compounds outside the AD according to the leverage approach. Thus, the third one (PLS-1), which is a PLS model constituted by five descriptors with three latent variables with only 38 congeners outside the AD, was selected to predict logBAF for all congeners. The descriptors used are the spectral mean absolute deviation from Burden matrix weighted by polarizability (SpMAD_B(p)) (Consonni and Todeschini, 2008), Geary autocorrelations on the molecular graph weighted by electro-negativity (GATS6e), Moran autocorrelation weighted by ionization potential (MATS5i), Geary autocorrelation weighted by ionization potential (GATS7i) and the frequency of Br–Br bonds at topological distance 4(F04[Br–Br]) (Todeschini and Consonni, 2009).

It was also confirmed to be the best model by two multivariate statistical process control (MSPC) tools that are often employed; the lack of model fit, or Q residual statistic, and the sample-to-model distance given by Hotelling's T^2 statistic (Jackson and Mudholkar, 1979). In Fig. 4A, T^2 values indicated that samples had low leverage on the model and did not exceeded the confidence limits of the model's hyperspace while residuals Q indicated that all samples were well reconstructed by the model.

To select the number of latent variables (LVs) to be included in the model we used the cross-validated root mean square error (RMSE) and Q^2 (Fig. 4B) that suggested the use of 3 latent variables. This was confirmed by the cumulated variance on the descriptor values that reached 83.6% and the response that reached 94.3% (Fig. 4C). The results were presented in Appendix A and congeners outside the AD are marked by asterisks.

3.3. Biomagnification factor

In the case of BMF, after applying the genetic algorithm on the dataset of 13 congeners and 30 variables, the choice of the best model was mostly based on the applicability domain because the one with the best statistics was associated with high number of congeners outside the AD according to the leverage approach. Thus, the model that presented the lower number of congeners that can be associated with low reliability of prediction was the third model MLR-10 (Table 3). The 4 descriptors were Moran autocorrelation weighted by Van der Waals volume (MATS6v), topological charge index of order 5 (GGI5), mean topological charge index of 2nd order (JGI2) and centered Moreau-Broto autocorrelation weighted by Van der Waals volume (ATSC1v) (Todeschini and Consonni, 2009). The high RMSE values is due to big standard deviation between training responses: 9.6 compared to 0.51 for logBCF.

The equation of the selected model is:

$$BMF = -12.8139 \times ATSC1v + 41.5349 \times MATS6v + 73.5575 \times GGI5 - 1.1780 \times 10^3 \times JGI2 + 109.7641 \quad (2)$$

Moreover, to better visualize the four variables of MLR-10 model into two dimensions Multi-Dimensional Scaling (MDS) was used. As it can be easily observed (Fig. 4D), most of the congeners are close to the training set while the 36 BDEs considered as outside of the model's AD were quite scattered and far from the training compounds and thus associated with lower reliability.

Calculated BMF by the selected model are in Appendix A. The biotransformation of the studied BDEs affected also the predictions resulting of low biomagnification rates. Thus, congeners with negative regression prediction values were considered not reliable (NR). Low reliability predictions, i.e. compound out the applicability domain are marked by asterisks.

4. Conclusions

This study was conducted in order to contribute to the risk assessment of PBDEs that are considered as persistent and harmful to the environment. For this purpose, we built different QSAR models to predict each of the three factors used to explain bioaccumulation of these POPs. The used datasets presented good coverage of the whole group of congeners.

Predicting bioaccumulation of PBDEs, like any other biological activity, is not an easy task due to lack of experimental data and information on biotransformation of these compounds. Metabolism of some congeners by debromination is a specie-specific and site specific process that can affect the reliability of model predictions (Table 1, Appendix A). This can explain the high predicted logBCF compared to experimental values for the two congeners that were removed from the training set (BDE-180 and BDE-183). Also the under estimated logBCF values in the case of the two congeners BDE-203 and BDE-207 can be due to the metabolisation of higher brominated congeners. However after applying the appropriate analysis tools and regression methods we were able to build models with good performance specifically fitted for BDEs that in the case of BCF was better than global models like CAESAR BCF and BCFBAF of EpiSuite. This simple linear model of two descriptors was based on a dataset of experiments on oligochaetes. These aquatic organisms were chosen for their important role in the environmental fate of the studied compounds as they play the role of a mediator between the sediments and the other aquatic species.

Besides the model prediction ability, the bioaccumulation models were selected in a way to reduce the used descriptors for more simplicity and the number of congeners outside the AD to get higher reliability of predictions. The models were then validated as required by the OECD principles. Most of the used variables are describing the electronic profile and relative bromine positions of the congeners.

Results from BCF model, together with the estimated logBAF and BMF, can be used as input data for environmental fate models and risk assessment of this group of compounds.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement n. 238701 of Marie Curie ITN Environmental Chemoinformatics (ECO) project.

Experiments were conducted in the framework the European project CADASTER under FP7.

Appendix A

See Table A1.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.chemosphere.2012.05.081.

References

Arnot, J.A., Gobas, F.A.P.C., 2006. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ. Rev. 14, 257–297.

ATSDR, 2004. Toxicological Profile of Polybrominated Biphenyls and Polybrominated Diphenyleters. <http://www.atsdr.cdc.gov/toxprofiles/tp68.pdf>.

Ballabio, D., Vasighi, M., Consonni, V., Kompany-Zareh, M., 2011. Genetic algorithms for architecture optimization of counter-propagation artificial neural networks. Chemometr. Intell. Lab. 105, 56–64.

Ballschmiter, K., Bacher, R., Mennel, A., Fischer, R., Riehle, U., Swerev, M., 1992. Determination of chlorinated biphenyls, chlorinated dibenzodioxins, and chlorinated dibenzofurans by GC-MS. J. High Res. Chromatog. 15, 260–270.

Braekevelt, E., Tittlemier, S.A., Tomy, G.T., 2003. Direct measurement of octanol–water partition coefficients of some environmentally relevant brominated diphenyl ether congeners. Chemosphere 51 (7), 563–567.

Burreau, S., Axelman, J., Broman, D., Jakobsson, E., 1997. Dietary uptake in pike (Esox lucius) of some polychlorinated biphenyls, polychlorinated naphthalenes and polybrominated diphenyl ethers administered in natural diet. E. Environ. Toxicol. Chem. 16, 2508.

Consonni, V., Todeschini, R., 2008. New spectral indices for molecule description. MATCH 60, 3–14.

Consonni, V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q^2 parameter for QSAR validation. J. Chem. Inf. Model 49, 1669–1678.

Consonni, V., Ballabio, D., Todeschini, R., 2010. Evaluation of model predictive ability by external validation techniques. J. Chemometr. 24, 194–201.

Darnerud, P.O., Atuma, S., Aune, M., Cnattingius, S., Wernroth, M.L., 1998. Polybrominated diphenyl ethers (PBDEs) in breast milk from primiparous women in Uppsala County, Sweden. Organohalogen Compd. 35, 411–414.

de Wit, C.A., 2002. An overview of brominated flame retardants in the environment. Chemosphere 46, 583–624.

Dimitrov, S.D., Mekenyan, O.G., Walker, J.D., 2002. Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals. SAR QSAR Environ. Res. 13, 177–184.

Environment Canada, 2004. Environmental Screening Assessment Report on Polybrominated Diphenyl Ethers (PBDEs).

Eriksson, P., Jakobsson, E., Fredriksson, A., 1998. Developmental neurotoxicity of brominated flame retardants, polybrominated diphenyl ethers and tetrabromo-bis-phenol A. Organohalogen Compd. 35, 375–377.

Eriksson, P., Jacobsson, E., Fredriksson, A., 2001. Developmental neurotoxicity of brominated flame retardants, polybrominated diphenyl ethers and tetrabromobis-phenol A. Organohalogen Compd. 35, 375–377.

Fang, L., Huang, J., Yu, G., Wang, L., 2008. Photochemical degradation of six polybrominated diphenyl ether congeners under ultraviolet irradiation in hexane. Chemosphere 71, 258–267.

Fernlof, G., Gadhasson, I., Podra, K., Darnerud, P.O., Thuvander, A., 1997. Lack of effects of some individual polybrominated diphenyl ether (PBDE) and polychlorinated biphenyl congeners on human lymphocyte functions in vitro. Toxicol. Lett. 90 (2–3), 189–197.

Fisk, A.T., Norstrom, R.J., Cymbalisty, C.D., Muir, D.C.G., 1998. Dietary accumulation and depuration of hydrophobic organochlorines: bioaccumulation parameters and their relationship with the octanol/water partition coefficient. Environ. Toxicol. Chem. 17, 951–961.

Gobas, F.A.P.C., Morrison, H.A., 2000. Bioconcentration and biomagnification in the aquatic environment. In: Boethling, R.S., Mackay, D. (Eds.), Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences. CRC Press, Boca Raton, Fla., USA, pp. 189–231.

Gobas, F.A.P.C., Zhang, X., 1992. Measuring bioconcentration factors and rate constants of chemicals in aquatic organisms under conditions of variable water concentrations and short exposure time. Chemosphere 25, 1961–1971.

Goodrich, M.S., Melancon, M.J., Davis, R.A., Lech, J.J., 1991. The toxicity, bioaccumulation, metabolism, and elimination of dioctyl sodium sulfosuccinate DSS in rainbow trout (Oncorhynchus mykiss). Water Res. 25, 119–124.

Ghose, A.K., Crippen, G.M., 1986. Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure–activity relationships. I. Partition coefficients as a measure of hydrophobicity. J. Comput. Chem. 7, 565–577.

Hawker, D.W., Connell, D.W., 1988. Octanol–water partition coefficients of polychlorinated biphenyl congeners. Environ. Sci. Technol. 22, 382–387.

He, J., Robrock, K.R., Alvarez-Cohen, L., 2006. Microbial reductive debromination of polybrominated diphenyl ethers (PBDEs). Environ. Sci. Technol. 40 (14), 4429–4434.

Hutzinger, O., Sundstrom, G., Safe, S., 1976. Environmental chemistry of flame retardants. Part I. Introduction and principles. Chemosphere 5, 3–10.

Isnard, P., Lambert, S., 1988. Estimating bioconcentration factors from octanolwater partition coefficient and aqueous solubility. Chemosphere 1, 21.

Jackson, J.E., Mudholkar, G.S., 1979. Control procedures for residuals associated with principal components analysis. Technometrics 21, 341–349.

Jakobsson, K., Athanasiadou, M., Christiansson, A., Athanassiadis, I., Bergman, A., Hagmar, L., 2005. Polybrominated diphenyl ethers (PBDEs) in serum from Swedish men 1988–2002. A longitudinal study. Organohalogen Compd. 67, 533–536.

Jiang-Ping, W., Xiao-Jun, L., Ying, Z., Yong, L., She-Jun, C., Bi-Xian, M., Zhong-Yi, Y., 2008. Bioaccumulation of polybrominated diphenyl ethers (PBDEs) and polychlorinated biphenyls (PCBs) in wild aquatic species from an electronic waste (e-waste) recycling site in South China. Environ. Internat. 34, 1109–1113.

Johnson, A., Seiders, K., Deligeannis, C., Kinney, K., Sandvik, P., Era-Miller, B., Alkire, D., 2006. PBDE flame retardants in Washington rivers and lakes: concentrations in fish and water, 2005–2006. Washington Ecology Section, Environmental Assessment Program, Washington State Department of Ecology, Olympia, Washington 98504-7710. Publication No. 06-03-027.

Kannan, K., Nakata, H., Stafford, R., Masson, G.R., Tanabe, S., Giesy, J.P., 1998. Bioaccumulation and toxic potential of extremely hydrophobic polychlorinated biphenyl congeners in biota collected at a superfund site contaminated with Aroclor 1268. Environ. Sci. Technol. 32, 1214–1221.

Mizukawa, K., Takada, H., Takeuchi, I., Ikemoto, T., Omori, K., Tsuchiya, K., 2009. Bioconcentration and biomagnification of polybrominated diphenyl ethers (PBDEs) through lower-trophic-level coastal marine food web. Mar. Pollut. Bull. 58, 1217–1224.

Kim, Y.M., Nam, I.H., Murugesan, K., Schmidt, S., Crowley, D.E., Chang, Y.S., 2007. Biodegradation of diphenyl ether and transformation of selected brominated congeners by Sphingomonas sp. PH-07. Appl. Microbiol. Biotechnol. 77, 187–194.

Konstantinov, A., Arsenault, G., Chittim, B., McAlees, A., McCrindle, R., Potter, D., Tashiro, C., Yeo, B., 2008. Identification of the minor components of Great Lakes DE-71™ technical mix by means of ^{1}H NMR and GC/MS. Chemosphere 73, S39–S43.

Kuramochi, H., Maeda, K., Kawamoto, K., 2007. Physicochemical properties of selected polybrominated diphenyl ethers and extension of the UNIFAC model to brominated aromatic compounds. Chemosphere 67, 1858–1865.

Leardi, R., Lupianez, A., 1998. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. Chemometr. Intell. Lab. 41, 195–207.

Leonards, P.E.G., Broekhuizen, S., de Voogt, P., Van Straalen, N.M., Brinkman, U.A.T., Cofino, W.P., van Hattum1, B., 1998. Studies of bioaccumulation and biotransformation of PCBs in mustelids based on concentration and congener patterns in predators and preys. Arch. Environ. Contam. Toxicol. 35, 654–665.

Lombardo, A., Roncaglioni, A., Boriani, E., Milan, C., Benfenati, E., 2010. Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. Chem. Cent. J. 4 (Suppl. 1), S1.

Mackay, D., 1982. Correlation of bioconcentration factors. Environ. Sci. Technol. 16, 274–278.

McDonald, T.A., 2002. A perspective of the potential health risks of PBDEs. Chemosphere 46, 745–755.

Meylan, W.M., Howard, P.H., Boethling, R.S., Aronson, D., Printup, H., Gouchie, S., 1999. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. Environ. Toxicol. Chem. 18, 664–672.

Mikula, P., Svobodova, Z., 2006. Brominated flame retardants in the environment: their sources and effects (a review). Acta Vet. Brno 75, 587–599.

Miles, J., Shelvin, M., 2001. Applying Regression and Correlation. Sage Publication, London.

Moermond, C.T.A., Zwolsman, J.J.G., Koelmans, A.A., 2005. Black carbon and ecological factors affect in site biota to sediment accumulation factors for hydrophobic organic compounds in flood plain lakes. Environ. Sci. Technol. 39, 3101–3109.

Moreau, G., Broto, P., 1980a. Autocorrelation of molecular structures. Application to SAR studies. Nouv. J. Chim. 4, 757–764.

Moreau, G., Broto, P., 1980b. The autocorrelation of a topological structure: a new molecular descriptor. Nouv. J. Chim. 4, 359–360.

Nichols, J.W., McKim, J.M., Andersen, M.E., Gargas, M.L., Clewell, H.J., Erickson, R.J., 1990. A physiology based toxicokinetic model for the uptake and disposition of waterborne organic chemicals in fish. Toxicol. Appl. Pharmacol. 106, 433–447.

Noren, K., Meironyte, D., 2000. Certain organochlorine and organobromine contaminants in Swedish human milk in perspective of past 20–30 years. Chemosphere 40, 1111–1123.

OECD, 1984. Earthworm, Acute Toxicity Tests. OECD 207. OECD Guidelines for testing of chemicals.

OECD, 1996. Bioconcentration: Flow-Through Fish Test. OECD guidelines for the testing chemicals No. 305E. Organization for Economic Co-operation and Development, Paris, France, p. 23.

OECD, 2008. Bioaccumulation in sediment-dwelling benthic oligochaetes. OECD 315. OECD Guidelines for the Testing of Chemicals.

Palm, A., Cousins, I.T., Mackay, D., Tysklind, M., Metcalfe, C., Alaee, M., 2002. Assessing the environmental fate of chemicals of emerging concern: a case study of the polybrominated diphenyl ethers. Environ. Pollut. 117, 195–213.

Patterson, D.G., Sjodin, A., Bergman, A., 2000. Brominated flame retardants in serum from US blood donors. Organohalogen Compd. 47, 45–48.

Pijnenburg, A.M.C.M., Everts, J.W., de Boer, J., Boon, J.P., 1995. Polybrominated biphenyl and diphenyl ether flame retardants: analysis, toxicity and environmental occurrence. Rev. Environ. Contam. Toxicol. 141, 1–26.

Safe, S., 1990. Polychlorinated biphenyls (PCBs), dibenzo-p-dioxins (PCDDs), dibenzofurans (PCDFs), and related compounds: environmental and mechanistic considerations which support the development of toxic equivalency factors (TEFs). Crit. Rev. Toxicol. 21 (1), 51–88.

Sharp, R., Lunder, S., 2004. In the dust. Toxic fire retardants in American homes. Environmental Working Group. <http://www.ewg.org/reports/inthedust>.

Shaw, G.R., Connel, D.W., 1982. Factors influencing concentrations of polychlorinated biphenyls in organisms from an estuarine ecosystem. Aust. J. Mar. Freshw. Res. 33, 1057–1070.

Slinker, B.Y., Glantz, S.A., 1985. Multiple regression for physiological data analysis: the problem of multicollinearity. Am. J. Phys. 249, R1–R12.

Stapleton, H.M., Alaee, M., Letcher, R.J., Li, J., Baker, J.E., 2004a. Debromination of the flame retardant decabromodiphenylether by juvenile carp (Cyprinus carpio) following dietary exposure. Environ. Sci. Technol. 38, 112–119.

Stapleton, H.M., Letcher, R.J., Li, J., Baker, J.E., 2004b. Dietary accumulation and metabolism of polybrominated diphenylethers by juvenile carp (Cyprinus carpio). Environ. Toxicol. Chem. 23, 1939–1946.

Stapleton, H.M., Letcher, R.J., Baker, J.E., 2004c. Debromination of polybrominated diphenyl ether congeners BDE 99 and BDE 183 in the intestinal tract of the common carp (Cyprinus carpio). Environ. Sci. Technol. 38, 1054–1061.

Stapleton, H.M., Letcher, R.J., Baker, J.E., 2002. Uptake, metabolism and depuration of polybrominated diphenyl ethers (PBDEs) by the common carp, (Cyprinus carpio). Organohalogen Compd. 58, 201–204.

Stapleton, H.M., Dodder, N.G., 2008. Photodegradation of decabromodiphenyl ether in house dust by natural sunlight. Environ. Toxicol. Chem. 27 (2), 306–312.

Streets, S.S., Henderson, S.A., Stoner, A.D., Carlson, D.L., Simcik, M.F., Swackhamer, D.L., 2006. Partitioning and bioaccumulation of PBDEs and PCBs in Lake Michigan. Environ. Sci. Technol. 40, 7263–7269.

Todeschini, R., Consonni, V., 2009. Molecular Descriptors for Chemoinformatics. Wiley, Weinheim.

Tomy, G.T., Palace, V.P., Halldorson, T., Braekevelt, E., Danell, R., Wautier, K., Evans, B., Lyndon, B., Fisk, A.T., 2004. Bioaccumulation, biotransformation and biochemical effects of brominated diphenyl ethers in juvenile lake trout (Salvelinus namaycush). Environ. Sci. Technol. 38, 1496–1504.

Toshima, S., Moriya, T., Yoshimura, K., 1992. Effects of polyoxyethylene (20) sorbitan monooleate on the acute toxicity of linear alkylbenzenesulfonate (C12-LAS) to fish. Ecotoxicol. Environ. Safe. 24, 26–36.

UNEP, 2009. Conference of the Parties of the Stockholm Convention on Persistent Organic Pollutants. Fourth Meeting. Recommendations of the Persistent Organic Pollutants Review Committee of the Stockholm Convention to amend Annexes A, B or C of the Convention Geneva, 4–8 May, 2009.

USEPA, 2000. Methodology for Deriving Ambient Water Quality Criteria for the Protection of Human Health. US Environmental Protection Agency, Washington, DC, USA. p. 180.

USEPA, 2010. An Exposure Assessment of Polybrominated Diphenyl Ethers. U.S. Environmental Protection Agency. National Center for Environmental Assessment Office of Research and Development. USEPA/600/R-08/086F.

Walker, S.L., Gobas, F.A.P.C., 1999. An investigation of the application of Canadian water quality guidelines. Environ. Toxicol. Chem. 18, 1323–1328.

Wang, Y.W., Li, X.M., Li, A., Wang, T., Zhang, Q.H., Wang, P., Fu, J., Jiang, G., 2007. Effect of municipal sewage treatment plant effluent on bioaccumulation of polychlorinated biphenyls and polybrominated diphenyl ethers in the recipient water. Environ. Sci. Technol. 41, 6026–6032.

Wania, F., Dugani, C., 2002. Assessing the long range transport potential of polybrominated diphenyl ethers: a comparison of four multimedia models, Final Report. University of Toronto at Scarborough, Scarborough, Ontario.

Watanabe, I., Tatsukawa, R., 1987. Formation of brominated dibenzofurans from the photolysis of flame retardant decabromobiphenyl ether in hexane by UV and sun light. B. Environ. Contam. Toxicol. 39, 953–959.

Weber, R., Kuch, B., 2003. Relevance of BFRs and thermal conditions on the formation pathways of brominated and brominated-chlorinated dibenzodioxins and dibenzofurans. Environ. Int. 29, 699–710.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. 58, 109–130.

Wolkers, H., Bavel, B.V., Derocher, A.E., Wiig, Ø., Kovacs, K.M., Lydersen, C., Lindström, G., 2004. Congener-specific accumulation and food chain transfer of polybrominated diphenyl ethers in two Arctic food chains. Environ. Sci. Technol. 38, 1667–1674.

Annex II: Ready biodegradability of chemicals

Mansouri, K.; Ringsted, T.; Ballabio, D.; Todeschini, R.; Consonni, V. Quantitative Structure-Activity Relationship Models for Ready Biodegradability of Chemicals. J. Chem. Inf. Model. 2013, 53, 867–878.

JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

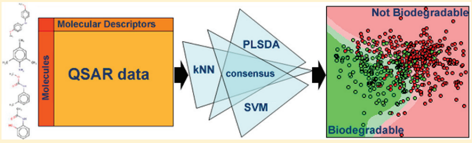
Quantitative Structure–Activity Relationship Models for Ready Biodegradability of Chemicals

Kamel Mansouri, Tine Ringsted, Davide Ballabio,* Roberto Todeschini, and Viviana Consonni

Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano Bicocca, Milano, Italy

Supporting Information

ABSTRACT: The European REACH regulation requires information on ready biodegradation, which is a screening test to assess the biodegradability of chemicals. At the same time REACH encourages the use of alternatives to animal testing which includes predictions from quantitative structure–activity relationship (QSAR) models. The aim of this study was to build QSAR models to predict ready biodegradation of chemicals by using different modeling methods and types of molecular descriptors. Particular attention was given to data screening and validation procedures in order to build predictive models. Experimental values of 1055 chemicals were collected from the webpage of the National Institute of Technology and Evaluation of Japan (NITE): 837 and 218 molecules were used for calibration and testing purposes, respectively. In addition, models were further evaluated using an external validation set consisting of 670 molecules. Classification models were produced in order to discriminate biodegradable and nonbiodegradable chemicals by means of different mathematical methods: *k* nearest neighbors, partial least squares discriminant analysis, and support vector machines, as well as their consensus models. The proposed models and the derived consensus analysis demonstrated good classification performances with respect to already published QSAR models on biodegradation. Relationships between the molecular descriptors selected in each QSAR model and biodegradability were evaluated.



1. INTRODUCTION

Substances which do not decay over a period of time pose a potential threat of accumulation and spread in the environment and organisms. Accumulation of persistent chemicals can, in the long run, show to be harmful because of the continuous exposure and the increasing chemical concentration in the surroundings.[1] The danger is, therefore, that the damages do not have to be immediate but can immerse after a longer period of time.

In Europe, legislators have consequently included persistency in the evaluation of chemicals in the framework of the European REACH regulation. REACH requires that chemicals produced or imported in quantities of more than 1 ton per year need information on ready biodegradation, which is a screening test for the assessment of biodegradability.[2] Thousands of chemicals exist in consumer products, and these can eventually end up in the environment. As an example, the EINECS list comprises more than 100 000 chemicals registered as being on the European Community market between 1971 and 1981.[3] Only a limited number of chemicals from the EINECS list have been tested for their biodegradability. Even for chemicals produced or imported at more than 1000 tons per year, the percentage of those with biodegradation data is merely 61%.[4] To increase the amount of data, REACH encourages the use of a weight-of-evidence approach, which means that all available information should be considered, including predictions from QSARs (quantitative structure–activity relationships) and read-across.

Several QSAR models have in the past been built to predict biodegradation with the use of different types of data, such as chemical half-life, expert judgment, and biodegradation screening tests. Table 1 collects some of the already published QSAR models which were built to classify molecules as ready or not ready biodegradable.[5−21] Both fingerprints (binary matrix stating the presence and absence of fragments/properties) and molecular descriptors have been used for modeling biodegradation.

In the study of Cheng, a consensus model was built where the average of several models were used to classify molecules.[22] The consensus model could correctly predict all ready biodegradable (RB) and not ready biodegradable (NRB) molecules in an external validation set of 27 molecules but the test set contained only four RB molecules. It was not evaluated if the external validation set was a good representation of the chemical space of the model, but four molecules may not be enough to cover the chemical domain.

Several structural features have been found to increase the time for biodegradation (for example halogens, chain branching, nitro groups, polycyclic residues, heterocyclic residues, and aliphatic ether bonds).[1] On the other hand,

Received: January 11, 2013
Published: March 7, 2013

Table 1. Classification Models on Ready Biodegradation Published in the Literature[a]

ref	method	descriptors	training set (RB/NRB)	test set (RB/NRB)	test set Sn	test set Sp
Loonen et al. 1999[16]	PLSDA	F (127 struct frag)	670 (Na/Na)	224 (Na/Na)	0.80	0.85
Tunkel et al. 2000[17]	MLR	D (43 struct frag + mw)	589 (254/335)	295 (131/164)	0.80	0.82
Tunkel et al. 2000[17]	LR	D (43 struct frag + mw)	589 (254/335)	295 (131/164)	0.79	0.83
Cheng et al. 2012[22]	NB	D (10 physicochemical)	1440 (529/911)	164 (62/102)	0.71	0.91
Cheng et al. 2012[22]	*k*NN	D (12 physicochemical)	1440 (529/911)	164 (62/102)	0.73	0.91
Cheng et al. 2012[22]	SVM	F (79 E-state keys)	1440 (529/911)	164 (62/102)	0.61	0.93

[a]The original reference, modelling method, type and number of descriptors, number of molecules included in training and test sets (as well as number of molecules for ready biodegradable and not ready biodegradable classes), sensitivity, and specificity obtained in the test set are reported for each model. Partial least squares discriminant analysis (PLSDA), multiple linear regression (MLR), logistic regression (LR), naive Bayes (NB), *k* nearest neighbours (*k*NN), support vector machines (SVM), fingerprints (F), molecular descriptors (D), structural fragments (struct frag), molecular weight (mw), ready biodegradable (RB), not ready biodegradable (NRB), not available (Na), sensitivity (Sn, correctly predicted ready biodegradable), specificity (Sp, correctly predicted not ready biodegradable). Consider that Sn and Sp are expressed as ratios, while some of the original papers report them as percentages.

some structural features have been found to enhance biodegradability. These features include esters, amides, hydroxyl groups, aldehyde groups, carboxylic acid groups, unbranched linear alkyne chains and phenyl rings.[1] However, the presence of one of these structural features does not indicate an RB or NRB molecule but should only be taken as generalizations.[1] A physicochemical property which have been found to correlate with the rate of biodegradation is water solubility where soluble molecules tend to be more easily biodegradable compared to insoluble molecules.[23] Molecular weight has also been indicated as an important factor in relation to biodegradation because molecules with a molecular weight higher than 500 cannot be transported into bacterial cells.[24] For some large molecules like proteins and polysaccharides, extracellular enzymes can degrade the molecules into smaller entities which can pass through the cell membrane.[1]

The aim of this study was to build QSAR models to predict ready biodegradation of chemicals by using different modeling methods and types of molecular descriptors. Particular attention was given to data screening and validation procedures in order to carry out predictive models.[25] A set of 837 molecules was used for calibration purposes, while 218 molecules were used to test the calibrated QSAR models. In addition, models were further evaluated using an external validation set consisting of 670 molecules. Data were carefully screened to ensure accurate models based on correct experimental values and molecular structures. The considered classification modeling methods included linear, nonlinear, and local models, as well as consensus models. These methods were coupled with genetic algorithms in order to select the optimal subsets of molecular descriptors. The proposed QSAR models were interpreted in connection to the current knowledge on biodegradation. Finally, since new molecular descriptors were introduced in the proposed models, their link to biodegradability was discussed.

2. MATERIALS AND METHODS

2.1. Biodegradability Experimental Data. Experimental data of the Japanese Ministry of International Trade and Industry (MITI) test (I) were collected from the webpage of the National Institute of Technology and Evaluation (NITE) of Japan.[26] The test is one of the six approved screening tests for ready biodegradation from the Organization for Economic Cooperation and Development (OECD).[27] The MITI test measures the biochemical oxygen demand (BOD) in aerobic aqueous medium for 28 days (the original OECD protocol used a 14 day test period).[17] Chemicals with a BOD value higher than 60% are considered as RB whereas those with a BOD lower than 60% are regarded as NRB.[22,27,28]

The initial data set contained 1309 molecules. BOD values and classification judgments from NITE were given for all the collected molecules. The data set was screened to ensure that it was in accordance with the OECD test protocol (301 C) and that the correct chemical structures were used.

2.2. Data Screening. The screening procedure was carried out on the basis of the steps described in the following paragraphs and summarized in Table 2.

Table 2. Results from the Screening Procedure of the MITI Data

reason for removal of molecules from the data set	number of removed molecules
chemical name and CAS number not in accordance	81
BOD replicates had more than 20% difference and classified differently	24
classification would change if nitrification was taken into account	4
experimental and NITE classification did not agree	54
disconnected structures	91
total number of removed molecules	254

2.2.1. Analysis of the Molecular Structures. The simplified molecular-input line-entry system (SMILES) format was used as the molecular structure representation. SMILES strings were collected from ChemSpider,[29] using a KNIME workflow.[30] When two CAS numbers were assigned to a chemical in the MITI database, then only "Biodegradation: CAS Registry No." was taken into consideration. When several names were specified in the MITI database, "Chemical Name in the Official Bulletin" was considered unless "Biodegradation: Name of chemical tested" was present. Here, 81 chemicals were removed because their CAS Registry Number and chemical names were not consistent in ChemSpider and the MITI database.

2.2.2. Handling of BOD Replicates. Replicate BOD values were given for 223 compounds in the MITI database. Most molecules with BOD replicates had three values. If one of the three values was significantly deviating from the two other BOD values according to Dixon's Q test with a 90% confidence limit,[31] then the deviating value was removed. If a molecule had a difference between BOD replicates higher than 20% and the replicate values classified the molecule into different categories (RB and NRB), then the molecule was removed. This was the case of 24 molecules. For the remaining molecules with replicate values, the average BOD was used.

2.2.3. Unifying the Test Duration. In the MITI data set, 427 molecules had a test period shorter than 28 days, while the rest had a test period of 28 days. The BOD values based on test periods shorter than 28 days were extrapolated to 28 days as proposed in the literature.[20] This extrapolation could over- or underestimate the BOD values. However, experimental data was only used if the BOD value and the judgment by NITE classified the molecule in the same class. It was therefore assumed that classification errors due to the extrapolation could be neglected.

2.2.4. Handling Molecules with Nitrification. If a molecule contains nitrogen, then there is a possibility for nitrification in the ready biodegradation test.[27] Nitrification involves the consumption of oxygen, and it is therefore necessary to exclude this consumption from the BOD value since the BOD should only measure the oxygen used by microorganisms. From the collected data, it was not possible to know the extent of nitrification. Four molecules which differed in their classification depending on the assumption of complete or no nitrification were removed.

2.2.5. Handling Divergences between Experimental and NITE Classification. As previously described, chemicals with BOD values higher than 60% are classified as biodegradable. When the experimental classification and the NITE judgment did not agree, molecules were removed. This was the case for 54 molecules.

2.2.6. Handling Disconnected Structures. The MITI data set included disconnected structures, such as salts, mixtures, isomer mixtures, and polymers. All 91 disconnected structures were removed because they could affect the subsequent calculation of molecular descriptors.

Table 2 summarizes the screening steps and the corresponding number of removed molecules. In the screening procedure, 254 molecules were removed. The remaining 1055 chemicals, with 356 RB and 699 NRB molecules, were used for modeling. The data set is provided in the Supporting Information file (SI) of this article.

2.3. Molecular Descriptors. The previously collected SMILES codes were used to calculate the molecular descriptors in DRAGON software version 6.[32] A two-dimensional structural representation was selected instead of 3D structures to avoid complex and irreproducible geometrical optimizations. The use of 3D descriptors could add valuable chemical information about the molecules. However, this type of descriptor requires a geometrical optimization and this can be an issue when applying the calibrated models to new molecules since the difference between the 3D conformers can affect the 3D descriptors values.

The calculated molecular descriptors were included in the following blocks of descriptors from DRAGON: constitutional indices, ring descriptors, topological indices, 2D matrix–based descriptors, functional group counts, atom-centered fragments, atom-type E-state indices, and 2D atom pairs (Table 3). A filtering of the descriptors was performed in DRAGON before exporting the descriptor values. Constant, near constant, and correlated descriptors were removed. In the latter case, for each pair of descriptors with a correlation coefficient higher than 98%, the one showing the largest pair correlation with all the other descriptors was excluded. A total number of 781 descriptors were exported from DRAGON for the subsequent modeling analysis.

Table 3. Number of Molecular Descriptors Initially Calculated by Using DRAGON

DRAGON block	number of descriptors
constitutional indices	32
ring descriptors	25
topological indices	34
2D matrix-based descriptors	84
functional group counts	88
atom centered fragments	69
atom-type E-state indices	37
2D atom pairs	412
total number of molecular descriptors	781

2.4. Modeling Methods. Three classification modeling methods were applied in order to find the appropriate relationship between molecular structures, encoded in molecular descriptors, and the biodegradability of chemicals: *k* nearest neighbors (*k*NN), partial least squares discriminant analysis (PLSDA), and support vector machines (SVM). The application of methods based on different mathematical strategies aimed to better explore the chemical space and balance potential biases related to each single modeling algorithm.

The *k*NN classification rule is conceptually quite simple:[33] a molecule is classified according to the classes of the *k* closest molecules, which means, it is classified according to the majority of its *k* nearest neighbors in the descriptors space. In this work, the Euclidean metric was used to measure distances between molecules. The *k* value giving the lowest classification error in cross-validation was selected as the optimal one.

PLSDA is a classification technique that profits the properties of partial least squares regression (PLS2-based method) with the discrimination power of a classification technique.[34,35] It finds fundamental relations between the matrix of descriptors and the class vector by calculating latent variables (LVs), which are orthogonal linear combinations of the original variables. PLSDA models were optimized in cross-validation to find a compromise between the classification performance and the number of selected LVs.

SVM define a decision boundary that optimally separates two classes by maximizing the distance between them.[36,37] The decision boundary can be described as an hyperplane that is expressed in terms of a linear combination of functions parametrized by support vectors, which consist in a subset of training molecules. SVM algorithms search for the support vectors that give the best separating hyperplane using a kernel function. During optimization, SVM search the decision boundary with maximal margin among all possible hyperplanes, where the margin can be intended as the distance between the hyperplane and the closest point for both classes. This procedure was carried out by means of a kernel based on a radial basis function.

Genetic algorithms (GAs) were applied to find the optimal subset of molecular descriptors.[38] GAs start from an initial random population of chromosomes, which are binary vectors representing the presence or absence of molecular descriptors. An evolutionary process is simulated to optimize a defined fitness function and new chromosomes are obtained by coupling the chromosomes of the initial population with genetic operations (crossover and mutation). The used fitness function was the classification error calculated in cross-validation.

Consensus analysis was also applied in order to combine information and predictions obtained by the three different modeling techniques. In fact, the consensus approach can improve the quality of models by increasing their prediction reliability.[39] Consensus modeling has also been shown to diminish the effects of noisy data. Individual models contain varying amounts of noise, which can be reduced by averaging the predictions of several models.[40] The generation of a consensus analysis can be based on different strategies such as averaging, scoring, and probabilities.[39−42] In this work, two different consensus algorithms were adopted: (a) Each molecule was assigned to the most frequent class out of the three predictions obtained with the considered classification methods (*k*NN, PLSDA, and SVM). (b) A molecule was assigned only if the three models classified it in the same class; otherwise, it was not assigned.

2.5. Model Validation. Molecules were randomly divided into training and test sets, containing 80% and 20% of the total number of considered molecules, respectively. The selection was performed maintaining the class proportions, that is, the number of test molecules of each class was proportional to the number of training molecules of that class. The training set was used to select molecular descriptors and to build the classification models. Molecules of the test set were used just to evaluate the predictive ability of the trained models.

During model optimization and descriptor selection, a cross-validation procedure with five cancellation groups was used. Classification models were evaluated on the basis of specificity and sensitivity, which are the ability to correctly predict RB and NRB molecules, respectively. In particular, specificity (Sp) and sensitivity (Sn) were calculated with the following equations:

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \qquad \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

where, TN and TP are the number of true negatives and true positives, and FN and FP are the number of false negatives and false positives, respectively. Being a two-class model, consider that the sensitivity of one class corresponds to the specificity of the other class. In addition, the nonerror rate (NER) was calculated as the average of specificity and sensitivity, while the classification error rate (ER) was calculated as the complement of NER. These indices were used in order to better estimate classification performances in presence of a data set with unequal number of molecules in each class. In this study error rate, specificity, and sensitivity are expressed as ratios and not as percentages.

The classification models were further evaluated using an external validation set. This set was built merging two sources: 464 molecules of the data set modeled by Cheng et al.[22] and 206 molecules of the Canadian DSL database (Domestic Substances List).[43] Initially, 1604 compounds were collected from the set modeled by Cheng. These molecules were screened in order to remove compounds already present in our training or test sets. Moreover, the screening procedure used for the molecules of the MITI data set (described in section 2.2) was applied on this set of molecules. Some CAS numbers were missing, and thus, they were retrieved from ChemSpider matching molecular SMILES and chemical names. After the screening, 464 compounds were selected and included in the external validation set.

The considered DSL list consisted of more than 3500 compounds which meet the categorization of the Canadian Environmental Protection Act.[44] According to this categorization, compounds which are classified as persistent and bioaccumulative can be considered as not biodegradable. Therefore, the 420 persistent and bioaccumulative compounds of the DSL list were considered for inclusion in the external validation set. The SMILES structures were collected using the KNIME workflow described in section 2.2. After removing inorganic compounds, polymers, salts, and disconnected structures, as well as 21 compounds overlapping with the Cheng data set, 206 NRB molecules were added to the external validation set. Summarizing, the external validation set included a total of 670 compounds. The number of RB and NRB molecules of training, test, and external validation sets are summarized in Table 4.

Table 4. Number of Molecules Included in Training, Test, and External Validation Set

data	ready biodegradable	not ready biodegradable	total
training set	284	553	837
test set	72	146	218
external validation set	191	479	670

2.6. Software. A KNIME workflow[30] was used to collect and check SMILES notations from ChemSpider database.[29] Molecular descriptors were calculated by means of DRAGON.[32] SVM were calibrated using the library LIBSVM 3.1 implemented in C[45] and compiled in MATLAB 7.13.[46] GAs, models fitting, and predictions were performed in MATLAB 7.13[46] by means of routines built by the authors.

3. RESULTS AND DISCUSSION

3.1. Descriptor Selection and Model Calibration. The selection of molecular descriptors was organized into two subsequent steps in order to handle the large number of calculated descriptors (781) and to avoid potential overfitting of the QSAR models. GAs were separately calculated on each block of molecular descriptors (Table 3). Descriptors selected from each block were then merged and again GAs were used to find the most appropriate subset of molecular descriptors to calibrate the final QSAR models. The final models were selected taking into consideration the ER in cross-validation and a balanced ratio between the specificity and the sensitivity. It was also important that the final models used a reduced number of selected descriptors and for the PLSDA model also a low number of latent variables. This was done in order to make the models easy to interpret and at the same time to decrease the risk of overfitting. The same procedure was used by coupling GAs with the three considered modeling methods. The classification model based on *k*NN, PLSDA, and SVM included 12, 23, and 14 molecular descriptors, respectively. The obtained QSAR models were validated using the molecules included in both test and external validation sets, which did not participate in the descriptor selection and model calibration. Values of the selected molecular descriptors are provided in the Supporting Information (SI) of this article.

3.2. Classification Performances of the QSAR Models. The classification performances of the three QSAR models are collected in Table 5. The three classification models showed comparable performances. The ER in fitting and cross-validation was equal to 0.14 for all the computed models, while the error on the test set was equal to 0.14 with SVM and slightly higher (0.15) with both PLSDA and *k*NN. This balance

Table 5. Classification Results in Fitting, Cross-Validation, and on the Test Set of the Proposed QSAR Models and Their Consensus Analysis[a]

			fitting			5-fold cross-validation			test set		
model	desc	k/LVs/c	ER	Sn	Sp	ER	Sn	Sp	ER	Sn	Sp
*k*NN	12	6	0.14	0.84	0.89	0.14	0.84	0.88	0.15	0.81	0.90
PLSDA	23	5	0.14	0.88	0.83	0.14	0.88	0.83	0.15	0.83	0.87
SVM	14	5	0.14	0.81	0.92	0.14	0.80	0.91	0.14	0.82	0.91
consensus 1	41		0.11	0.86	0.91	0.11	0.87	0.91	0.13	0.82	0.92
consensus 2	41		0.07	0.91	0.95	0.07	0.91	0.95	0.09	0.88	0.94
				19% not assigned			20% not assigned			15% not assigned	

[a]For each model, the number of included descriptors, error rate (ER), specificity (Sp, correctly predicted not ready biodegradable), and sensitivity (Sn, correctly predicted ready biodegradable) are provided. The optimal parameters, *k* for *k*NN, the number of latent variables (LVs) for PLSDA, and the cost (c) for SVM, are reported in the table.

between model performances on the training and test sets can indicate the absence of overfitting, which is a possibility when dealing with variable selection on high dimensional data. Moreover, the quality of a QSAR classification model should also be evaluated on the basis of its ability to correctly predict each modeled class. Compared to other QSAR models on ready biodegradation,[16,17,22] the proposed models showed a good balance between specificity and sensitivity, which never had a difference higher than 0.11. In addition, specificity and sensitivity values calculated on the training and test sets were comparable, indicating robustness and reliability of the proposed models. The SVM and *k*NN models showed higher specificity than sensitivity for the RB class, that is, errors associated to NRB molecules predicted as biodegradable were lower. PLSDA, on the other hand, had the opposite behavior in the fitting and cross-validation results but the test set result showed the same tendency as the *k*NN and SVM models. The reason for the PLSDA model to have a higher sensitivity than specificity in the fitting and cross-validation results and the opposite scenario for the test set was not known but since the test set was chosen randomly, it is possible that this result is due to random variation.

Afterward, predictions obtained by the three classification QSAR models were merged and models based on the two different consensus approaches previously described were calculated. Classification performances of the consensus models are shown in Table 5. When assigning molecules to the most frequent class (consensus 1), classification results were improved, maintaining a reasonable balance between specificity and sensitivity on both the training and test sets. On the other hand, when not assigning molecules in the presence of divergence between the three classification models (consensus 2), ERs were further decreased. The classification ER on the test molecules was equal to 0.09, with a percentage of not assigned test molecules equal to 15%. This could mean that not assigned molecules were associated with lower reliability of prediction. In any case, the presence of 15% not classified molecules was balanced by the good classification performances of the consensus 2 model. As an example of the improved performance it can be mentioned that consensus 2 gave a specificity on the test set (ratio of correctly predicted NRB test molecules) equal to 0.94. The presence of not classified molecules is a matter of choice between high reliability with less molecules predicted or predictions for all the molecules with a lower reliability.

The external validation set supported the results for the three QSAR and consensus models by giving results relatively close to the cross-validation and test set validation. Results are collected in Table 6. The ERs of *k*NN, SVM, PLSDA, and consensus 1 were in the range between 0.17 and 0.18 and thus comparable with those obtained on the training and test sets (Table 5). The slightly higher ERs were expected due to the different sources of the external validation molecules, which could have slightly different classification thresholds with respect to the MITI data set used to train the models. Consensus 2 gave again the lowest ER on the external validation set (0.13) and 13% of not assigned molecules. Finally, all considered models showed the same conservative behavior on the external validation set, that is, specificity was always higher than sensitivity. Thus, NRB molecules were more accurately predicted and models did not tend to classify them as biodegradable.

Table 6. Classification Results on the External Validation Set of the Proposed QSAR Models and Their Consensus Analysis[a]

	ER	Sn	Sp
*k*NN	0.17	0.75	0.91
PLSDA	0.17	0.80	0.86
SVM	0.18	0.74	0.91
consensus 1	0.17	0.76	0.91
consensus 2	0.13	0.81	0.94
		13% not assigned	

[a]For each model, error rate (ER), specificity (Sp, correctly predicted not ready biodegradable), and sensitivity (Sn, correctly predicted ready biodegradable) are provided.

Considering all the obtained results in classification (summarized in Tables 5 and 6), as well as the model complexity (represented by the number of selected molecular descriptors), the proposed QSAR models and their consensus models had equal or better classification performances with respect to models already published in the literature (Table 1). In particular, models selected in this study showed balanced results on training, test, and external validation sets, suggesting reliable predictions and the absence of potential overfitting. This is in contrast to some of the models published in the literature which showed greater difference between the calibration and validation results, that is, ER equal to 0.00 and 0.18 on the training and test molecules, respectively.[22]

3.3. Descriptor Interpretation.

One of the fundamentals of QSAR is that models should be reduced to a set of descriptors which is information rich but as small as possible, in order to ensure stability of the model and reliability of its predictions.[47,48] The symbols of the descriptors, the descriptor

Table 7. List of Molecular Descriptors Selected in the QSAR Models

symbol	description	DRAGON block	model
B01[C-Br]	presence/absence of C−Br at topological distance 1	2D atom pairs	PLSDA
B03[C-Cl]	presence/absence of C−Cl at topological distance 3	2D atom pairs	PLSDA
B04[C-Br]	presence/absence of C−Br at topological distance 4	2D atom pairs	PLSDA
C%	percentage of C atoms	constitutional indices	*k*NN−PLSDA
C-026	R−CX−R	atom centered fragments	SVM
F01[N-N]	frequency of N−N at topological distance 1	2D atom pairs	*k*NN
F02[C-N]	frequency of C−N at topological distance 2	2D atom pairs	SVM
F03[C-N]	frequency of C−N at topological distance 3	2D atom pairs	*k*NN
F03[C-O]	frequency of C−O at topological distance 3	2D atom pairs	PLSDA
F04[C-N]	frequency of C−N at topological distance 4	2D atom pairs	*k*NN−PLSDA
HyWi_B(m)	hyper-Wiener-like index (log function) from Burden matrix weighted by mass	2D matrix-based	PLSDA
J_Dz(e)	Balaban-like index from Barysz matrix weighted by Sanderson electronegativity	2D matrix-based	*k*NN
LOC	lopping centric index	topological indices	PLSDA
Me	mean atomic Sanderson electronegativity (scaled on Carbon atom)	constitutional indices	PLSDA
Mi	mean first ionization potential (scaled on carbon atom)	constitutional indices	PLSDA
N-073	Ar2NH/Ar3N/Ar2N−Al/R···N···R	atom centered fragments	PLSDA
nArCOOR	number of esters (aromatic)	functional group counts	SVM
nArNO2	number of nitro groups (aromatic)	functional group counts	PLSDA
nCb-	number of substituted benzene C(sp2)	functional group counts	*k*NN−SVM
nCIR	number of circuits	ring descriptors	PLSDA
nCp	number of terminal primary C(sp3)	functional group counts	*k*NN
nCrt	number of ring tertiary C(sp3)	functional group counts	SVM
nCRX3	number of CRX3	functional group counts	PLSDA
nHDon	number of donor atoms for H-bonds (N and O)	functional group counts	SVM
nHM	number of heavy atoms	constitutional indices	*k*NN
nN	number of nitrogen atoms	constitutional indices	SVM
nN-N	number of N hydrazines	functional group counts	PLSDA−SVM
nO	number of oxygen atoms	constitutional indices	*k*NN−PLSDA
NssssC	number of atoms of type ssssC	atom-type E-state indices	*k*NN−SVM
nX	number of halogen atoms	constitutional indices	SVM
Psi_i_1d	intrinsic state pseudoconnectivity index−type 1d	topological indices	PLSDA
Psi_i_A	intrinsic state pseudoconnectivity index—type S average	topological indices	SVM
SdO	sum of dO E-states	atom-type E-state indices	PLSDA
SdssC	sum of dssC E-states	atom-type E-state indices	*k*NN
SM6_B(m)	spectral moment of order 6 from Burden matrix weighted by mass	2D matrix-based	SVM
SM6_L	spectral moment of order 6 from Laplace matrix	2D matrix-based	PLSDA
SpMax_A	leading eigenvalue from adjacency matrix (Lovasz−Pelikan index)	2D matrix-based	PLSDA
SpMax_B(m)	leading eigenvalue from Burden matrix weighted by mass	2D matrix-based	SVM
SpMax_L	leading eigenvalue from Laplace matrix	2D matrix-based	*k*NN−PLSDA−SVM
SpPosA_B(p)	normalized spectral positive sum from Burden matrix weighted by polarizability	2D matrix-based	PLSDA
TI2_L	second Mohar index from Laplace matrix	2D matrix-based	PLSDA

blocks, and a brief description of the molecular descriptors from DRAGON which were selected in this study are collected in Table 7. The numbers of molecular descriptors included in each of the proposed models are comparable to those published in the literature (Table 1). In order to evaluate how the selected descriptors related to ready biodegradability, principal component analysis (PCA) was performed separately on the descriptors selected in the *k*NN and SVM models. PCA models were calculated on the training set, while test set molecules were projected onto the PCA model. Scores and loadings plots were used to discuss the behavior of the selected descriptors in relation to the knowledge on biodegradability found in the literature. The descriptors included in the PLSDA model were directly analyzed by means of the latent variables calculated by PLSDA.

3.3.1. Molecular Descriptors of the kNN Model. The *k*NN model included 12 descriptors (Table 7). The results of the PCA analysis on the set of 12 descriptors are shown in Figure 1. The score plot of the first and fourth principal components (PC1 and PC4) explained together 34% of the variance and is shown in Figure 1A, while the corresponding loading plot is presented in Figure 1B.

The combination of PC1 and PC4 gave a reasonable separation between RB and NRB molecules. The majority of NRB molecules had positive scores on PC1 (Figure 1A). As shown in the loading plot (Figure 1B), the descriptors which were responsible for the highest positive values on PC1 were nCb-, F01[N-N], F04[C-N], and F03[C-N]. These descriptors encode information on substituted benzene and nitrogen. This fits with the fact that NRB molecules contain more cyclic and nitro groups than RB molecules.[1] The descriptor “number of heavy atoms” (nHM) is also located in the positive side of PC1, and this can be related to the fact that RB molecules do not contain heavy atoms. On the other hand, most of the RB molecules had a low value on PC1. One of the descriptors which were correlated with low values on PC1 was the

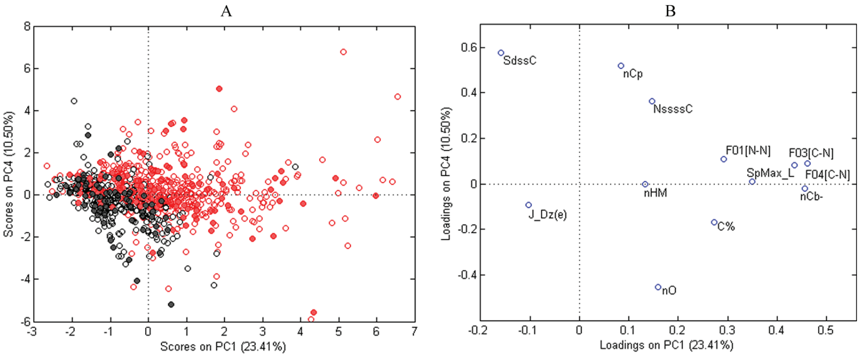


Figure 1. PCA of the descriptors used in the *k*NN model. Scores plot (A) and loadings plot (B) of the first and fourth principal components (explained variance equal to 34%). Ready biodegradable molecules are colored in black, and not ready biodegradables are in red; training molecules are marked with empty circles, and test molecules are marked with full circles. Labels of the molecular descriptors refer to symbols listed in Table 7.

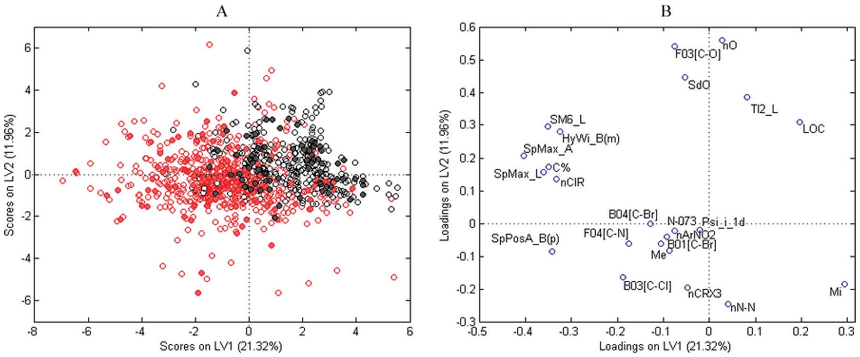


Figure 2. Scores plot (A) and loadings plot (B) of the first and second latent variables of the PLSDA model (explained variance equal to 33%). Ready biodegradable molecules are colored in black, and not ready biodegradables are in red; training molecules are marked with empty circles, test molecules are marked with full circles. Labels of the molecular descriptors refer to symbols listed in Table 7.

descriptor giving information about carbon with two single bonds and one double bond (SdssC). SdssC might be correlated with RB molecules because according to the literature, this class of molecules tend to be less branched compared to the NRB ones.[1] Also PC4 contained information on molecular branching, since two of the most important descriptors on this component were quaternary carbon and carbon bound to three terminal atoms (NssssC, nCp). Having the same upper right side orientation as the NRB, these two descriptors are therefore negatively correlated with biodegradability, thus confirming that branching decreases biodegradation.

3.3.2. Molecular Descriptors of the PLSDA Model. The PLSDA model included 23 descriptors (Table 7). The score plot of the first and second latent variables (LV1 and LV2), explaining together 33% of variance, is shown in Figure 2A.

RB molecules were grouped in the upper right side of the score plot, having positive scores on both LV1 and LV2. Matrix-based descriptors were placed on the extreme left side of the loadings plot (Figure 2B), thus correlating with NRB. These descriptors contain information on the molecular branching, and they might therefore be connected with NRB molecules since the degree of branching has an influence on a molecule's ability to biodegrade.[1] Descriptors with information on cycles, halogens, and nitrogen (nCIR, B03[C-Cl], F04[C-N], B04[C-Br], B01[C-Br], N-073, and nCRX3) had negative loadings on LV1. Cycles, halogens, and nitrogen are more often seen in NRB compared to RB compounds, and their connection with NRB molecules is therefore in alignment with knowledge from the literature.[1]

Descriptors related to the presence of oxygen (nO, F03[C-O], and SdO) had positive loadings on LV2 and, thus, were responsible for the separation of the RB and NRB classes on this latent variable. On the other side, descriptors related to the presence of nitrogen and halogens (B03[C-Cl], nCRX3, nNN) had negative loadings on LV2. This result fits with the knowledge on biodegradation, since the presence of functional groups with oxygen atoms increase biodegradability, while NRB molecules tend to have more nitrogen and halogens.[1]

3.3.3. Molecular Descriptors of the SVM Model. The SVM model included 14 molecular descriptors (Table 7). The results of the PCA analysis on this set of 14 descriptors are shown in

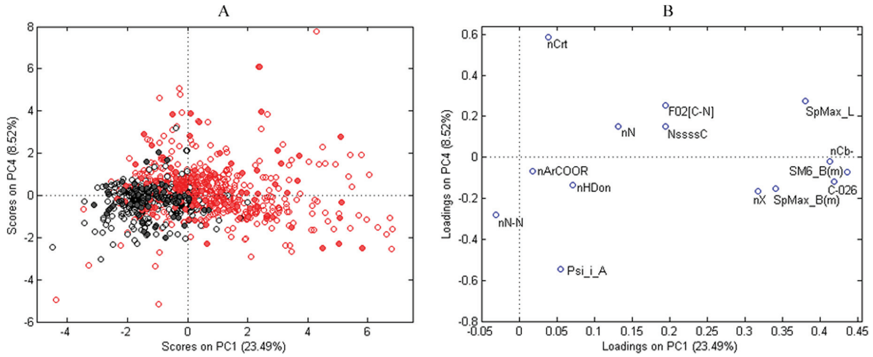


Figure 3. PCA of the descriptors used in the SVM model. Scores plot (A) and loadings plot (B) of the first and fourth principal components (explained variance equal to 32%). Ready biodegradable molecules are colored in black, and not ready biodegradables are in red; training molecules are marked with empty circles, and test molecules are marked with full circles. Labels of the molecular descriptors refer to symbols listed in Table 7.

Figure 3. The score plot of the first and fourth principal components (PC1 and PC4), explaining together 32% of variance, is shown in Figure 3A.

The first principal component was able to separate the two classes in the presence of some overlap (Figure 3A). In particular, RB molecules were characterized by negative scores, while the majority of NRB molecules were placed on the right side of PC1, having positive scores. The majority of the descriptors had positive loadings on PC1, as shown in Figure 3B. The most important descriptors for PC1 had information on molecular branching, aromatic groups and halogens (SpMax_L, SM6_B(m), C-026, nCb-, nX), as well as the presence of nitrogen (NssssC, F02[C-N], nN). All these descriptors are related to the NRB class of molecules, which is characterized by positive scores on PC1. This is in accordance with the literature since NRB molecules in general have more nitrogen groups and aromatic groups with halogens compared to RB molecules.[1] PC4 is less successful in separating the two classes. However, the nCrt descriptor, which encodes information about rings, had the greatest positive loading value on PC4. PC4 seems to show a tendency for lower values among the RB molecules compared to the second class, which is expected, since rings are more often seen in NRB.

Summarizing, descriptors selected in each QSAR model encoded similar information about the presence of halogens, chain branching, nitro groups, rings and some functional groups, which are related to biodegradability. Moreover, it was seen that the information on the relationships between chemical structures and biodegradability was consistent in the three QSAR models even though the descriptors were selected independently in each of the three proposed models.

3.4. Improving Interpretability of the Models.

The proposed models for ready biodegradability are based on several substructure descriptors but also on some matrix-based molecular descriptors (HyWi_B(m), J_Dz(e), SM6_B(m), SM6_L, SpMax_A, SpMax_B(m), SpMax_L, SpPosA_B(p), TI2_L). These descriptors are calculated from Laplacian (**L**), Barysz (**Dz**), and Burden (**B**) matrices, which are derived from the adjacency matrix (**A**). The adjacency matrix, or vertex adjacency matrix, is an important source for the calculation of molecular descriptors.[49] This is one of the fundamental graph theoretical matrices and represents the whole set of connections between adjacent pairs of atoms.[50] The adjacency matrix gives information about branching, which is demonstrated to be relevant for biodegradation modeling. This was confirmed in the PCA plots, where these matrix-based descriptors were always related to NRB compounds. Nevertheless, matrix-based molecular descriptors were never included before in already published biodegradation QSAR models and thus their relationship with biodegradation could not be directly verified and needs further investigation.

2D matrix-based descriptors are topological indices calculated by applying a set of basic algebraic operators to different graph-theoretical matrices representing the H-depleted molecular graph of molecules.[49]

Laplace matrix **L** is obtained by the difference between a diagonal vertex degree matrix and the adjacency matrix **A**:

$$[\mathbf{L}]_{ij} = \begin{cases} -1 & \text{if } (i, j) \in \mathrm{E(G)} \\ \delta_i & \text{if } i = j \\ 0 & \text{if } (i, j) \notin \mathrm{E(G)} \end{cases}$$

where δ_i is the ith vertex degree, that is, the number of vertices adjacent to vertex i and E(G) is the set of graph edges.

Burden matrices $\mathbf{B}(w)$ are augmented adjacency matrices defined to account for heteroatoms and bond multiplicity as the following:

$$[\mathbf{B}(w)]_{ij} = \begin{cases} \sqrt{\pi_{ij}^{*}} & \text{if } (i, j) \in \mathrm{E(G)} \\ \dfrac{w_i}{w_\mathrm{C}} & \text{if } i = j \\ 0.001 & \text{if } (i, j) \notin \mathrm{E(G)} \end{cases}$$

The diagonal elements are atomic carbon-scaled properties (e.g., mass (m), polarizability (p)); the off-diagonal elements corresponding to pairs of bonded atoms are the square roots of conventional bond orders π^* (i.e., 1, 2, 3, and 1.5 for single, double, triple, and aromatic bonds, respectively); all other matrix elements are set at 0.001.

Barysz matrices $\mathbf{Dz}(w)$ are weighted distance matrices that were defined on the basis of a generalization of Barysz

Table 8. List of OLS Models Built to Describe the Matrix-Based Descriptors[a]

descriptor	R^2	Q^2	model equations
HyWi_B(m)	0.94	0.94	−0.297 + 0.001MW + 0.745MWC01 + 3.446Eta_alpha_A
J_Dz(e)	0.82	0.79	−1.712 − 0.376nCIC + 2.212Xindex − 0.215NRS + 1.143piPC02
SM6_B(m)	0.87	0.86	3.432 + 0.005ZM1Mad + 1.927B01[C-Br] + 5.88Eta_alpha_A + 0.566piPC02
SM6_L	0.99	0.99	−0.416 + 1.04SRW08 + 3.683X0A
SpMax_A	0.96	0.95	0.495 + 0.267SRW08 − 0.087RDCHI
SpMax_B(m)	0.78	0.78	3.267 + 6.509B01[C-X] + 3.193B01[C-Br] + 0.135piPC04
SpMax_L	0.93	0.93	−8.339 + 1.678SRW08 + 9.693X1A − 1.49MWC01
SpPosA_B(p)	0.86	0.86	1.613 + 0.858Eta_alpha_A − 0.831Mi + 0.004C%
TI2_L	0.94	0.94	1.763 + 1.747MSD + 2.408RDCHI

[a]The squared correlation coefficient in fitting (R^2), in cross-validation with five cancellation groups (Q^2), and the model equations are provided.

weighting scheme in terms of conventional bond orders π^* and any atomic property:[51]

$$[\mathbf{Dz}(w)]_{ij} = \begin{cases} d_{ij}(w, \pi^*) & \text{if } i \neq j \\ 1 - \dfrac{w_C}{w_i} & \text{if } i = j \end{cases}$$

$$d_{ij}(w, \pi^*) = \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \frac{{w_C}^2}{w_{b(1)} w_{b(2)}} \right)$$

where w_C is any atomic property (e.g., Sanderson electronegativity (e)) of the carbon atom and w_i the corresponding value of the ith atom; $d_{ij}(w,\pi^*)$ is a weighted topological distance calculated by summing the edge weights over all bonds involved in the shortest path between vertices v_i and v_j; the subscripts $b(1)$ and $b(2)$ represent the two vertices incident to the considered bth edge.

The topological indices that were derived from these graph matrices and found to be related to ready biodegradability of organic compounds are briefly defined below.

The hyper-Wiener-like index, HyWi_B(m), is calculated according to the following formula:

$$\text{HyWi\_B(m)} = \ln\left\{ 1 + \frac{1}{2} \sum_{i=1}^{\text{nSK}} \sum_{j=i}^{\text{nSK}} \left([\mathbf{B}(\text{m})]_{ij}^{\,2} + [\mathbf{B}(\text{m})]_{ij} \right) \right\}$$

where nSK is the number of non-H atoms and defines the matrix dimension. This index is sensitive to molecule size and for a given size it takes minimum values for linear hydrocarbons while increases both with the number of heavy atoms and branching involving multiple bonds.

The Balaban-like index, J_Dz(e), is similar to Randić connectivity index but calculated with a normalization factor that makes it independent of the molecule size and cyclicity degree:

$$\text{J\_Dz(e)} = \frac{\text{nBO}}{\text{nCIC} + 1} \sum_{i=1}^{\text{nSK}-1} \sum_{j=i+1}^{\text{nSK}} a_{ij} [VS_i(\mathbf{Dz};\text{ e}) VS_j(\mathbf{Dz};\text{ e})]^{-1/2}$$

where the vertex degrees are replaced by the matrix row sums VS and elements a_{ij} of the adjacency matrix are introduced to account only for contributions from bonded atom pairs; nBO is the number of graph edges, and nCIC the number of independent rings in the molecule.

SpMax_A, SpMax_B(m), SpMax_L, SpPosA_B(p), SM6_B(m), SM6_L, and TI2_L are spectral indices calculated as function of the matrix eigenvalues.[52] In particular, SpMax is the leading eigenvalue, that is, the largest eigenvalue of the matrix spectrum, and SpPosA is the normalized spectral positive sum index, that is, the sum of the positive eigenvalues divided by the number of non-H atoms in order to reduce molecule size influence.[53] The leading eigenvalue of the adjacency matrix **A** (SpMax_A) is the well-known Lovasz−Pelikan index,[54] which was demonstrated to be related to molecular branching. Both SpMax_A and SpMax_L demonstrated to be able to characterize a large group of NRB molecules containing halogens (especially F and Cl) as the substituents in nonterminal positions along the molecular structure.

The spectral moment of sixth order (SM6) is the sum of the sixth power of all of the matrix eigenvalues. Since SM6_B(m) and SM6_L are derived from modified adjacency matrices, these indices are to some extent related to the number of self-returning walks of length six in the molecule, which can also be expressed as linear combinations of counts of certain fragments contained in the molecular graph.[55] These indices tend to increase with molecular branching and cyclicity. Moreover, the index SM6_B(m) is able to characterize a group of about 50 NRB compounds including heavy atoms (e.g., Sn and Br) and with large ramification or number of rings. The second Mohar index (TI2_L) is calculated as the inverse of the smallest nonzero eigenvalue of the Laplace matrix, which is weighted by the number of non-hydrogen atoms.[56] This index does not account for the presence of different heteroatoms in molecules but is very sensitive to structural features such as branching and cyclicity. It increases with the number of non-H atoms, and in a series of equal-sized molecules it discriminates between linear chains (high values) and branched/cyclic structures that typically are not ready biodegradable.

In order to improve the interpretability of the proposed QSAR models, matrix-based descriptors were further analyzed to elucidate the information they encode. For this purpose, ordinary least squares (OLS) regression was used to investigate the existing relationships between these targeted matrix-based descriptors and other DRAGON molecular descriptors, which were used as the independent variables in the regression models. A variable selection procedure based on GAs was carried out to search for the optimal subset of DRAGON descriptors related to each matrix-based descriptor. Regression models were optimized on the basis of the squared correlation coefficient Q^2 calculated in cross-validation with five cancellation groups.[57,58]

In addition to the already calculated descriptors (Table 3), the following DRAGON blocks were considered for this analysis: connectivity indices, topological information indices, walk and self-returning walk counts, and Extended Topochemical Atom (ETA) indices. These indices were selected as they encode to different extent information about molecular branching.[49] Statistics of the OLS models calculated for each matrix-based descriptor are collected in Table 8. Regression models included a maximum number of four descriptors; high and balanced performance in fitting (R^2) and cross-validation (Q^2) demonstrated good consistency as well as ability in describing the chemical information encoded by matrix-based descriptors. In particular, the regression models for HyWi_B-(m), SM6_L, SpMax_A, SpMax_L, and TI2_L had R^2 and Q^2 higher than 0.9; the models for SM6_B(m) and SpPosA_B(p) gave R^2 and Q^2 higher than 0.8, while just two models (J_Dz(e), SpMax_B(m)) had Q^2 between 0.78 and 0.79.

On the basis of these results, it could be concluded that the considered matrix-based descriptors mainly encode chemical information related to branching, cyclicity, and molecular size, which were demonstrated to be important factors related to biodegradability. In addition, the obtained OLS models also proved that matrix-based descriptors are highly information rich, since they were modeled by several other descriptors, each encoding different chemical information (Table 9). This feature makes matrix-based descriptors particularly interesting to QSAR modeling, since using descriptors able to encode different molecular features can lead to more parsimonious models including a limited number of variables.

Table 9. Molecular Descriptors Selected in the OLS Models Describing the Matrix-Based Descriptors

symbol	description	DRAGON Block
B01[C-Br]	presence/absence of C−Br at topological distance 1	2D atom pairs
B01[C-X]	presence/absence of C−X at topological distance 1	2D atom pairs
C%	percentage of C atoms	constitutional indices
Eta_alpha_A	eta average core count	ETA indices
Mi	mean first ionization potential (scaled on Carbon atom)	constitutional indices
MSD	mean square distance index (Balaban)	topological indices
MW	molecular weight	constitutional indices
MWC01	molecular walk count of order 1	walk and path counts
nCIC	number of rings (cyclomatic number)	ring descriptors
NRS	number of ring systems	ring descriptors
piPC02	molecular multiple path count of order 2	walk and path counts
piPC04	molecular multiple path count of order 4	walk and path counts
RDCHI	reciprocal distance sum Randic-like index	connectivity indices
SRW08	self-returning walk count of order 8	walk and path counts
X0A	average connectivity index of order 0	connectivity indices
X1A	average connectivity index of order 1	connectivity indices
Xindex	Balaban X index	information indices
ZM1Mad	first Zagreb index by Madan vertex degrees	topological indices

4. CONCLUSIONS

The aim of this study was to develop reliable classification QSAR models for ready biodegradability. Experimental values were collected from the MITI database and screened to obtain a consistent data set that meets the requirements of the OECD guidelines. The structure representations of the compounds, as well as the collected experimental data, were accurately verified. The resulting data set was split into training and test sets before modeling. Genetic algorithms coupled with three different classification algorithms (*k*NN, PLSDA, and SVM) were applied in order to select the optimal subset of molecular descriptors. The three models and the derived consensus analysis demonstrated good statistics in fitting and cross-validation as well as high accuracy in prediction for the test set with respect to already published models on biodegradation. The lowest ER in classification was reached by means of *k*NN, which gave an ER equal to 0.12 for the test set, and consensus analysis, which gave an error rate equal to 0.06 with 23% of not assigned molecules. The developed models were further validated using an external validation set collected from different sources, and good classification performances were obtained. The proposed models showed a balance between specificity and sensitivity values, as well as similar performances in training, test, and external validation sets, which can indicate the absence of overfitting. The potential relationships between the selected molecular descriptors and biodegradability were evaluated by comparing with information from the literature.

Matrix-based molecular descriptors, which were used for the first time to model biodegradability, were further analyzed. The information they encoded was evaluated by means of regression OLS models based on other types of molecular descriptors. Relationships between matrix-based descriptors and biodegradability were highlighted, since they contained information about molecular branching and size. In general, this family of molecular descriptors appeared to be interesting for QSAR modeling, since they were information rich and thus by using them the total number of descriptors to be used to model a defined endpoint could be reduced.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Training, test, and external validation sets, CAS numbers, SMILES structures, and experimental class (ready/not ready biodegradable: RB/NRB) of the molecules as well as the values of molecular descriptors included in the QSAR models. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: davide.ballabio@unimib.it. Mailing address: Department of Environmental Sciences, University of Milano-Bicocca

P.zza della Scienza, 1-20126 Milano, Italy. Telephone: +39-02-6448.2801. Fax: +39-02-6448.2839.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007-2013] under Grant Agreement no. 238701 of the Marie Curie ITN Environmental Chemoinformatics (ECO) project: http://www.eco-itn.eu/.

■ REFERENCES

(1) Boethling, R. S. Designing Biodegradable Chemicals. In *Designing Safer Chemicals*; DeVito, S. C., Garrett, R. L., Eds.; American Chemical Society: Washington, DC, 1996; Vol. *640*, pp 156−171.

(2) European Commission. Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission Regulation (EC) No 1488/94 as well as Council Directive 76/769/EEC. *Off. J. Eur. Union* **2006**, *L396*, 1−849.

(3) Rorije, E.; Loonen, H.; Müller, M.; Klopman, G.; Peijnenburg, W. J. G. M. Evaluation and application of models for the prediction of ready biodegradability in the MITI-I test. *Chemosphere* **1999**, *38*, 1409−1417.

(4) Allanou, R.; Hansen, B. G.; Van der Bilt, Y. *Public availability of data on EU high production volume chemicals*; European Communities: Italy, 1999; Report EUR 18996 EN.

(5) Pavan, M.; Worth, A. P. Review of estimation models for biodegradation. *QSAR Comb. Sci.* **2008**, *27*, 32−40.

(6) Raymond, J. W.; Rogers, T. N.; Shonnard, D. R.; Kline, A. A. A review of structure-based biodegradation estimation methods. *J. Hazard. Mater.* **2001**, *84*, 189−215.

(7) Jaworska, J. S.; Boethling, R. S.; Howard, P. H. Recent developments in broadly applicable structure-biodegradability relationships. *Environ. Toxicol. Chem.* **2003**, *22*, 1710−1723.

(8) Baker, J. R.; Gamberger, D.; Mihelcic, J. R.; Sabljić, A. Evaluation of artificial intelligence based models for chemical biodegradability prediction. *Molecules* **2004**, *9*, 989−1003.

(9) Geating, J. *Literature study of the biodegradability of chemicals in water*; U.S. EPA: Cincinnati, OH, 1981; Report EPA-600/2-81-175.

(10) Niemi, G. J.; Veith, G. D.; Regal, R. R.; Vaishnav, D. D. Structural features associated with degradable and persistent chemicals. *Environ. Toxicol. Chem.* **1987**, *6*, 515−527.

(11) Boethling, R. S.; Sabljić, A. Screening-level model for aerobic biodegradability based on a survey of expert knowledge. *Environ. Sci. Technol.* **1989**, *23*, 672−679.

(12) Howard, P. H.; Boethling, R. S.; Stiteler, W.; Meylan, W.; Beauman, J. Development of a predictive model for biodegradability based on BIODEG, the evaluated biodegradation data base. *Sci. Total Environ.* **1991**, *109−110*, 635−641.

(13) Howard, P. H.; Boethling, R. S.; Stiteler, W. M.; Meylan, W. M.; Hueber, A. E.; Beauman, J. A.; Larosche, M. E. Predictive model for aerobic biodegradability developed from a file of evaluated biodegradation data. *Environ. Toxicol. Chem.* **1992**, *11*, 593−603.

(14) Boethling, R. S.; Howard, P. H.; Meylan, W.; Stiteler, W.; Beauman, J.; Tirado, N. Group contribution method for predicting probability and rate of aerobic biodegradation. *Environ. Sci. Technol.* **1994**, *28*, 459−465.

(15) Gamberger, D.; Horvatić, D.; Sekušak, S.; Sabljić, A. Applications of experts' judgement to derive structure-biodegradation relationships. *Environ. Sci. Pollut. Res. Int.* **1996**, *3*, 224−228.

(16) Loonen, H.; Llndgren, F.; Hansen, B.; Karcher, W.; Nlemela, J.; Hiromatsu, K.; Takatsuki, M.; Peunenburg, W.; Rorije, E.; Struijs, J. Prediction of biodegradability from chemical structure: Modeling of ready biodegradation test data. *Environ. Toxicol. Chem.* **1999**, *18*, 1763−1768.

(17) Tunkel, J.; Howard, P. H.; Boethling, R. S.; Stiteler, W.; Loonen, H. Predicting ready biodegradability in the Japanese Ministry of International Trade and Industry test. *Environ. Toxicol. Chem.* **2000**, *19*, 2478−2485.

(18) Huuskonen, J. Prediction of biodegradation from the atom-type electrotopological state indices. *Environ. Toxicol. Chem.* **2001**, *20*, 2152−2157.

(19) Jaworska, J.; Dimitrov, S.; Nikolova, N.; Mekenyan, O. Probabilistic assessment of biodegradability based on metabolic pathways: catabol system. *SAR QSAR Environ. Res.* **2002**, *13*, 307−323.

(20) Sedykh, A.; Klopman, G. Data analysis and alternative modelling of MITI-I aerobic biodegradation. *SAR QSAR Environ. Res.* **2007**, *18*, 693−709.

(21) Alikhanidi, S.; Takahashi, Y. Pesticide Persistence in the Environment-Collected Data and Structure-Based Analysis. *J. Comp. Chem. (Japan)* **2004**, *3*, 59−70.

(22) Cheng, F.; Ikenaga, Y.; Zhou, Y.; Yu, Y.; Li, W.; Shen, J.; Du, Z.; Chen, L.; Xu, C.; Liu, G.; Lee, P. W.; Tang, Y. In silico assessment of chemical biodegradability. *J. Chem. Inf. Model.* **2012**, *52*, 655−669.

(23) Klopman, G.; Balthasar, D. M.; Rosenkranz, H. S. Application of the computer-automated structure evaluation (CASE) program to the study of structure-biodegradation relationships of miscellaneous chemicals. *Environ. Toxicol. Chem.* **1993**, *12*, 231−240.

(24) Nendza, M. Prediction of Persistence. In *Predicting Chemical Toxicity and Fate*; Cronin, M., Livingstone, D., Eds.; CRC Press: Boca Raton, FL, 2004.

(25) Gramatica, P.; Cassani, S.; Roy, P. P.; Kovarich, S.; Yap, C. W.; Papa, E. QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae. *Mol. Inform.* **2012**, *31*, 817−835.

(26) CHRIP National Institute of Technology and Evaluation (NITE) of Japan, Chemical Risk Information Platform http://www.safe.nite.go.jp/english/kizon/KIZON_start_hazkizon.html (accessed Jan 16, 2012).

(27) Organisation for Economic Co-operation and Development. *Test No. 301: Ready Biodegradability*; OECD Publishing: Paris, 1992.

(28) *EPA 712-C-98-076 Fate, Transport, and Transformation Test Guidelines*; EPA: Washington, D.C., 1998; OPPTS 835.3110.

(29) *ChemSpider*; Royal Society of Chemistry: Cambridge, http://www.chemspider.com/ (accessed Oct 29, 2012).

(30) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer: New York, 2007.

(31) Rorabacher, D. B. Statistical treatment for rejection of deviant values: critical values of Dixon's "Q" parameter and related subrange ratios at the 95% confidence level. *Anal. Chem.* **1991**, *63*, 139−146.

(32) *DRAGON* (Software for Molecular Descriptor Calculations) version 6.0.28; Talete srl: Milano, Italy, 2012.

(33) Kowalski, B. R.; Bender, C. F. The K-nearest neighbor classification rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation. *Anal. Chem.* **1972**, *44*, 1405−1411.

(34) Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109−130.

(35) Ståhle, L.; Wold, S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *J. Chemom.* **1987**, *1*, 185−196.

(36) Cortes, C.; Vapnik, V. Support-Vector Networks. *In Machine Learning* **1995**, 273−297.

(37) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory; COLT '92*, Pittsburgh, PA, July 27−29; ACM: New York, 1992; pp 144−152.

(38) Leardi, R.; Lupiáñez González, A. Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemom. Intell. Lab. Syst.* **1998**, *41*, 195−207.

(39) Baurin, N.; Mozziconacci, J.-C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.

(40) Votano, J. R.; Parham, M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A.; Xie, Q.; Tong, W. Three new consensus QSAR models for the prediction of Ames genotoxicity. *Mutagenesis* **2004**, *19*, 365−377.

(41) Ganguly, M.; Brown, N.; Schuffenhauer, A.; Ertl, P.; Gillet, V. J.; Greenidge, P. A. Introducing the Consensus Modeling Concept in Genetic Algorithms: Application to Interpretable Discriminant Analysis. *J. Chem. Inf. Model.* **2006**, *46*, 2110−2124.

(42) Hewitt, M.; Cronin, M. T. D.; Madden, J. C.; Rowe, P. H.; Johnson, C.; Obi, A.; Enoch, S. J. Consensus QSAR Models: Do the Benefits Outweigh the Complexity? *J. Chem. Inf. Model.* **2007**, *47*, 1460−1468.

(43) Environment Canada DSL (Domestic Substances List) http://www.ec.gc.ca/lcpe-cepa/default.asp?lang=En&n=5F213FA8-1&wsdoc=D031CB30-B31B-D54C-0E46-37E32D526A1F (accessed Nov 4, 2012).

(44) CEPA 1999 Canadian Environmental Protection Act http://laws-lois.justice.gc.ca/eng/regulations/SOR-2000-107/page-1.html#footnotea_e-ID0EFBCA (accessed Nov 14, 2012).

(45) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; National Taiwan University, Department of Computer Science: Taipei, Taiwan, 2001.

(46) *MATLAB*, version 7.13.0.564; MathWorks: Natick, MA, 2011; www.mathworks.com.

(47) Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(48) Manallack, D. T.; Livingstone, D. J. Artificial neural networks: Application and chance effects for QSAR data analysis. *Med. Chem. Res.* **1992**, *2*, 181−190.

(49) Todeschini, R.; Consonni, V. *Molecular descriptors for chemoinformatics*; Wiley-VCH: New York, 2009.

(50) Trinajstić, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992; pp 225−273.

(51) Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412−1422.

(52) Consonni, V.; Todeschini, R. New spectral indices for molecule description. *Match* **2008**, *60*, 3−14.

(53) Balaban, A. T.; Ciubotariu, D.; Medeleanu, M. Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 517−523.

(54) Lovasz, L.; Pelikan, J. On the eigenvalue of trees. *Period Math Hung* **1973**, *3*, 175−182.

(55) Barysz, M.; Trinajstić, N. A novel approach to the characterization of chemical structures. Int. *J. Quantum Chem. Quant. Chem. Symp.* **1984**, *18*, 661−673.

(56) Trinajstić, N.; Babic, D.; Nikolić, S.; Plavšić, D.; Amić, D.; Mihalić, Z. The Laplacian matrix in chemistry. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 368−376.

(57) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the Definition of the Q2 Parameter for QSAR Validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669−1678.

(58) Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* **2010**, *24*, 194−201.

Annex III: Applicability domain approaches

Sahigara, F.; Mansouri, K.; Ballabio, D.; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules 2012, 17, 4791–4810.

Molecules **2012**, *17*, 4791-4810; doi:10.3390/molecules17054791

OPEN ACCESS

molecules

ISSN 1420-3049

www.mdpi.com/journal/molecules

Article

Comparison of Different Approaches to Define the Applicability Domain of QSAR Models

Faizan Sahigara, Kamel Mansouri, Davide Ballabio, Andrea Mauri, Viviana Consonni and Roberto Todeschini *

Milano Chemometrics and QSAR Research Group, Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1-20126 Milano, Italy;
E-Mails: faizan.sahigara@unimib.it (F.S.); kamel.mansouri@unimib.it (K.M.); davide.ballabio@unimib.it (D.B.); andrea.mauri@unimib.it (A.M.); viviana.consonni@unimib.it (V.C.)

* Author to whom correspondence should be addressed; E-Mail: roberto.todeschini@unimib.it; Tel.: +39-02-6448-2820; Fax: +39-02-6448-2839.

Received: 14 March 2012; in revised form: 13 April 2012 / Accepted: 16 April 2012 / Published: 25 April 2012

Abstract: One of the OECD principles for model validation requires defining the Applicability Domain (AD) for the QSAR models. This is important since the reliable predictions are generally limited to query chemicals structurally similar to the training compounds used to build the model. Therefore, characterization of interpolation space is significant in defining the AD and in this study some existing descriptor-based approaches performing this task are discussed and compared by implementing them on existing validated datasets from the literature. Algorithms adopted by different approaches allow defining the interpolation space in several ways, while defined thresholds contribute significantly to the extrapolations. For each dataset and approach implemented for this study, the comparison analysis was carried out by considering the model statistics and relative position of test set with respect to the training space.

Keywords: QSAR; model validation; Applicability Domain; interpolation space

1. Introduction

The quantitative relationship between chemical structures and their properties can be established mathematically by means of QSARs and thus, given that the structural information is available, QSAR models can be used theoretically to predict the properties for those chemicals [1]. Due to increasing application of such QSAR models, there had been rising concerns with respect to their predictions [2]. Derivation of QSAR models is based primarily on training sets which are structurally limited and thus their applicability to the query chemicals is limited. In other words, the model can provide more reliable prediction for the external compounds that fall within these structural limitations [3].

A new European legislation on chemicals—REACH (Registration, Evaluation, Authorization and restriction of Chemicals) came into force in 2007, which deals with risk assessment of chemicals for their safe use, thus contributing to the human health and environment [4]. This law allows and encourages the use of QSAR model predictions when the experimental data are not sufficiently available or as supplementary information, provided validity of the model is justified [5]. Five OECD principles for QSAR validation adopted in November 2004 are the requisites of any given model proposed for regulatory use and can be significant to demonstrate the validity of QSAR models, which is crucial for REACH implementation.

According to these OECD principles, the QSAR model should have: (1) a defined end point; (2) an unambiguous algorithm; (3) a defined domain of applicability; (4) appropriate measures for goodness-of-fit, robustness and predictivity and (5) a mechanistic interpretation, if possible [6]. The principles, in general, provide user with all the essential information regarding end-point being predicted, model algorithm used, scope of the model and associated limitations, model performance and understanding of how the model descriptors are associated with predicted endpoint [5]. This paper primarily focuses on the third OECD principle that deals with defining the Applicability Domain (AD) of a QSAR model.

The principle of Applicability Domain requires users to define the model limitations with respect to its structural domain and response space. As discussed above, the reliable QSAR predictions are limited generally to the chemicals that are structurally similar to ones used to build that model [7–9]. The query chemicals that satisfy the scope of the model are considered as within the AD and classified as interpolated whereas the rest are extrapolations and thus, outside the AD. Reliability in a given model is higher for predictions falling within the AD and it is most likely to be unreliable for the extrapolations. This implies that the fourth OECD principle dealing with model accuracy is highly dependable on the model's AD since here the chemical space associated with reliable predictions is identified. Molecular descriptors used to build the model also play a significant role in defining the AD. Thus, if a query chemical differs in terms of the structural limitations defined by the training set, it can be considered as an outlier for that chemical space.

Defining a model's AD is essential in order to determine the subspace of chemical structures that could be predicted reliably. In other words, the degree of generalization of a predictive model depends on how broad the domain of applicability is. If the domain is too restricted, this implies the model is capable of giving reliable predictions only for limited chemical structures. Also, for regulatory purposes, like in REACH, it is essential for the user to provide all possible documentation about the model's AD. This is beneficial for the user to see if the endpoint for the chemical structures under evaluation can be reliably predicted. Also, for the cases where several QSAR models are available for chemicals of interest, the knowledge of AD can be applied to compare how reliable the predictions could be for different models [1].

Characterization of the interpolation space is very significant to define the AD for a given QSAR model. Several AD approaches have been already proposed and primarily they all differ in the way how they characterize the interpolation space defined by the descriptors used. They can be classified into following four major categories based on the methodology used for interpolation space characterization in the model descriptor space: Range-based methods, Geometric methods, Distance-based methods and Probability Density Distribution based methods [1–5].

In this study, the above mentioned AD approaches are discussed and compared, focusing on the methodology used and criteria followed to consider a query structure to be within (or outside) the Applicability Domain. The major goal of this paper is to provide a detailed comparison of the results obtained, using these different AD approaches on some selected datasets. Two models from the CAESAR project, which predict the bioconcentration factor (BCF), were chosen as the case study [10,11]. Apart from their own test sets, an alternative test set from EPI Suite package BCFBAF v3.00 was chosen to facilitate further evaluation of AD approaches [12,13]. The number of test compounds considered outside AD for different approaches was calculated and the reliability of these results was further interpreted by analyzing both, the prediction statistics and the relative position of test compounds with respect to the training space. For all distance measures in this study, the pattern of test compounds considered outside the AD was understood by implementing the distance-based approaches with several threshold defining strategies that considered both, the distances of training compounds from their mean as well as the average distances of training compounds from their first 5 nearest neighbors. Finally, comparing the results derived with this analysis, most preferred thresholds for distance-based approaches were chosen for their overall comparison with other AD approaches.

2. Applicability Domain Methods

The basis for interpolation is to predict the function value at a given point when the values at neighboring points are known. There are several descriptor based approaches by which the interpolation regions in multivariate space can be estimated for QSAR models. In a given p-dimensional descriptor space, estimations for new query chemicals are then obtained using the training data [1]. All the approaches used for this study were implemented using MATLAB [14] and are discussed briefly in this section informing their main features to define the interpolation space as well as the thresholds criterion used.

2.1. Range-Based and Geometric Methods

These are considered as the simplest methods to characterize a model's interpolation space.

2.1.1. Bounding Box

This approach considers the range of individual descriptors used to build the model. Assuming a uniform distribution, resulting domain of applicability can be imagined as a Bounding Box which is a p-dimensional hyper-rectangle defined on the basis of maximum and minimum values of each descriptor used to build the model. The sides of this hyper-rectangle are parallel with respect to the coordinate axes. However, there are several drawbacks associated with this approach: since only descriptor ranges are taken into consideration, empty regions in the interpolation space cannot be identified and also the correlation between descriptors cannot be taken into account [1,2].

2.1.2. PCA Bounding Box

PCA transforms the original data into a new coordinate system by the rotation of axes, such that the new axes are orthogonal to each other and aligned in the direction having maximum variance within the data. These new axes are called Principal Components (PCs) representing the maximum variance within the dataset [15]. A *M*-dimensional hyper-rectangle (where *M* is the number of significant components) is obtained similar to the previous approach by considering the projection of the molecules in the principal component space, however taking into account the maximum and minimum values for the PCs. The implementation of Bounding Box with PCA can overcome the problem of correlation between descriptors but empty regions within the interpolation space still remains an issue [1,2,5]. Moreover, selection of appropriate number of components is significant to implement this approach.

2.1.3. Convex Hull

With this approach, interpolation space is defined by the smallest convex area containing the entire training set. Implementing a Convex Hull can be challenging with increasing data complexity [16]. For two or three dimensional data, several algorithms are proposed; however, increase in dimensions contribute to order of complexity. In addition, set boundaries are analyzed without considering the actual data distribution. Similar to the Range-based approaches, Convex Hull cannot identify the potential internal empty regions within the interpolation space [1,2].

2.2. Distance-Based Methods

These approaches calculate the distance of query compounds from a defined point within the descriptor space of the training data. The general idea is to compare distances measured between defined point and the dataset with a pre-defined threshold. The threshold is a user defined parameter and is set to maximize the separation of dense regions within the original data. However, the cut-off value does not entirely reflect the actual data density [1–5]. No strict rules were evident from the literature about defining thresholds for distance-based approaches and thus it is up to the user how to define them. In this study, for all the distance measures, several possible threshold defining strategies were considered, the derived results were compared and finally the appropriate thresholds were chosen to overall compare their results with the ones derived from Range-based, Geometric and Probability Density Distribution based approaches. Some commonly used and most useful distance measures in QSAR studies include Mahalanobis, Euclidean and City Block distances [2,5].

The unique feature associated with Mahalanobis measure is the co-variance matrix which can handle the correlated descriptors. The other two distance measures lack this characteristic and thus require an additional treatment; for example, PC rotation to correct for the correlated axes. Iso-distance contours constitute the regions having constant distance measures and generally their shape differs with approaches according to the distance measure considered, for example, ellipsoids for Mahalanobis and spherical in case of Euclidean distances [2].

Apart from them, similar approaches based on leverage are quite recommended for defining AD of a QSAR model [17]. Leverage of a query chemical is proportional to its Mahalanobis distance measure from the centroid of the training set. The leverages are calculated for a given dataset **X** by obtaining the leverage matrix (**H**) with the equation below:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}} \qquad (1)$$

where **X** is the model matrix while $\mathbf{X}^{T}$ is its transpose matrix.

Diagonal values in the **H** matrix represent the leverage values for different points in a given dataset. Compounds far from the centroid will be associated with higher leverage and are considered to be influential in model building. Leverage is proportional to Hotellings T^2 statistic and Mahalanobis distance measure but can be applied only on the regression models. The approach can be associated with a warning leverage, generally three times the average of the leverage that corresponds to *p/n* where *p* is the number of model parameters while *n* is the number of training compounds. A query chemical with leverage higher than the warning leverage can be associated with unreliable predictions. Such chemicals are outside the descriptor space and thus be considered outside the AD [1,2,5]. In this study, the corresponding Mahalanobis measures were used.

K nearest Neighbors Approach

This approach is based on providing similarity measure for a new chemical with respect to the compounds within the training space. The similarity is accessed by finding the distance of a query chemical from nearest training compound or its distances from *k* nearest neighbors in the training set. If these distance values are within the user defined threshold, the query chemical with higher similarity is indicated to have higher number of training neighbors and therefore, is considered to be reliably predicted. Thus, similarity to the training set molecules is significant for this approach in order to associate a query chemical with reliable prediction [9].

2.3. Probability Density Distribution-Based Method

Considered as one of the most advanced approaches for defining AD, these methods are based on estimating the Probability Density Function for the given data. This is feasible by both, parametric methods that assume standard distribution and non parametric methods which do not have any such assumptions concerning the data distribution. A main feature of these approaches is their ability to identify the internal empty regions. Moreover, if needed, the actual data distribution can be reflected by generating concave regions around the interpolation space borders [1,2].

Generally these approaches are implemented by estimating probability density of the dataset followed by identifying Highest Density Region that consists of a known fraction (given as user input) from the total probability mass [1].

Potential is created for each molecule in the training set such that it is highest for that molecule and decreases with distance. Once the potential is calculated for all the compounds, global potential is obtained by adding the individual potentials thus indicating the probability density [18,19].

There are several types of potential functions; however, for this study Gaussian function was considered. Given two molecules x_i and x_j, it can be determined as below:

$$\Phi\left(x_i, x_j\right) = \frac{1}{\sqrt{(2\pi)}s} \cdot exp\left[\frac{-1}{\left(2s^2\right)\left(x_i - x_j\right)^2}\right] \tag{2}$$

where $\Phi\left(x_i, x_j\right)$ is the potential induced on x_j by x_i and width of the curve is defined by smoothing parameter s. The cut off value associated with Gaussian potential functions, namely f_p, can be calculated by methods based on sample percentile [18]:

$$f_p = f_i + (\mathrm{q} - j)\left(f_{j+1} - f_j\right) \tag{3}$$

with $\mathrm{q} = \mathrm{p} \times \frac{n}{100}$, where p is the percentile value of probability density, n is the number of compounds in the training set and j is the nearest integer value of q. Test compounds with potential function values lower than this threshold are considered outside the AD.

2.4. Other AD Approaches

Apart from the AD strategies discussed above, several other approaches were published in literature to define the AD of QSAR models, some of which are briefly discussed below. These approaches were not considered for this comparative study since the analysis was limited to the classical AD methodologies used for interpolation space characterization in the model descriptor space.

2.4.1. Decision Trees and Decision Forests Approach

Based on the consensus prediction of Decision Trees (DT), this approach specifies the AD in terms of prediction confidence and domain extrapolation. The main idea here is to minimize the overfitting which can be achieved by combining the DTs and keeping the differences within different DTs to maximum possible. Predictions from all the combined DTs are averaged in order to determine the prediction confidence for a given compound while domain extrapolation provides the prediction accuracy for that compound outside the training space [1,20,21].

2.4.2. Stepwise Approach to Determine Model’s AD

This approach is divided into four stages applied in a sequential manner. In the first stage, a query chemical is checked to fall within the range of variation of the physicochemical properties of training set compounds. During the second stage, structural similarity is found within the chemicals that are correctly predicted by the model. The third deals with mechanistic check while the reliability of simulated metabolism is taken into account in the final stage. To be considered within the AD, a query compound is required to satisfy all the conditions specified within these four stages. As a part of this rigorous approach, a chemical is evaluated for similarity, metabolic and mechanistic check, thus addressing the reliability of predictions and allowing a better assessment of model’s AD [3,5].

2.5. Models and Test Sets

This section deals with models and datasets selected for the comparison of the different AD approaches.

2.5.1. CAESAR Models

Bioconcentration factor, which is one of the most important endpoints for environmental fate of chemicals, was chosen for comparing the results derived from the different AD approaches considered in this study. As the procedure requires deep knowledge of the model and also information about its datasets and building methods, two already existing models to predict BCF were considered [10,11].

The QSAR models (Model 2 and Model 5) used in this study were the selected best two BCF models developed under the EU project CAESAR taking into account the REACH requirements [10]. These two models based on Radial Basis Function Neural Network (RBFNN) [22] were rebuilt, each with five descriptors that were calculated using Dragon 5.5 [23].The obtained statistics are summarized in Table 1.

Table 1. An overview of selected CAESAR models.

Model	Training set		Test set	
	***R*^{2} [(a)]**	***RMSE* [(b)]**	***Q*^{2} [(c)]**	***RMSEP* [(d)]**
1) Model 2	0.804	0.591	0.797	0.600
2) Model 5	0.810	0.581	0.774	0.634

[(a)] Determination coefficient *R*^{2}; [(b)] Root-mean-square error *RMSE*; [(c)] Predictive squared correlation coefficient *Q*^{2}; [(d)] Root-mean-square error of prediction *RMSEP*.

2.5.2. CAESAR and EPI Suite Test Sets

The CAESAR dataset consisted of 473 compounds, randomly divided into a training set of 378 compounds and a test set of 95 compounds, as explained in the original study [10]. The *Q*^{2} and *RMSEP* values for the test sets of CAESAR Model 2 and Model 5 are reported in Table 1.

For a better evaluation of AD approaches, in addition to the CAESAR test set, the validation set of the BCF model from EPI Suite package BCFBAF was selected as an additional test set [12,13]. This test set was comprised of 158 compounds, from which one compound was discarded due to structure inadequacy while other 49 compounds were not considered due to overlapping with the CAESAR training set compounds.

3. Results and Discussion

For the AD approaches discussed earlier, general rules to define thresholds are discussed in the literature except for distance-based approaches. Thresholds can be defined in several ways for the distance-based approaches, thus resulting in an ambiguity over selection of appropriate thresholds for this study. As a result, before an overall comparison of results with different AD approaches could be performed, thresholds for distance-based approaches had to be finalized.

To decide upon appropriate thresholds for distance-based approaches, several threshold defining strategies were implemented for the different distance measures considered in this study. All these strategies discussed below required calculating distances of training compounds from their centroid. To evaluate further possibilities, the study was extended implementing these strategies however considering average distance of each training compound from their first 5 nearest neighbors. Model statistics were recorded each time and the most appropriate distance based thresholds were then selected from above mentioned results for all distance measures considered in this study. Until this point, all the four categories of AD approaches were associated with appropriate thresholds and finally subjected to overall comparison of results.

The results were tabulated informing the model's statistics for each AD approach on the compounds considered inside the applicability domain using the following parameters:

i) Number of test compounds considered outside the domain of applicability;

ii) Predictive squared correlation coefficient Q^2 [24,25]:

$$Q^2 = 1 - \frac{\left[\sum_{i=1}^{n_{EXT}} (\hat{y}_i - y_i)^2 \right] / n_{EXT}}{\left[\sum_{i=1}^{n_{TR}} (y_i - \overline{y}_{TR})^2 \right] / n_{TR}} \tag{4}$$

where $\hat{y}_i$ is the predicted value for the *i*-th compound and y_i its experimental value; n_{TR} is the number of compounds in the training set and n_{EXT} the number in the test set; $\overline{y}_{TR}$ is the mean response of the training set. Moreover, in order to somehow quantify the role of the compounds considered inside and outside AD, $\Delta RMSEP$ was defined by the following equation:

$$\Delta RMSEP = RMSEP_{OUT} - RMSEP_{IN} \tag{5}$$

where $RMSEP_{OUT}$ is the root mean square error in prediction for the test compounds outside AD, while $RMSEP_{IN}$ is the root mean square error in prediction for the test compounds inside AD. Negative values indicate that the compounds detected outside AD are predicted better than the compounds inside AD, thus highlighting some possible drawbacks in the definition of interpolation space. On the contrary, positive values of $\Delta RMSEP$ indicate a reliable partition for the compounds detected as inside and outside AD.

Multi Dimensional Scaling (MDS) was used to visualize the relative position of test compounds with respect to the training space. MDS enables the representation of *p*-dimensional data by means of a 2D plot. The implementation allowed a better understanding of how the interpolation space was characterized and if the compounds outside the AD were more concentrated around the training set extremities or not.

3.1. Defining Thresholds for Distance-Based AD Approaches

Initially, the distances of training compounds from their centroid were calculated and from this resulting vector, the maximum and average distance value (*maxdist* and *d*) were derived. The first threshold strategy defined the AD considering *maxdist* as threshold [2]. The second and third strategies considered twice and thrice the values of *d* as their thresholds, respectively. The fourth strategy performed percentile approach on the above derived vector of distances sorted in ascending order and the distance value corresponding to 95 percentile (*p95*) was chosen as the threshold. Finally, the fifth strategy (*dsz*) considered average distance *d* as well as the standard deviation from the distance vector (*std*) and the threshold was then defined as $d + std \times z$, where *z* is the arbitrary parameter and is set to 0.5 as default value [26].

For all the cases, distance of a test compound from the training set centroid is compared with the defined threshold. If the distance of this test compound from the training set centroid is less than or equal to the threshold value, it is considered inside the AD. Thus, these approaches differ the way in which thresholds are derived, however the principle behind considering a given test compound to be inside or outside AD remains the same. Results derived with all the four threshold strategies are shown in Table 2 for CAESAR Model 2 considering different distance measures.

Table 2. Statistics for CAESAR Model 2 implementing distance-based approaches with different thresholds. For the acronyms *maxdist*, *d*, *p95*, *dsz*, and *ΔRMSEP*, refer to text.

		Compounds outside the AD		Q^2		***ΔRMSEP***	
Approach	***Thresholds***	***CAESAR out of 95 (%)***	***EPI Suite out of 108 (%)***	***CAESAR***	***EPI Suite***	***CAESAR***	***EPI Suite***
Euclidean (*maxdist*)	0.942	0 (0.0)	4 (3.7)	0.797	0.703	-	1.436
Euclidean (*3*d*)	1.018	0 (0.0)	1 (0.9)	0.797	0.676	-	0
Euclidean (*2*d*)	0.679	7 (7.4)	12 (11.1)	0.802	0.718	0.146	0.753
Euclidean (*p95*)	0.663	7 (7.4)	12 (11.1)	0.802	0.718	0.146	0.753
Euclidean (*dsz*)	0.423	15 (15.8)	36 (33.3)	0.791	0.741	−0.064	0.381
CityBlock (*maxdist*)	1.472	0 (0.0)	1 (0.9)	0.797	0.676	-	2.713
CityBlock (*3*d*)	1.863	0 (0.0)	0 (0.0)	0.797	0.616	-	-
CityBlock (*2*d*)	1.242	3 (3.1)	6 (5.5)	0.804	0.699	0.267	−1.049
CityBlock (*p95*)	1.084	8 (8.4)	11 (10.1)	0.801	0.705	0.068	0.717
CityBlock (*dsz*)	0.748	18 (18.9)	38 (35.1)	0.786	0.739	−0.093	0.361
Mahalanobis (*maxdist*)	6.614	0 (0.0)	0 (0.0)	0.797	0.616	-	-
Mahalanobis (*3*d*)	6.027	0 (0.0)	0 (0.0)	0.797	0.616	-	-
Mahalanobis (*2*d*)	4.018	6 (6.3)	5 (4.6)	0.791	0.624	−0.174	0.162
Mahalanobis (*p95*)	4.034	6 (6.3)	5 (4.6)	0.791	0.624	−0.174	0.162
Mahalanobis (*dsz*)	2.497	21 (22.1)	27 (25.0)	0.778	0.706	−0.138	0.354

No test compounds emerged outside the AD with first two strategies considering CAESAR test set, due to the higher threshold values; however, comparing the model statistics with the other approaches, this probably implies some possible drawbacks of these strategies in defining the interpolation space. Comparable results were derived considering the third and fourth strategies which imply the thresholds corresponding to twice the value of *d* and that corresponding to 95 percentile converged significantly for both the test sets. Model statistics improved in most of the cases, thus reflecting a reasonable choice of compounds outside AD. The final strategy taking into account also the standard deviation provided the maximum number of test compounds outside the AD, however with no (or significant) improvement to the model statistics for both the test sets. A similar pattern was observed for compounds considered outside the AD with both the test sets, however, with respect to the number of compounds considered outside the AD with different threshold strategies, the values were comparatively higher with EPI Suite test set. This reflected how diverse both the test sets were in terms of their compounds and indicating that the CAESAR test set comprised of compounds more similar to the training data as compared to the other test set. None of the strategies performed well with Mahalanobis distance measure for CAESAR test set resulting in a negative *ΔRMSEP*. Similar pattern for compounds outside AD was observed for CAESAR model 5 and the corresponding results can be found in Table 3.

Table 3. Statistics for CAESAR Model 5 implementing distance-based approaches with different thresholds. *Maxdist*: Maximum distance between training compounds and centroid of the training set; *d*: Average distance of training compounds from their mean; *ΔRMSEP*: Difference between *RMSEP* for compounds outside and inside the AD.

Approach	*Thresholds*	*Compounds outside the AD*		*Q^2*		*ΔRMSEP*	
		CAESAR out of 95 (%)	*EPI Suite out of 108 (%)*	*CAESAR*	*EPI Suite*	*CAESAR*	*EPI Suite*
Euclidean (*maxdist*)	0.942	0 (0.0)	2 (1.8)	0.774	0.647	-	0.598
Euclidean (*3*d*)	0.958	0 (0.0)	2 (1.8)	0.774	0.647	-	0.598
Euclidean (*2* d*)	0.639	3 (3.1)	9 (8.3)	0.783	0.665	0.329	0.354
Euclidean (*p95*)	0.614	4 (4.2)	11 (10.1)	0.783	0.673	0.266	0.367
Euclidean (*dsz*)	0.393	23 (24.2)	32 (29.6)	0.753	0.646	−0.128	0.044
CityBlock (*maxdist*)	1.472	0 (0.0)	2 (1.8)	0.774	0.647	-	0.598
CityBlock (*3*d*)	1.791	0 (0.0)	1 (0.9)	0.774	0.634	-	0.037
CityBlock (*2*d*)	1.194	1 (1.0)	5 (4.6)	0.772	0.657	−0.417	0.457
CityBlock (*p95*)	1.085	4 (4.2)	11 (10.1)	0.767	0.665	0.309	0.308
CityBlock (*dsz*)	0.723	21 (22.1)	32 (29.6)	0.751	0.639	−0.156	0.022
Mahalanobis (*maxdist*)	6.957	0 (0.0)	0 (0.0)	0.774	0.633	-	-
Mahalanobis (*3*d*)	6.121	0 (0.0)	0 (0.0)	0.774	0.633	-	-
Mahalanobis (*2*d*)	4.081	3 (3.1)	6 (5.5)	0.767	0.621	−0.445	−0.275
Mahalanobis (*p95*)	3.859	5 (5.2)	6 (5.5)	0.764	0.621	−0.327	−0.275
Mahalanobis (*dsz*)	2.495	23 (24.2)	18 (16.6)	0.760	0.637	−0.081	0.035

The study was further extended by implementing the above mentioned threshold strategies for each distance measure, but considering average distance of each training compound from its first 5 nearest neighbors. Given a *n* by *n* distance matrix where *n* is total number of training compounds, in all the cases, average distance of each training sample from its first five nearest training neighbors is found. Later, the gross average is derived from these average distance values which will be denoted henceforth as *D*. In the first and second case, twice and thrice the value of *D* is considered as threshold, respectively. For the third case, percentile approach discussed earlier in potential density distribution methods, is applied on the sorted average distances of all training compounds (used to calculate *D*) and the value corresponding to 95 percentile (*p95*) is considered as threshold [27]. For the last strategy (*DSZ*), besides calculating the gross average distance *D* from the first five nearest neighbors, also the standard deviation (*Std*) is calculated on the average distances. Finally, the threshold is defined as $D + Std \times z$, where *z* is the arbitrary parameter and is set to 0.5 as default value [26]. For all the cases, average distance of a test compound from its first five nearest neighbors in the training set is compared with the defined threshold. If the average distance for this test compound is less than or equal to the threshold value, it is considered inside the AD.

Results derived with all the four threshold strategies are shown in Tables 4 and 5 for CAESAR Model 2 and Model 5, respectively, considering different distance measures.

Table 4. Statistics for CAESAR Model 2 implementing different 5NN based threshold strategies. For the acronyms *D*, *p95*, *DSZ*, and *ΔRMSEP*, refer to text.

Approach	*Thresholds*	*Compounds outside the AD*		*Q^2*		*ΔRMSEP*	
		CAESAR out of 95(%)	*EPI Suite out of 108(%)*	*CAESAR*	*EPI Suite*	*CAESAR*	*EPI Suite*
Euclidean (*3*D*)	1.522	2 (2.1)	1 (0.9)	0.804	0.676	0.394	2.713
Euclidean (*2* D*)	1.015	9 (9.5)	16 (14.8)	0.795	0.750	−0.037	0.765
Euclidean (*p95*)	1.164	8 (8.4)	13 (12.0)	0.797	0.745	0.859	1.342
Euclidean (*DSZ*)	0.693	14 (14.7)	31 (28.7)	0.787	0.767	−0.113	0.517
CityBlock (*3*D*)	2.371	4 (4.2)	5 (4.6)	0.803	0.679	0.187	0.968
CityBlock (*2*D*)	1.581	10 (10.5)	18 (16.7)	0.794	0.742	−0.042	0.664
CityBlock (*p95*)	1.918	7 (7.4)	11 (10.2)	0.799	0.741	0.034	0.944
CityBlock (*DSZ*)	1.083	16 (16.8)	27 (25.0)	0.801	0.731	0.037	0.446
Mahalanobis (*3*D*)	1.718	3 (3.2)	4 (3.7)	0.803	0.628	0.221	0.295
Mahalanobis (*2*D*)	1.145	9 (9.5)	18 (16.7)	0.794	0.748	−0.045	0.691
Mahalanobis (*p95*)	1.388	6 (6.3)	11 (10.2)	0.801	0.735	0.908	1.183
Mahalanobis (*DSZ*)	0.786	19 (20.0)	29 (26.9)	0.795	0.745	−0.019	0.470

Table 5. Statistics for CAESAR Model 5 implementing different 5NN based threshold strategies. D: The gross average distance of training set compounds from their 5NN; *ΔRMSEP*: Difference between *RMSEP* for compounds outside and inside the AD.

Approach	*Thresholds*	*Compounds outside the AD*		*Q^2*		*ΔRMSEP*	
		CAESAR out of 95 (%)	*EPI Suite out of 108 (%)*	*CAESAR*	*EPI Suite*	*CAESAR*	*EPI Suite*
Euclidean (*3*D*)	1.681	0 (0.0)	2 (2.8)	0.774	0.644	-	0.364
Euclidean (*2* D*)	1.121	7 (7.4)	13 (12.0)	0.781	0.690	0.130	0.437
Euclidean (*p95*)	1.331	1 (1.0)	7 (6.5)	0.772	0.656	−0.331	0.126
Euclidean (*DSZ*)	0.782	18 (18.9)	22 (20.4)	0.784	0.743	0.072	0.512
CityBlock (*3*D*)	2.684	1 (1.1)	5 (4.6)	0.772	0.648	−0.456	0.307
CityBlock (*2*D*)	1.789	9 (9.5)	12 (11.1)	0.788	0.690	0.190	0.462
CityBlock (*p95*)	2.302	2 (2.1)	8 (7.4)	0.785	0.657	0.529	0.310
CityBlock (*DSZ*)	1.232	19 (20.0)	30 (27.8)	0.782	0.753	0.055	0.433
Mahalanobis (*3*D*)	2.006	0 (0.0)	4 (3.7)	0.774	0.624	−0.326	−0.149
Mahalanobis (*2*D*)	1.337	6 (6.3)	10 (9.3)	0.779	0.683	0.115	0.482
Mahalanobis (*p95*)	1.668	2 (2.1)	6 (5.6)	0.771	0.631	−0.193	−0.043
Mahalanobis (*DSZ*)	0.933	21 (22.1)	24 (22.2)	0.792	0.713	0.110	0.356

As obvious from Table 4, lowest number of test compounds were considered outside AD with the strategy considering 3**D* as threshold. When the thresholds were lowered to 2**D*, several other test compounds were considered outside the AD, however, the model performed worse with CAESAR test set. Same pattern was observed considering EPI Suite test set however, without lowering the model statistics and the number of test compounds outside the AD were comparatively higher in this case. Strategy taking into account also the standard deviation, was associated with the lowest threshold value thus, restricting the AD. Large number of compounds were considered outside the AD without improving the model statistics. The percentile approach considered reasonable number of test compounds outside AD without any major impact on the model statistics and the results were comparatively better with EPI Suite test set. Similar results and considerations were derived with CAESAR model 5.

The next and the final step was to finalize upon one threshold strategy for distance-based approaches. All the four above mentioned strategies behaved differently depending on the distance measure considered. A strategy that improved the model statistics for one distance measure couldn't have similar impact for another distance measure. This observation couldn't allow an easy interpretation towards finalizing upon one strategy. However, considering improved model statistics with reasonable number of test compounds considered outside the AD, the percentile approach was a preferred choice. Moreover, when the methodologies for different AD methods were described earlier, Probability Density Distribution method reflected the statistical significance of defining percentiles. These considerations concluded finalizing upon the percentile approach for overall comparison of the results. This approach was implemented initially considering the distance of training compounds from their centroid (*p95*) and in the later case, based on average distance of training compounds from their 5 nearest neighbors (*p95*). Both the considerations were different in defining the interpolation space and thus, resulted in different number of compounds outside the AD with the same distance measure. Information derived in both the cases was significant and thus was retained for the overall comparison of the results.

3.2. Overall Comparisons

The distance-based approaches were then compared with other previously discussed AD approaches, considering the both CAESAR (95 compounds) and EPI suite (108 compounds) test sets. The results are summarized in Tables 6 and 7 for CAESAR Model 2 and Model 5, respectively.

As shown in Table 6, by performing PCA analysis along with Bounding Box approach on Model 2, two test compounds were considered outside the AD. Convex Hull and Probability Density approach led to maximum number of test compounds outside the AD, thus decreasing the generalization ability of the models. *p95* approach lowered the model statistics for Mahalanobis distance measure. Q^2 slightly lowered for Convex Hull that considered several test compounds outside the AD. On the other hand, model statistics improved for Probability Density Distribution approach which was associated with the maximum number of test compounds outside the AD (42.6%). As a general remark, the model statistics improved for several approaches with increase in number of test compounds considered outside the AD. Since the CAESAR test set comprised compounds more similar to the training set, not many test compounds emerged outside the AD; however, the EPI suite test set is comparatively different from the training data and thus considerably more compounds were outside the AD by different approaches. *ΔRMSEP* remained positive considering most of the AD approaches. Similar pattern for compounds outside the AD was derived for CAESAR model 5 and the corresponding results are reported in Table 7.

Table 6. Statistics for CAESAR Model 2 applied to CAESAR and EPI Suite test sets for different AD approaches.

Approach	***Compounds outside the AD***		***Q^2***		***ΔRMSEP***	
	CAESAR out of 95 (%)	***EPI Suite out of 108 (%)***	***CAESAR***	***EPI Suite***	***CAESAR***	***EPI Suite***
Euclidean Dist. *(p95)*	7 (7.4)	12 (11.1)	0.802	0.718	0.146	0.753
City Block Dist. *(p95)*	8 (8.4)	11 (10.1)	0.801	0.705	0.068	0.717
Mahalanobis Dist. *(p95)*	6 (6.3)	5 (4.6)	0.791	0.624	−0.174	0.162
5NN-Euclidean Dist. (*p95*)	8 (8.4)	13 (12.0)	0.797	0.745	0.859	1.342
5NN-CityBlock Dist. (*p95*)	7 (7.4)	11 (10.2)	0.799	0.741	0.034	0.944
5NN-Mahalanobis Dist. (*p95*)	6 (6.3)	11 (10.2)	0.801	0.735	0.908	1.183
Bounding Box	0 (0.0)	2 (1.8)	0.797	0.678	-	1.798
PCA Bounding Box	2 (2.1)	3 (2.8)	0.804	0.688	0.371	1.533
Convex Hull	22 (23.2)	31 (28.7)	0.789	0.721	−0.052	0.368
Potential Function	29 (30.5)	46 (42.6)	0.831	0.766	0.156	0.374

Table 7. Statistics for CAESAR Model 5 applied to CAESAR and EPI Suite test sets for different AD approaches.

Approach	***Compounds outside the AD***		***Q^2***		***ΔRMSEP***	
	CAESAR out of 95 (%)	***EPI Suite out of 108 (%)***	***CAESAR***	***EPI Suite***	***CAESAR***	***EPI Suite***
Euclidean Dist. *(p95)*	4 (4.2)	11 (10.1)	0.783	0.673	0.266	0.367
City Block Dist. *(p95)*	4 (4.2)	11 (10.1)	0.767	0.665	0.309	0.308
Mahalanobis Dist. *(p95)*	5 (5.2)	6 (5.5)	0.764	0.621	−0.327	−0.275
5NN-Euclidean Dist. (*p95*)	1 (1.0)	7 (6.5)	0.772	0.656	−0.331	0.126
5NN-CityBlock Dist. (*p95*)	2 (2.1)	8 (7.4)	0.785	0.657	0.529	0.310
5NN-Mahalanobis Dist. (*p95*)	2 (2.1)	6 (5.6)	0.771	0.631	−0.193	−0.043
Bounding Box	0 (0.0)	1 (0.9)	0.774	0.634	-	0.037
PCA Bounding Box	0 (0.0)	2 (1.8)	0.774	0.634	-	0.021
Convex Hull	16 (16.8)	21 (19.4)	0.780	0.643	0.049	0.051
Potential Function	28 (29.5)	47 (43.5)	0.787	0.813	0.062	0.455

To visualize where test set compounds were located with respect to the training compounds, multidimensional scaling (MDS) was performed. This enabled the representation of 5 dimensional data (the molecular descriptors defining the CAESAR models) by means of a two dimensional plot.

From the MDS plots in Figure 1, it is clear that several test compounds that were localized towards the extremities of training set were considered outside the AD with most of the approaches. For example, CAESAR test compound 33 and EPI Suite test compound 60 were considered outside on the basis of 7 and 9 AD approaches, respectively. However, there were several compounds that were quite close to the training space but still falling outside the AD, especially with Convex Hull and Probability Density approaches (for example, CAESAR test compound 38 and EPI Suite test compound 33). Since the internal empty regions within chemical space cannot be easily detected and correlation between descriptors cannot be explained with Bounding Box, this approach failed to consider any test compound outside the AD. When the same approach was implemented on this dataset after PCA analysis, the correlation between descriptors was taken into account and as a result, two compounds from the test set were considered outside the AD. With respect to the EPI Suite test set, the MDS plots showed how most of test compounds outside the AD were lying in the training set extremities and were almost the same for different AD approaches. Those compounds were further more distant from training set than in the CAESAR test set. Similar results were derived for CAESAR model 5 and the corresponding plots are shown in Figure 2.

Figure 1. CAESAR test set (**a**) and Epi Suite test set (**b**) projected in the training space of Model 2. Training set (+); test set (♦); compounds outside the AD with different approaches; distance based *p95* (○), distance based 5NN (○), Bound. Box and PCA Bound. Box (○), Conv. Hull (○), Pot. Funct. (○).

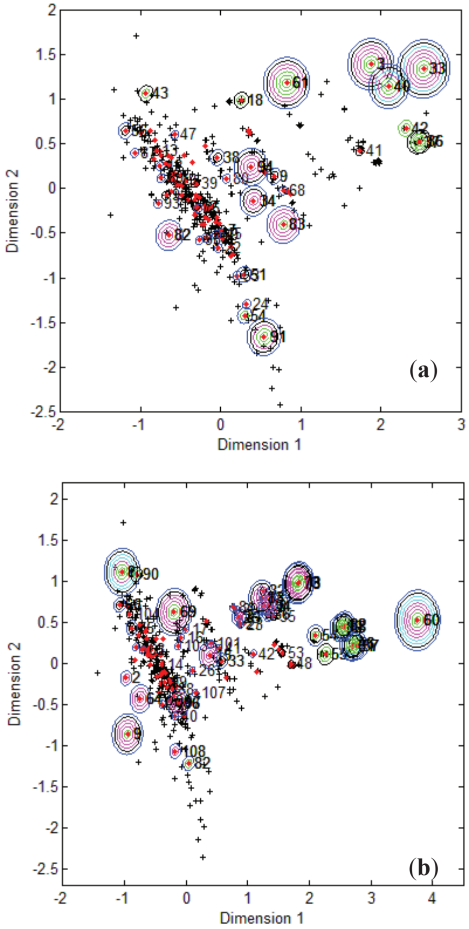
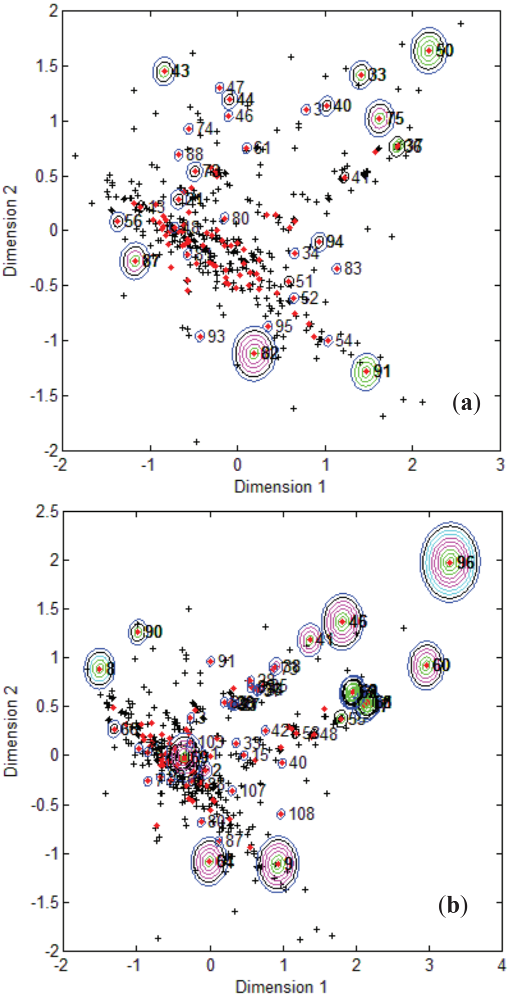


Figure 2. CAESAR test set (**a**) and Epi Suite test set (**b**) projected in the training space of Model 5. Training set (+); test set (♦); compounds outside the AD with different approaches; distance based *p95* (○), distance based 5NN (○), Bound. Box and PCA Bound. Box (○), Conv. Hull (○), Pot. Funct. (○).



It was observed for both the CAESAR models that some compounds very close to the training compounds were considered outside the AD while others lying further were considered inside it. This could be explained by the fact that most of the implemented approaches considered only interpolation by simply excluding all test compounds in the extremities and including all those surrounded by training set compounds even if they are situated within empty regions of the chemical space.

Figure 3 provides the calculated logBCF values from the CAESAR Model 2 plotted against the experimental log BCF values (Exp logBCF). It can be noted that several test compounds not so reliably predicted were considered outside the AD. On the other hand, well predicted test compounds like 34 in

CAESAR test set and 59 in EPI Suite test set were considered outside by 2 and 5 AD approaches respectively. This indicates that the strategy used by different AD approaches might have considered some well predicted compounds outside the AD, thus affecting the model statistics. As seen earlier in Tables 6 and 7, Convex Hull and Probability Density Distribution approaches had considerable number of test compounds outside the AD; however, both the approaches differed significantly with respect to the model statistics. The results corresponding to CAESAR model 5 are plotted in Figure 4.

Figure 3. Predicted Vs observed log BCF values for CAESAR test set (**a**) and Epi Suite test set (**b**) with Model 2. Training set (+); test set (♦); compounds outside the AD with different approaches; distance based *p95* (○), distance based 5NN (○), Bound. Box and PCA Bound. Box (○), Conv. Hull (○), Pot. Funct. (○).

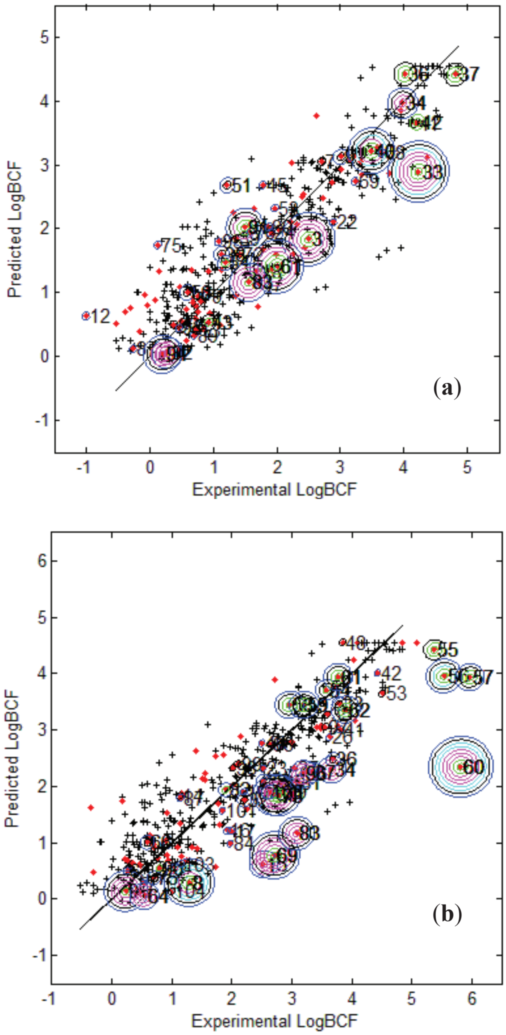
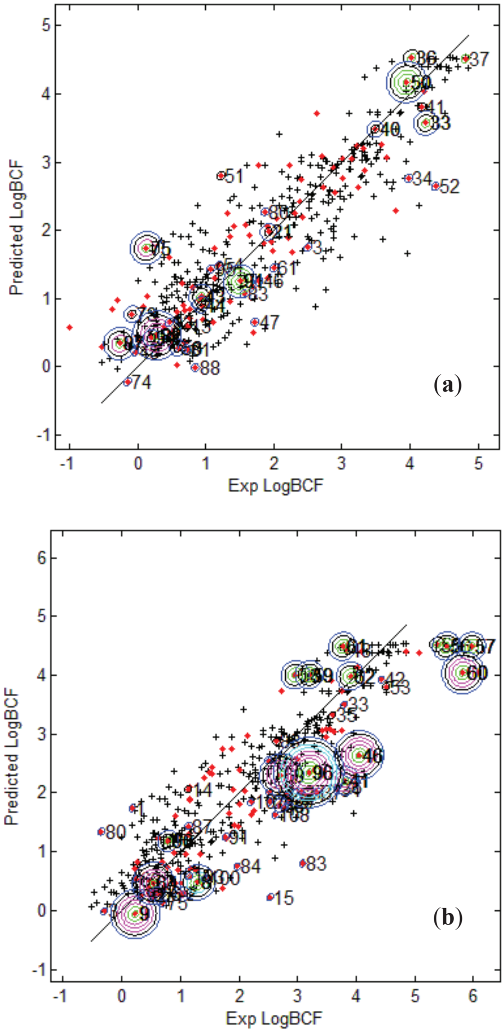


Figure 4. Predicted Vs observed log BCF values for CAESAR test set (**a**) and Epi Suite test set (**b**) with Model 5. Training set (+); test set (♦); compounds outside the AD with different approaches; distance based *p95* (○), distance based 5NN (○), Bound. Box and PCA Bound. Box (○), Conv. Hull (○), Pot. Funct. (○).



The plots indicate that several test compounds unreliably predicted were localized on the extremities of the training space and considered outside the AD while several well predicted test compounds were also considered outside with different approaches. This observation holds true for both the test sets however, the number of test compounds considered outside the AD were considerably higher for EPI Suite test set. Figure 3b shows that the three compounds 56, 57 and 60 considered outside the AD by several approaches were underestimated, and thus the model statistics highly improved with AD approaches not considering them within the domain of applicability.

4. Conclusions

The characterization of interpolation space varied depending on the Applicability Domain approach implemented. Approaches compared in this study suffered from several limitations, some concerning the complexity of algorithm while some related to the algorithm used for defining interpolation space. Addition of PCA did not contribute significantly to the Bounding Box approach with the first test set however, with respect to the second validation set, performing PCA analysis had a significant impact on improving the model statistics. Probability Density Distribution approach and Convex Hull were associated with the highest number of test compounds outside the AD and thus allowing only a limited use of the models. Distance-based approaches considered reasonable number of test compounds outside the AD, however model statistics lowered for some distance measures. As expected, most of the test compounds considered outside the AD with most of the approaches were concentrated towards the training set extremities. It was clearly evident from the MDS plots that the distance from training space was significant in defining the model's AD. Also, several test compounds badly predicted by the model were considered as outside the AD with most of the approaches. The results from the alternative test set provided were similar; however, number of test compounds outside the AD increased. When various thresholds were subjected to distance-based approaches, it was noted, however with some exceptions, that increase in the number of test compounds outside AD also improved the model's statistics. Finally, all the implemented AD approaches had their own strengths and limitations and thus, it is up to the model builder to choose most appropriate applicability domain approach for his model. For instance, in this study, one of the aspects considered to evaluate a given AD approach was the number of test compounds outside the AD and its resulting impact on the model performance. It is important to note that the results derived with different AD approaches may vary for the same dataset and none of these approaches can be considered sufficient enough to be applied to all the cases; therefore, considering the present state of the art, it would be preferable to evaluate the results from all possible strategies before assessing a new compound set.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme [FP7/2007–2013] under Grant Agreement n. 238701 of Marie Curie ITN Environmental Chemoinformatics (ECO) project. We would like to thank Emilio Benfenati at Mario Negri Institute, Milan, Italy for providing us with the CAESAR data.

References and Notes

1. Netzeva, T.I.; Worth, A.; Aldenberg, T.; Benigni, R.; Cronin, M.T.D.; Gramatica, P.; Jaworska, J.S.; Kahn, S.; Klopman, G.; Marchant, C.A.; *et al.* Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Altern. Lab. Anim.* **2005**, *33*, 155–173.
2. Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicabilty domain estimation by projection of the training set descriptor space: A review. *Altern. Lab. Anim.* **2005**, *33*, 445–459.

3. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O.A. Stepwise approach for defining the applicability domain of SAR and QSAR models. *J. Chem. Inf. Model.* **2005**, *45*, 839–849.
4. REACH. European Community Regulation on chemicals and their safe use. Available online: http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm (accessed on 3 February 2012).
5. Worth, A.P.; Bassan, A.; Gallegos, A.; Netzeva, T.I.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Vracko, M. *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*. ECB Report EUR 21866 EN, European Commission, Joint Research Centre; Ispra, Italy, 2005; p. 95.
6. OECD. Quantitative Structure-Activity Relationships Project [(Q)SARs]. Available online: http://www.oecd.org/document/23/0,3746,en_2649_34377_33957015_1_1_1_1,00.html (accessed on 3 February 2012).
7. Worth, A.P.; van Leeuwen, C.J.; Hartung, T. The prospects for using (Q)SARs in a changing political environment: high expectations and a key role for the Commission's Joint Research Centre. *SAR QSAR Environ. Res.* **2004**, *15*, 331–343.
8. Nikolova-Jeliazkova, N.; Jaworska, J. An approach to determining applicability domains for QSAR group contribution models: an analysis of SRC KOWWIN. *Altern. Lab. Anim.* **2005**, *33*, 461–470.
9. Sheridan, R.; Feuston, R.P.; Maiorov, V.N.; Kearsley, S. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 1912–1928.
10. Zhao, C.; Boriani, E.; Chana, A.; Roncaglioni, A.; Benfenati, E. A new hybrid QSAR model for predicting bioconcentration factor (BCF). *Chemosphere* **2008**, *73*, 1701–1707.
11. Lombardo, A.; Roncaglioni, A.; Boriani, E.; Milan, C.; Benfenati, E. Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. *Chem. Cent. J.* **2010**, *4* (Suppl 1), doi:10.1186/1752-153X-4-S1-S1.
12. Meylan, W.M.; Howard, P.H.; Aronson, D.; Printup, H.; Gouchie, S. *Improved Method for Estimating Bioconcentration Factor (BCF) from Octanol-Water Partition Coefficient*, 2nd Update; SRC TR-97-006; Syracuse Research Corp., Environmental Science Center: North Syracuse, NY, USA, 1997; Prepared for: Robert S. Boethling, EPA-OPPT.
13. Meylan, W.M.; Howard, P.H.; Boethling, R.S.; Aronson, D.; Printup, H.; Gouchie, S. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. *Environ. Toxicol. Chem.* **1999**, *18*, 664–672.
14. MATLAB. The Language of Technical Computing. Available online: http://www.mathworks.com/products/matlab/ (accessed on 3 February 2012).
15. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemometr. Intell. Lab.* **1987**, *2*, 37–52.
16. Preparata, F.P.; Shamos, M.I. Convex Hulls: Basic Algorithms. In *Computational Geometry: An Introduction*; Preparata, F.P., Shamos, M.I., Eds.; Springer-Verlag: New York, NY, USA, 1991; pp. 95–148.

17. Tropsha, A.; Gramatica, P.; Gombar, V. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69–77.
18. Jouan-Rimbaud, D.; Bouveresse, E.; Massart, D.L.; de Noord, O.E. Detection of prediction outliers and inliers in multivariate calibration. *Anal. Chim. Acta* **1999**, *388*, 283–301.
19. Forina, M.; Armanino, C.; Leardi, R.; Drava, G. A class-modelling technique based on potential functions. *J. Chemometr.* **1991**, *5*, 435–453.
20. Tong, W.; Hong, H.; Fang, H.; Xie, Q. Perkins, R. Decision forest: Combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525–531.
21. Tong, W.; Hong, H.; Xie, Q.; Xie, L.; Fang, H.; Perkins, R. Assessing QSAR limitations: A regulatory perspective. *Curr. Comput. Aid. Drug Des.* **2004**, *1*, 65–72.
22. Wan, C.; Harrington, P.B. Self-configuring radial basis function neural networks for chemical pattern recognition. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1049–1056.
23. DRAGON (Software for Molecular Descriptor Calculations). Talete srl, Milano, Italy. Available online: http://www.talete.mi.it (accessed on 3 February 2012).
24. Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, *49*, 1669–1678.
25. Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemometr.* **2010**, *24*, 104–201.
26. Tetko, I.V.; Sushko, I.; Pandey, A.K.; Zhu, H.; Tropsha, A.; Papa, E.; Oberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **2008**, *48*, 1733–1746.
27. Weaver, S.; Gleeson, M.P. The importance of the domain of applicability in QSAR modeling. *J. Mol. Graph. Model.* **2008**, *26*, 1315–1326.

Sample Availability: The CAESAR data sets used in this study can be requested directly from the CAESAR project (http://www.caesar-project.eu).