



Final report of the research activity

Project Dates

Start Date: 29 December, 2010

End Date: 28 September, 2013

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

Early Stage researcher:

Kamel Mansouri

Supervisor:

Prof. Roberto Todeschini

Research Institution:


Milano Chemometrics and QSAR Research Group

Department of Environmental Sciences

University of Milano Bicocca. Milan, Italy

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.

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.

Contents

Contents.....	iv
List of Figures	vii
List of Tables	viii
Preface.....	ix
<i>The ECO project</i>	ix
<i>Thesis goals and structure</i>	ix
Part I: Introduction.....	1
1. POPs and pathways to the environment.....	2
1.1. General properties of POPs.....	3
1.2. Pathways to the environment	4
2. Regulation of chemicals in Europe	5
2.1. REACH, the European legislation about chemicals	5
2.2. The European Chemicals Agency (ECHA).....	6
2.3. Mode of action within REACH	6
3. QSARS for regulatory purposes	7
3.1. QSARs and REACH.....	7
3.2. OECD Principles for the Validation of QSARs.....	7
Part II: Tools and Methods.....	9
1. Introduction	10
2. Data acquisition and curing.....	11
2.1. Data sources.....	11
2.2. Data curing.....	12
3. Molecular descriptors	14

3.1.	Introduction	14
3.2.	Analysis of new molecular descriptors.....	15
3.3.	Software for descriptor calculation.....	20
4.	Variable selection techniques.....	22
4.1.	Stepwise forward selection.....	22
4.2.	Genetic Algorithms (GAs).....	23
5.	Modeling methods in QSAR	24
5.1.	Unsupervised methods for exploratory data analysis	24
5.2.	Supervised learning methods for modeling.....	25
6.	Goodness of fit measures and validation methods.....	29
6.1.	Validation methods	29
6.2.	Regression parameters	30
6.3.	Classification parameters.....	31
7.	Applicability domain of models	32
8.	Multi-criteria decision making in model selection.....	34
Part III: Results and Discussion		36
1.	Introduction.....	37
2.	Octanol/Water Partition Coefficient	39
2.1.	Case study 1: the logP-1000 contest.....	40
2.2.	Case study 2: modeling PHYSPROP dataset for logP	43
3.	Bioaccumulation	49
3.1.	Definitions	49
3.2.	Assessing bioaccumulation by QSARs	50
3.3.	Case study: QSARs for assessing bioaccumulation.....	51
4.	Biodegradability	53
4.1.	QSARs for assessing biodegradability of chemicals.	53
4.2.	Summary of the published study on biodegradability	54
4.3.	Substructural keys for predicting biodegradability.....	54
4.4.	Predicting biodegradability from the BOD values.....	55
5.	Applicability domain of QSARs.....	57
5.1.	Different approaches for defining the AD.....	57
5.2.	Summary of the published study on the AD approaches	58
6.	Structure-activity landscapes	59
6.1.	The Structure-Activity Landscape Index (SALI)	59

Contents

6.2.	Graphical methods for characterizing SAR landscapes.....	60
6.3.	Metric distances for investigating SAR landscapes.....	61
7.	Conclusion	65
	References	67

List of Figures

Figure 1: <i>different levels of structural representation.</i>	14
Figure 2: <i>Choosing the hyperspace with the optimal margin</i>	28
Figure 3: <i>The overlapping conditions for the adequacy of QSARs in regulatory purposes.</i>	32
Figure 4: <i>The correlation between logP and the molecular weights.</i>	41
Figure 5: <i>The KNIME workflow used to prepare the dataset.</i>	43
Figure 6: <i>The frequency of descriptors' selection during 20 runs (a) and the obtained models (b) and their parameters Q2 (red points) and U scores (blue points).</i>	46
Figure 7: <i>The evolution of Q2 (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.</i>	47
Figure 8: <i>Predicted versus observed BOD values of the training set (black points) and test set (red points).</i>	55
Figure 9: <i>The activity cliffs of the simulated dataset using SALI. The x and y axis represent the number of samples while the z-axis represents the difference in activity.</i>	60
Figure 10: <i>SAS map applied on the simulated dataset using the Euclidean distance.</i>	60
Figure 11: <i>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. ...</i>	62
Figure 12: <i>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.</i>	63
Figure 13: <i>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. ...</i>	63

List of Tables

Table 1: The confusion matrix in classification	31
Table 2: REACH regulatory endpoints associated with the OECD test guidelines.	37
Table 3: QSPR models for logP using different modeling methods.	41
Table 4: Statistics of MlogP and AlogP for the training and test sets.	42
Table 5: Benchmarking the predictions of the selected models.	42
Table 6: Ranks and weights of the considered parameters.	44
Table 7: The 10 dCV performed during the first GA run of the third step.	45
Table 8: Nine models obtained by means of stepwise forward selection performed after the 10 dCVs of the first GA run.	45
Table 9: Evaluation of models resulting from the stepwise forward selection.	47
Table 10: The molecular descriptors included in the model M16.	47
Table 11: Statistics of model M16.	48
Table 12: The selected k NN models using different combinations of structural keys and distance measures.	54
Table 13: Statistics of weighted and non weighted k NN regression models.	55
Table 14: Statistics of weighted and non weighted k NN classification models.	56

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 ($\log K_{OW}$, $\log P_{OW}$ 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

3. QSARS for regulatory purposes

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.

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 [45]. In addition, REACH requires a set of 4 conditions in alignment with the OECD principles to be fulfilled for QSAR modeling [46]:

- 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 [47], PubChem [48,49], ChemExper [50]. These databases have useful searching options, such as chemical name, CAS-RN (Chemical Abstract Registration Number) [51,52], PubMed ID [53] and/or structure representations such as SMILES and INCHI codes [54].

In addition to the information about chemicals, other online sources provide also access to modeling tools designed for QSAR, such as VCCLAB [55], OCHEM (Online Chemical Modeling Environment) [56], OpenTox [57], QSARdb [58], SPARC [59] and PBT profiler [60], 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 [61]. Another relevant data source for QSAR is the online freely available database of the United States Environmental Protection Agency (US-EPA) [62]. The datasets used to build the physicochemical and environmental fate models implemented in EPI (Estimation Program Interface) Suite are available online [63]. 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) [54], or unique SMILES [64].

However, several errors in the SMILES notations can be faced during the structures checks [65,66]. 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 [45].

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) [67]. 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 [68]. 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) [69]. 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) [70].

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 [71]. OCHEM also offers several API services for uploading data as well as creating and applying QSAR models [72]. 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 [73,74].

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 [75].

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 [76]. 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.

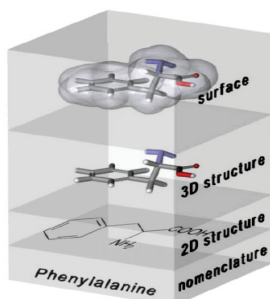


Figure 1: *different levels of structural representation.*

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 [77,78]. 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*” [76].

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 IUPAC International chemical Identifier (InChI) is also used as a descriptor to predict properties of chemicals [79].
- 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 [80].
- 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 [81].
- 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) [82–84].

A comprehensive review of molecular descriptors has been published by Todeschini and Consonni [76].

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 [85]. 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 [76,86,87]. 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 [86,88–91]. 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) [92–94]. It is also called the Laplacian graph energy when it's calculated from the Laplacian matrix [95,96]. $SpPos$ 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 [97]. $SpMaxA$ is the maximum absolute value of the spectrum [76].

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 [76]. These descriptors are calculated as following:

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

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 [98–101].

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 [102].

Two new families of spectral indices have been recently developed and published in the literature [85]. 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 [103].

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 [98]. 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 [104]. In addition to QSAR analysis, these descriptors can also be useful in similarity/dissimilarity studies of chemicals [98].

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 [76]. All the calculations were performed by the software DRAGON [105].

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 [106,107]. 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 [108].

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 [109]. 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 [76]. 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**:

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

where δ_i 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 **B**(*w*) are augmented adjacency matrices defined to account for heteroatoms and bond multiplicity calculated as the following:

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

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 [110]:

$$[\mathbf{Dz}(w)]_{ij} = \begin{cases} d_{ij}(w, \pi^*) & \text{if } i \neq j \\ 1 - \frac{w_C}{w_i} & \text{if } i = j \end{cases} \quad d_{ij}(w, \pi^*) = \sum_{b=1}^{d_{ij}} \left(\frac{1}{\pi_b^*} \cdot \frac{w_C^2}{w_{b(1)} \cdot w_{b(2)}} \right)$$

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 (χ) are two examples of the topological indices that can be derived from the previously described matrices ($\mathbf{B}(w)$ and $\mathbf{Dz}(w)$) [111,112]. Variances of these indices calculated using the mass (m) and electronegativity (e) as weighting schemes have shown interesting modeling properties [102].

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 $\mathbf{M}(w)$ [76,113].

The general formula for calculating the hyper-Wiener-type index is the following [111]:

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

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 [114]. It was the first descriptor proposed for molecular branching [115].

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 [116–118]. 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 [76].

The Balaban-like index general formula is given by [112]:

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

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 [119,120]. Recently, these bit-strings started to be used as descriptors for QSAR modeling [121–124]. 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 [125,126]. 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 [127]. 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 [54].

Following the general classification pattern for molecular descriptors, these string representations of chemical structures are categorized in 2D, 3D and 4D accordingly [123,124,128–131].

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 [121].

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 [132]. 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 [133].

The PubChem binary substructure keys are developed to be used by PubChem database in order to perform the searching queries [49]. 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 [54]. These sections encode hierarchic atom-type counts, rings, atom pairs, atom nearest neighbours, atom connections, simple and complex SMARTS patterns [134].

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 [105]. It was the main tool of molecular descriptors calculations used in this work. It calculates almost 5000 molecular descriptors [135]. 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 [136]. 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 [137]. 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 [132]. The substructure searching method is based on the complete atom-atom and bond-bond matching [138,139]. The developers of the software have also provided a list of 1365 substructures covering a wide range of fragments based on mass-spectrometry fragmentation [140]. 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 [141]. It is available under the terms of the GNU Lesser General Public License (LGPL) [142]. Thus it is freely available for use and modification by academic and industrial institutions and may be integrated in proprietary packages [143]. Subsequently, its libraries started to be a basis for several software projects [141]. 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 [144].

A dedicated graphical user interface was designed for the molecular descriptor calculations [145]. 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 [146]. 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 [54,141]. 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 [147]. 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 [148–150]. 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 [151].

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) [152–154], stepwise forward/backward selection [155], particle swarms [156], simulated annealing and ant colony algorithms [157,158]. 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 [155]. 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 [159]. GAs are widely used in the fields of Chemometrics and Chemoinformatics [154,160,161].

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 [152,153].

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 [152]. 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 [153,154,162].

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 [163,164]. 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:

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

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 ($k\text{NN}$), Support Vector Machines (SVM) and Partial Least Squares Discriminant Analysis (PLSDA) [165–173].

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 [174–176]. 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 [165,166]. 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

k NN 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 k NN modeling is the Euclidean distance. Other metrics such as Manhattan distance and Mahalanobis distance can also be applied [177]. 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 k NN approach often presents good results, however, its predictive ability in regression can be altered in the case of high-dimensional data [178].

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:

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

where y_i is the observed response, $x_{i1}, x_{i2}, \dots, x_{ip}$ are the independent variables for the i^{th} 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 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 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 $b_0, b_1, b_2, \dots, b_p$ 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 $\hat{\mathbf{y}}$ is obtained as following:

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

where \mathbf{b} is the vector of estimated parameters $b_0, b_1, b_2, \dots, b_p$ 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 [168]. A major advantage of this method is its ability to overcome the problem of singularity of $(\mathbf{X'X})$ 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 \mathbf{X} into orthogonal scores \mathbf{T} and loadings \mathbf{P} as follows;

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

Then, \mathbf{y} is correlated to the first columns of the scores instead of the original variables of \mathbf{X} . In this way, PLS includes information from both, \mathbf{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:

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}' + \mathbf{F}$$

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 [179,180].

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 k NN 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, k NN performs better in classification than in regression for with such high dimensionality [181].

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 [173,182]. 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 [183–185]. 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 [186,187]. This library was implemented in MATLAB to be coupled with the GAs for the variable selection and modeling steps.

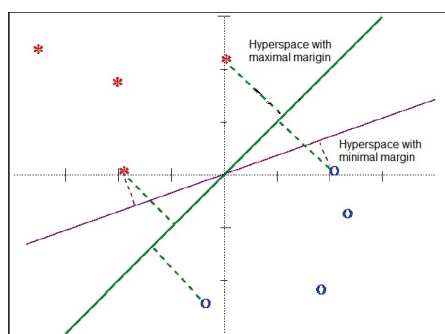


Figure 2: *Choosing the hyperspace with the optimal margin*

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” [188].

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 [45,189,190]. 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 [191–194].

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 [195]. This problem can be solved by applying the more robust LMO approach [196]. 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 [76].

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 [197,198]. 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 [76,199]. 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 [46].

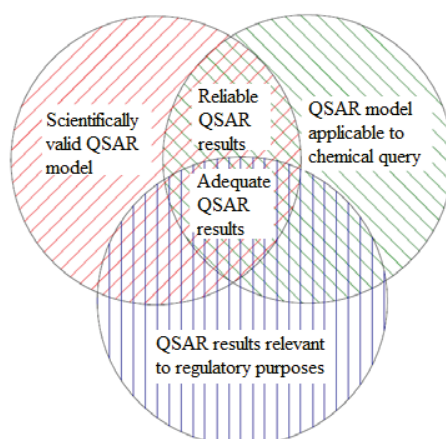


Figure 3: *The overlapping conditions for the adequacy of QSARs in regulatory purposes.*

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 [200–202]. 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 k NN 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 [203].

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 \mathbf{X} are obtained from the Hat matrix \mathbf{H} calculated as follows:

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

The diagonal values of \mathbf{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 [200,201]. 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 [204–207]. 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 U_i of each i th alternative for the non-weighted and weighted cases are given by:

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

where p is the number of criteria t .

The Desirability D_i of each i th alternative for the non-weighted and weighted cases are given by:

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

The weight constraint is:

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

The weights are calculated using the method of normalized weights for ranked criteria [208,209]:

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

where 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 [210]. 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 i th alternative for the j th criterion into a value between 0 and 1 [209].

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 $\tau = 1$ 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^2Test .

*(d) Rank chromosomes according to $\{D, U\} = f(Q^2, LVs, nVars, nOutliers, R^2Test)$.
→ 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.

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 [44]. 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 [44].

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 [39,44].

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

1. Introduction

	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 [\[44\]](#).

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 [211]. The behavior and fate of a chemical substance are mostly depending on its physicochemical properties [212]. In absence of reliable experimental data, non-testing methods such as QSPR estimations can be used to provide such required information about chemicals [211].

The octanol/water partition coefficient (k_{ow}), usually expressed in log values ($\log k_{ow}$ or $\log P$) is a key parameter in environmental assessment of chemicals since it is related to lipophilicity/hydrophobicity [213–217]. It is used as the basic predictor in many estimation models for water solubility, bioavailability, bioaccumulation, toxicity/ecotoxicity and PBT assessment/screening [213,218–222]. In REACH regulation, providing a $\log P$ value is required for all tonnage bands of chemicals [39,211].

$\log P$ is defined as the ratio of the concentrations of a dissolved chemical in two immiscible phases, octanol and water, at the equilibrium [223]. 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 $\log P$ can be considered in regulatory purposes as more reliable than a single test [44].

Several QSPR models using different methods have been developed and published in the literature [219,224–228]. These models and their results have been compared in several reviews [229,230].

A comparison study of different methods for predicting $\log P$ was published by Mannhold and Dross [230]. 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.* [229].

There are two OECD test protocols for $\log P$, OECD Guideline 107 and OECD Guideline 117 [44]. 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 $\log D$, and could differ from its neutral form by a factor of 4 to 5 orders of magnitude [231].

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 [232–234]. 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 [235].

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, Chempidspider 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) [105]. 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 *k*NN for regression. The number of Latent Variables (LVs) for PLS and the number of nearest neighbors for *k*NN 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.

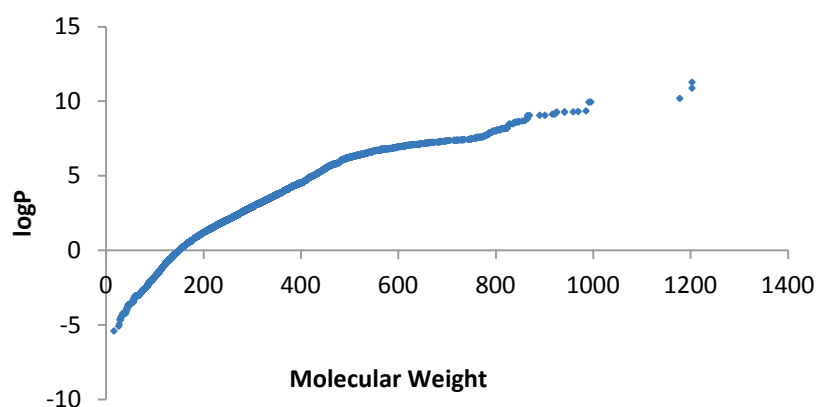


Figure 4: The correlation between logP and the molecular weights.

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
kNN_1	255	5	-	0.86	0.88	-	0.74	0.68
kNN_2	30	5	-	0.84	0.85	-	0.80	0.75
kNN_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) [236,237]. 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	R^2	RMSEC	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 *k*NN_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		<i>k</i> 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 [63,238]. This dataset was originated from the PHYSPROP database [239,240]. The same dataset was used for the development of KOWWIN, the EpiSuite's model for estimating logP [227].

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 [136]. 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 [69]. The developed workflow (Figure 5) was used to run a series of queries through the web-services of the online databases ChemSpider and CIR [73,241].

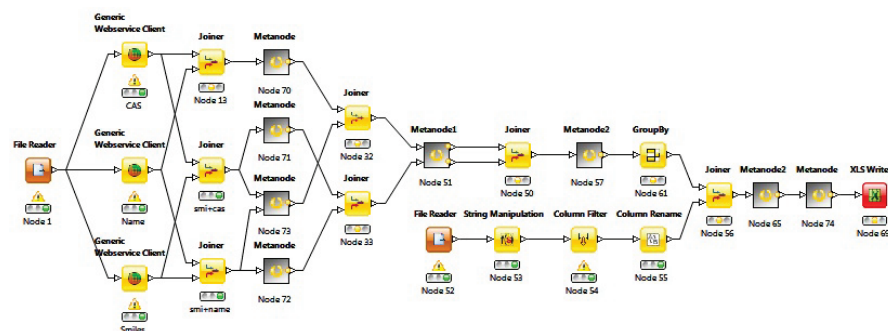


Figure 5: The KNIME workflow used to prepare the dataset.

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 [105]. 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.

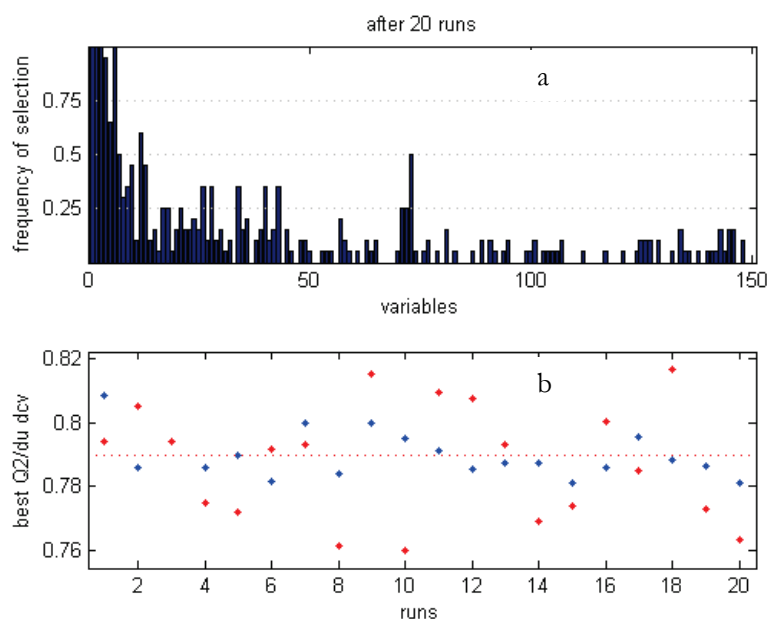


Figure 6: The frequency of descriptors' selection during 20 runs (a) and the obtained models (b) and their parameters Q^2 (red points) and U scores (blue points).

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.

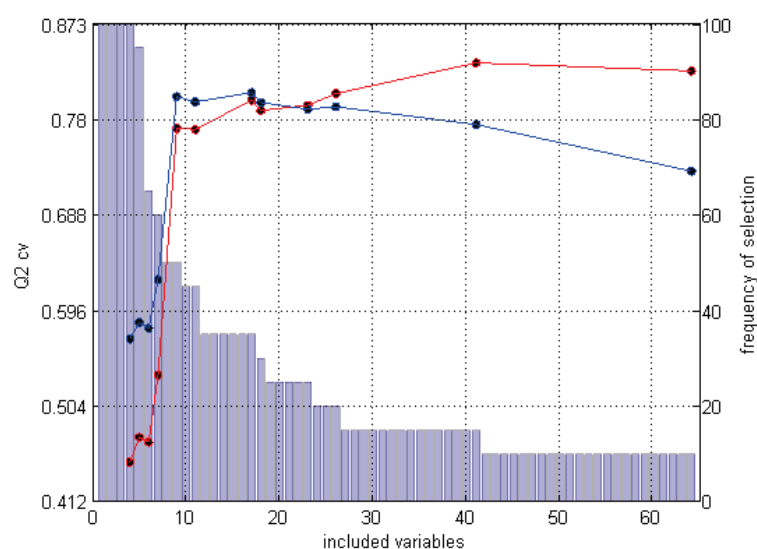


Figure 7: The evolution of Q^2 (red line) and U (blue line) during the final Stepwise forward selection. The histogram represents the frequency of selection of the descriptors in percentage over the number of total runs.

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 led 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 [242]. 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 [243,244].

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: $BCF = C_o/C_w$

where BCF is the bioconcentration factor (L/kg), C_o is the chemical concentration in the whole organism (mg/kg, wet weight) and 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: $BMF = C_o/C_d$

where BMF is dimensionless, C_o is the steady-state chemical concentration in the organism (mg/kg), 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 [242].

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 [244–246].

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 [247].

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 [242].

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 [248–254]. Some of these models have already been used in regulatory applications of a number of chemicals [242].

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 [255]. 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 [247,255].

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 [252,256]. 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 [215].

However, the logP based predictions for high hydrophobic compounds remain uncertain for regulatory use [242].

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 [257–260]. As for the previous experimental descriptor, BCF models based on S may have accuracy problems for specific chemical groups [260].

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 [261–265].

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 [44].

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 [266,267]. 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 [267].

PBDEs are known for their long range atmospheric transport, in fact, they are usually detected in different geographical regions distant from their original sources [268]. 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 [267,269].

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 (IUPAC) nomenclature. Similar toxic properties have also been noticed between CBs and PBDEs [270–272]. However, the second group of chemicals are more lipophilic than their corresponding chlorinated compounds [273].

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 [103]. 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.

3. Bioaccumulation

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 publication Mansouri et al.[\[103\]](#).

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 [274]. 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 [274]. 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 [274].

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 [275]. 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 [276].

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 CO₂ production. It provides also an attempt to the metabolic pathways and the plausible biodegradation products that may arise [277].

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 [278].

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 [279,280]. 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 [102].

More details can be found in the published article of the study provided in the publication Mansouri et al. [102].

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 Mansouri et al. [102].

This QSAR study used only binary descriptors based on several structural keys calculated by PADEL and SubMat (Table 12). For this purpose, a *k*NN 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 ($1 - \text{Similarity}$) [281]. 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 [102].

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

Structural keys (number)	Distance	<i>k</i>	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 k NN 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 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 RBs while the correctly predicted NRBs 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.

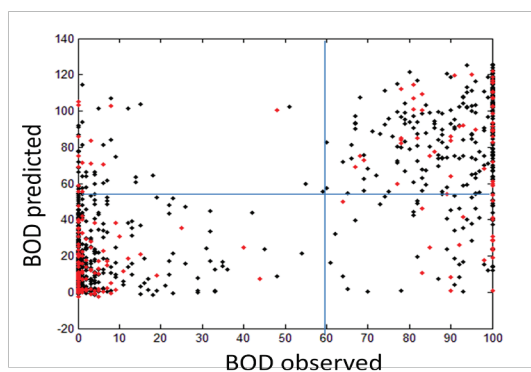


Figure 8: Predicted versus observed BOD values of the training set (black points) and test set (red points).

4. Biodegradability

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 [200–202]. 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 [200].

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 [200].

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 [\[200\]](#).

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 [\[203\]](#).

The complete study is published and the article is provided in the Sahigara et al. [\[203\]](#).

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*” [282].

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) [282]. Later several different studies using the SALI index and graphical methods for characterizing the activity landscapes have been published [283–291].

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 [282]. 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 i th and the j th 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 (\mathbf{X}) and a simulated activity (\mathbf{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.

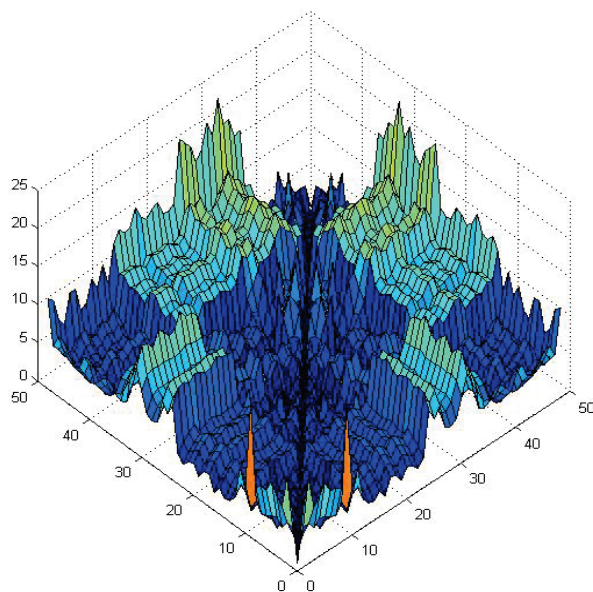


Figure 9: The activity cliffs of the simulated dataset using SALI. The x and y axis represent the number of samples while the z -axis represents the difference in activity.

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 [292–294]. 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.

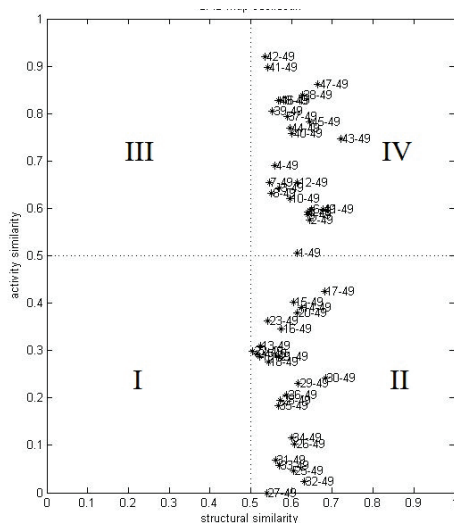


Figure 10: SAS map applied on the simulated dataset using the Euclidean distance.

6.2.2. The Patterson plot

The Patterson plot is a method to graphically investigate the structurally similar compounds and their relative activity similarity [295]. 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 $1 - \textit{Similarity}$.

If the dataset obey to the congenericity principle, the pairs of molecules will appear in the lower triangle of the plot [296]. 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 $1 - \textit{Similarity}$ 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:

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

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:

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

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

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

where p is the number of variables.

For binary data, the Soergel distance is the complement of the Jaccard-Tanimoto [297,298]. 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.

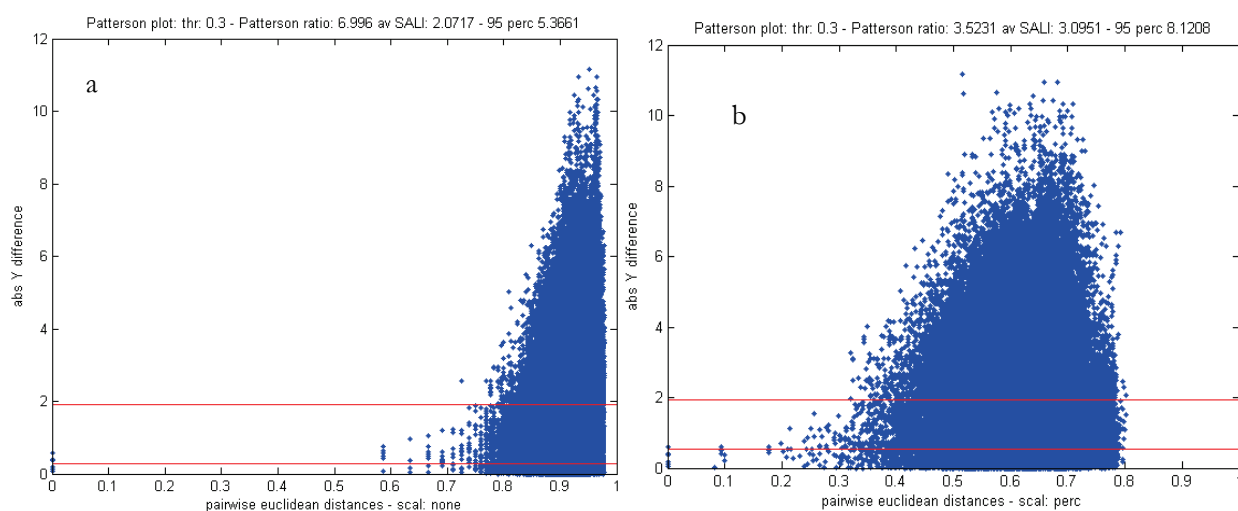


Figure 11: The pairwise Euclidean distance without scaling (a) and scaled (b). thr: the used threshold for calculating the Patterson ratio; av SALI: the average value of the SALI index on all pairs; 95 perc: the 95 percentile of the SALI index on all pairs.

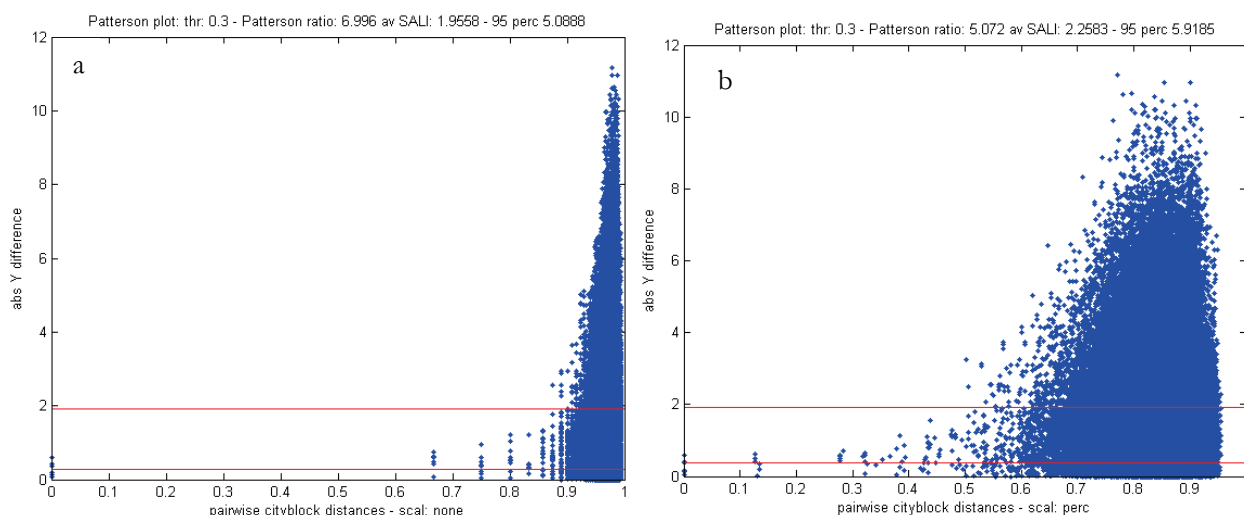


Figure 12: The pairwise Manhattan distance without scaling (a) and scaled (b). *thr*: the used threshold for calculating the Patterson ratio; *av SALI*: the average value of the SALI index on all pairs; *95 perc*: the 95 percentile of the SALI index on all pairs.

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.

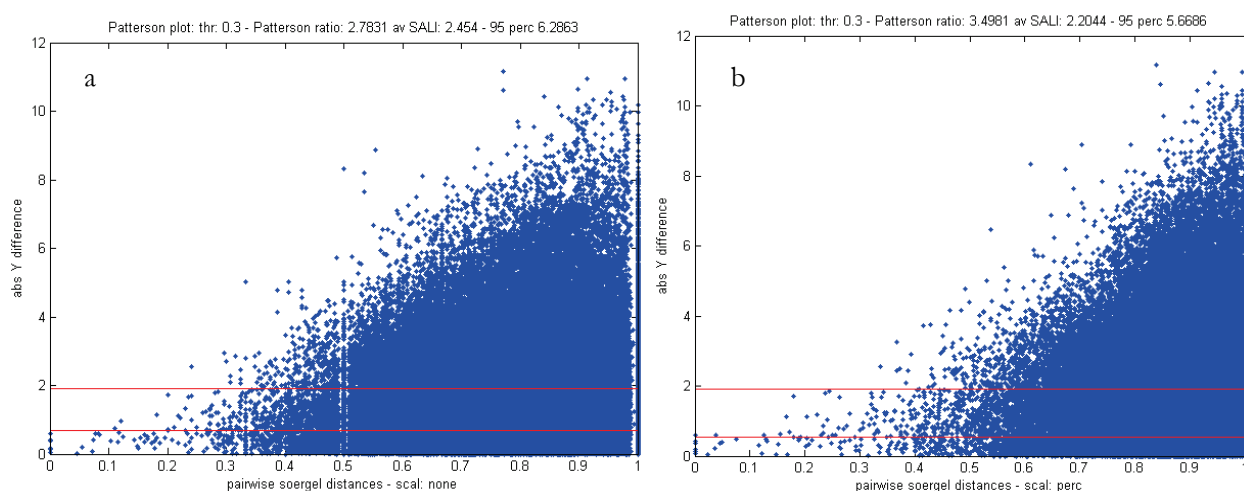


Figure 13: The pairwise Soergel distance on non-scaled (a) and scaled data (b). *thr*: the used threshold for calculating the Patterson ratio; *av SALI*: the average value of the SALI index on all pairs; *95 perc*: the 95 percentile of the SALI index on all pairs.

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 [103]. 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, *k*NN 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 [102]. 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 [203].

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. 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.; 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. *N. Engl. J. Med.* **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. Heal. Part 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. *Environ. Sci. Pollut. Res.* **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 Rep. Prep. Contract Chem. Manuf. Assoc.* **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. *Integr. Environ. Assess. Manag.* **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. *J. Phys. Chem. Ref. 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. *Environ. Sci. Technol.* **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. *Integr. Environ. Assess. Manag.* **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. *Environ. Sci. Technol.* **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. *Environ. Pollut.* **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. *Environ. Sci. Pollut. Res.* **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.I.], 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. 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*.
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.
45. 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.
46. ECHA Guidance on Information Requirements and Chemical Safety Assessment. Chapter R6. 2008.
47. ChemSpider | The free chemical database <http://www.chemspider.com/> (accessed Apr 26, 2013).
48. 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.
49. The PubChem Project <http://pubchem.ncbi.nlm.nih.gov/> (accessed Apr 26, 2013).

50. ChemExper - catalog of chemicals suppliers, physical characteristics and search engine <http://www.chemexper.com/> (accessed Apr 26, 2013).
51. 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.
52. CAS, Chemical Abstracts Service <http://www.cas.org/> (accessed Apr 26, 2013).
53. PubMed - NCBI <http://www.ncbi.nlm.nih.gov/pubmed> (accessed Apr 26, 2013).
54. James, C. A.; Weininger, D.; Delany, J. *Daylight Theory Manual*; Chemical Information Systems: Aliso Viejo, CA, USA, 2008.
55. 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.
56. 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.
57. The OpenTox project <http://www.opentox.org/> (accessed Apr 26, 2013).
58. QSAR DataBank <http://qsardb.org/> (accessed Apr 26, 2013).
59. SPARC <http://ibmlc2.chem.uga.edu/sparc/test/login.cfm?CFID=264604&CFTOKEN=52651371> (accessed Apr 26, 2013).
60. PBT Profiler <http://www.pbtprofiler.net/> (accessed Apr 26, 2013).
61. OECD *QSAR Toolbox*; Oasis, 2011.
62. ECOTOX | MED | US EPA http://cfpub.epa.gov/ecotox/data_download.cfm (accessed May 29, 2013).
63. EPI Suite Data <http://esc.syrres.com/interkow/EpiSuiteData.htm> (accessed Apr 26, 2013).
64. D, W.; A, W.; SMILES, W. J. L. Algorithm for Generation of Unique SMILES Notation. *J Chem Inf Comput Sci* **1989**, *29*, 97.
65. 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.
66. 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.
67. *ChemBioFinder*; PerkinElmer Informatics Databases.
68. *ChemBioOffice Ultra 13.0 Suite*; Desktop Software; PerkinElmer Informatics.
69. 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.
70. GNU General Public License - GPL - Free Software Foundation (FSF) <http://www.gnu.org/licenses/gpl.html> (accessed May 1, 2013).
71. ChemSpider API Services <http://www.chemspider.com/AboutServices.aspx> (accessed Apr 28, 2013).
72. OCHEM web-services - user's manual <http://docs.eadmet.com/display/MAN/Using+web-services> (accessed Apr 28, 2013).
73. CIR Chemical Identifier Resolver NCI/CADD <http://cactus.nci.nih.gov/chemical/structure> (accessed May 16, 2013).
74. Talete, S. R. L. *CIR node for KNIME*; Talete srl, <http://www.talete.mi.it>.
75. *Guide to intelligent data analysis: how to intelligently make sense of real data*; Texts in computer science; Springer: London ; New York, 2010.
76. Todeschini, R.; Consonni, V. *Molecular descriptors for chemoinformatics*; Wiley-VCH, 2009.
77. 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.
78. 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.

79. 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.
80. Benigni, R.; Bosa, C. Structural alerts of mutagens and carcinogens. *Curr Comput -Aided Drug* **2006**, *2*, 169–176.
81. 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.
82. Kubinyi, H. Comparative molecular field analysis (CoMFA). In *Handbook of Chemoinformatics*; Wiley-VCH Verlag GmbH, Weinheim, Germany,, 2003; Vol. 4.0, pp. 1555–1575.
83. Kim, K. H. Comparative Molecular Field Analysis (CoMFA). In; Chapman & Hall London, UK, 1995; pp. 291–331.
84. 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.
85. Consonni, V.; Todeschini, R. New spectral indices for molecule description. *Match* **2008**, *60*, 3–14.
86. Ivanciuc, O.; Balaban, A. T. The graph description of chemical structures. *Topol. Indices Relat. Descriptors Qsar Qspr* **1999**, 59–167.
87. Hall, G. G. Eigenvalues of molecular graphs. *Bull Inst Math Appl* **1981**, *17*, 70–72.
88. 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.
89. Ivanciuc, O. Design of topological indices. Part 27. Szeged matrix for vertex- and edgeweighted molecular graphs as a source of structural descriptors for QSAR models. *Rev Roum Chim* **2002**, *47*, 479–492.
90. Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. The complementary distance matrix, a new molecular graph metric. *Act - Models Chem* **2000**, *137*, 57–82.
91. Ivanciuc, O. QSAR and QSPR molecular descriptors computed from the resistance distance and electrical conductance matrices. *Act - Models Chem* **2000**, *137*, 607–631.
92. Gutman, I. Topological formulas for freevalency index. *Croat Chem Acta* **1978**, *51*, 29–33.
93. Gutman, I. The energy of a graph: Old and new results. *Algebr. Comb. Appl.* **2001**, 196–211.
94. Gutman, I. Topology and stability of conjugated hydrocarbons. The dependence of total π -electron energy on molecular topology. *J. Serbian Chem. Soc.* **2005**, *70*, 441–456.
95. Gutman, I.; Zhou, B. Laplacian energy of a graph. *Linear Algebra Its Appl.* **2006**, *414*, 29–37.
96. Zhou, B.; Gutman, I. On Laplacian energy of graphs. *Match* **2007**, *57*, 211–220.
97. Lovasz, L.; Pelikan, J. On the eigenvalue of trees. *Period Math Hung* **1973**, *3*, 175–182.
98. 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.
99. Estrada, E. Spectral moments of the edge adjacency matrix of molecular graphs. 2. Molecules containing heteroatoms and QSAR applications. *J Chem Inf Comput Sci* **1997**, *37*, 320–328.
100. 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.
101. 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.
102. 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.
103. 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.
104. 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.
105. DRAGON (Software for Molecular Descriptor Calculations); Talete srl, <http://www.talete.mi.it>; Milano, Italy, 2012.
106. Bonchv, D. an R. *Chemical Graph Theory: Reactivity and Kinetics*; Gordon and Breach Science Publishers: New York, 1992.
107. Devillers, J. *Topological indices and related descriptors in QSAR and QSPR*; Gordon & Breach: Amsterdam, 1999.
108. Kier, L. B.; Hall, L. H. *Molecular connectivity in structure-activity analysis*; Research Studies Press ; Wiley:

- Letchworth, Hertfordshire, England; New York, 1986.
109. Trinajstić, N. *Chem. Graph Theory* **1992**, 225–273.
 110. Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1412–1422.
 111. Ivanciuc, O.; Ivanciuc, T.; Diudea, M. V. Molecular Graph Matrices and Derived Structural Descriptors. *Sar Qsar Env. Res* **1997**, *7*, 63–87.
 112. 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.
 113. 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.
 114. Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc.* **1947**, *69*, 17–20.
 115. Bonchev, D.; Trinajstic, N. Information theory, distance matrix, and molecular branching. *J Chem Phys* **1977**, *67*, 4517.
 116. Balaban, A. T. Enumeration of cyclic graphs. In *Chemical Applications of Graph Theory*; Academic Press, London, UK, 1976; pp. 63–105.
 117. 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.
 118. Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404.
 119. Hagadone, T. R. Molecular substructure similarity searching: efficient retrieval in twodimensional structure databases. *J Chem Inf Comput Sci* **1992**, *32*, 515–521.
 120. Barnard, J. M. Substructure searching methods: old and new. *J Chem Inf Comput Sci* **1993**, *33*, 532–538.
 121. 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.
 122. 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.
 123. 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.
 124. 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.
 125. 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.
 126. 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.
 127. 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.
 128. 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.
 129. 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.
 130. Pozzan, A. 3D pharmacophoric hashed fingerprints. In *Rational Approaches to Drug Design*, 2001; pp. 224–228.
 131. 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.
 132. MDL Information Systems, Inc.; 14600 Catalina Street, San Leandro, CA 94577, 2004.
 133. 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.
 134. PubChem Substructure Fingerprint ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt (accessed May 1, 2013).
 135. Talete srl. Via V. Pisani, 13 - 20124 Milano - Italy <http://www.talete.mi.it/index.htm> (accessed May 1, 2013).
 136. dProperties (software for molecular property calculation); Talete srl., <http://www.talete.mi.it/>; Milano, Italy, 2012.
 137. Scsibraný, H.; Varmuza, K. *Software SubMat (Generation of Binary Substructure Descriptors)*; Laboratory

- for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology: Vienna, 2004.
138. Structure and Substructure Searching. In *Encyclopedia of computational chemistry*; New York: J. Wiley: Chichester, 1998; pp. 2764–2771.
139. 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.
140. Varmuza, K.; Demuth, W.; Karlovits, M.; Scsibrany, H. Binary substructure descriptors for organic compounds. *Croat. Chem. Acta* **2005**, *78*, 141–149.
141. 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.
142. 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).
143. The Chemistry Development Kit (CDK) SourceForge http://sourceforge.net/apps/mediawiki/cdk/index.php?title=Main_Page (accessed May 1, 2013).
144. 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.
145. CDK Descriptor Calculator GUI <http://rguha.net/code/java/cdkdesc.html> (accessed May 1, 2013).
146. Guha, R. Using R to provide statistical functionality for QSAR modeling in CDK. *Cdk News* **2005**, *2*, 2–6.
147. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466–1474.
148. Laggner, C. *SMARTS Patterns for Functional Group Classification*; Inte:Ligand Software-Entwicklungs und Consulting GmbH, 2005.
149. Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinforma. Oxf. Engl.* **2008**, *24*, 2518–2525.
150. Hall, L. H.; Kier, L. B. Electrotological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem Inf Comput Sci* **1995**, *35*, 1039–1045.
151. PaDEL-Descriptor <http://padel.nus.edu.sg/software/padeldescriptor/> (accessed May 2, 2013).
152. Goldberg, D. E. *Genetic algorithms in search, optimization, and machine learning*; Addison-Wesley: Reading, MA., 1988.
153. Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom* **1992**, *6*, 267–281.
154. Leardi, R. Genetic algorithms in chemometrics and chemistry: A review. *J. Chemom.* **2001**, *15*, 559–569.
155. Jennrich, R. I. Stepwise discriminant analysis. *Stat. Methods Digit. Comput.* **1977**, 76–95.
156. Agrafiotis, D. K.; Cedeño, W. Feature Selection for Structure-Activity Correlation Using Binary Particle Swarms. *J. Med. Chem.* **2002**, *45*, 1098–1107.
157. 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.
158. Izrailev, S.; Agrafiotis, D. K. Variable selection for QSAR by artificial ant colony systems. *Sar Amp Qsar Env. Res* **2001**, *13*, 417–423.
159. Forrest, S. Genetic algorithms: principles of natural selection applied to computation. *Science* **1993**, *261*, 872–878.
160. 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.
161. Lavine, B. K.; Moores, A. J. Genetic Algorithms in Analytical Chemistry. *Anal. Lett.* **1999**, *32*, 433–445.
162. MATLAB Version 7.13.0.564; MathWorks www.mathworks.com, 2011.
163. Hansch, C.; Fujita, T. ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86*, 1616–1626.
164. Free, S. M.; Wilson, J. W. A mathematical contribution to structure-activity studies. *J. Med. Chem.* **1964**, *7*, 395–399.
165. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **1964**, *29*, 1–27.
166. Winsberg, S.; Carroll, J. D. A quasi-nonmetric method for multidimensional scaling VIA an extended euclidean model. *Psychometrika* **1989**, *54*, 217–229.

167. Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; Wiley, 1986.
168. 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.
169. Geladi, P.; Kowalski, B. R. Partial least squares regression: a tutorial. *Anal Chim Acta* **1986**, *185*, 1–17.
170. 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.
171. 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.
172. 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.
173. Louwerse, D. J.; Tate, 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.
174. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
175. Anderson, T. W. Asymptotic theory for principal component analysis. *Ann Math Stat* **1963**, *34*, 122–148.
176. Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
177. Weinberger, K. Q.; Saul, L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *J Mach Learn Res* **2009**, *10*, 207–244.
178. Hastie, T.; Tibshirani, R.; Friedman, J. H. *The elements of statistical learning data mining, inference, and prediction*; Springer: New York, 2009.
179. De Jong, S. SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.
180. Mevik, B. H.; Wehrens, R. The pls package: Principal component and partial least squares regression in R. *J. Stat. Softw.* **2007**, *18*, 1–24.
181. Cover, T.; Hart, P. Nearest neighbor pattern classification. *Ieee Trans. Inf. Theory* **1967**, *13*, 21–27.
182. Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173.
183. Vapnik, V. N. *Statistical Learning Theory*; 1st ed.; Wiley-Interscience, 1998.
184. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; 1st ed.; Cambridge University Press, 2000.
185. Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; 1st ed.; The MIT Press, 2001.
186. Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; National Taiwan University, Department of Computer Science: Taipei 106, Taiwan, 2001.
187. Hsu, C.-W.; Lin, C.-J. A comparison of methods for multiclass support vector machines. *Ieee Trans. Neural Networks* **2002**, *13*, 415–425.
188. 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.
189. 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.
190. Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J Chem Inf Comput Sci* **2003**, *43*, 579–586.
191. Topliss, J. G.; Edwards, R. P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.
192. 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.
193. Clark, M.; Cramer, R. D. The Probability of Chance Correlation Using Partial Least Squares (PLS). *Quant. Struct.-Act. Relationships* **1993**, *12*, 137–145.
194. Aptula, A. O.; Jeliaskova, N. G.; Schultz, T. W.; Cronin, M. T. D. The Better Predictive Model: High q^2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set? *Qsar Comb. Sci.* **2005**, *24*, 385–396.
195. Golbraikh, A.; Tropsha, A. Beware of q^2 ! *J. Mol. Graph. Model.* **2002**, *20*, 269–276.
196. Burden, F. R.; Brereton, R. G.; Walsh, P. T. Cross-validated selection of test and validation sets in multivariate calibration and neural networks as applied to spectroscopy. *Analyst* **1997**, *122*, 1015–1022.
197. 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.

198. Consonni, V.; Ballabio, D.; Todeschini, R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* **2010**, *24*, 194–201.
199. Frank, I. E.; Todeschini, R. The data analysis handbook. *Data Handl. Sci. Technol.* **1994**, *14*, 366.
200. 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.
201. 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.
202. 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.
203. 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.
204. Keller, H. R.; Massart, D. L.; Brans, J. P. Multicriteria decision making: A case study. *Chemom. Intell. Lab. Syst.* **1991**, *11*, 175–189.
205. 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.
206. 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.
207. 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.
208. Pavan, M.; Todeschini, R. Chapter 2 Total-Order Ranking Methods. *Data Handl. Sci. Technol.* **2008**, *27*, 51–72.
209. Pavan, M.; Todeschini, R. Multicriteria Decision Making Methods. In *Comprehensive Chemometrics*; Elsevier, 2009; Vol. 1, pp. 591 – 629.
210. Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated double cross validation. *J. Chemom.* **2009**, *23*, 160–171.
211. 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.
212. 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.
213. Lyman, W. J.; Reehl, W. F.; Rosenblatt, D. H. *Handb. Chem. Prop. Estim. Methods* **1990**.
214. 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.
215. 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.
216. 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.
217. 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*.
218. 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.
219. Hansch, C.; Leo, A. *Substituent Constants for Correlation Analysis in Chemistry and Biology*; John Wiley & Sons, Inc.: New York, 1979.
220. 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.
221. 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.

222. Oszwaldowski, 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.
223. 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.
224. Leo, A. J. Calculating log Poct from structures. *Chem Rev* **1993**, *93*, 1281–1306.
225. Rekker, R. F. *Hydrophobic Fragm. Constant* **1977**.
226. 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.
227. Meylan, W. M.; Howard, P. H. Atom/fragment contribution method for estimating octanol-water partition coefficients. *J. Pharm. Sci.* **1995**, *84*, 83–92.
228. Petrauskas, A.; Kolovanov, E. A. ACD/log P method description. *Persp Drug Disc* **2000**, *19*, 99–116.
229. 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.
230. Mannhold, R.; Dross, K. Calculation Procedures for Molecular Lipophilicity: a Comparative Study. *Quant. Struct.-Act. Relationships* **1996**, *15*, 403–409.
231. 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.
232. 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.
233. 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.
234. ZINC- A database of commercially-available compounds <http://zinc.docking.org/> (accessed Jun 5, 2013).
235. 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.
236. 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.
237. 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.
238. US Environmental Protection Agency <http://www.epa.gov/> (accessed May 16, 2013).
239. 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.
240. 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.
241. ChemSpider Royal Society of Chemistry, Thomas Graham House (290), Science Park, Milton Road, Cambridge CB4 0WF <http://www.chemspider.com/> (accessed Oct 29, 2012).
242. 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.
243. Ecetoc Persistence of chemicals in the environment. *Persistence Chem. Environ.* **2003**.
244. Boethling, R. S.; Mackay, D. *Handb. Prop. Estim. Methods Chem. Environ. Heal. Sci.* **2000**.
245. Dearden, J. QSAR Modeling of Bioaccumulation. In *Predicting Chemical Toxicity and Fate*; CRC Press, 2004.
246. 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.
247. Nendza, M. *Structure Activity Relationships in Environmental Sciences*; Springer, 1998.
248. Schüürmann, G.; Klein, W. Advances in bioconcentration prediction. *Chemosphere* **1988**, *17*, 1551–1574.

249. 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.
250. 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.
251. Mackay, D. Correlation of Bioconcentration Factors. *Es T Contents* **1982**, *16*, 274–278.
252. 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.
253. 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.
254. 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.
255. 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.
256. 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.
257. 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.
258. 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.
259. Jorgensen, S. E.; Halling-Sorensen, B.; Mahler, H. *Handb. Estim. Methods Ecotoxicol. Environ. Chem.* **1998**.
260. Isnard, P.; Lambert, S. Estimating bioconcentration factors from octanol-water partition coefficient and aqueous solubility. *Chemosphere* **1988**, *17*, 21–34.
261. Sabljic, A. The prediction of fish bioconcentration factors of organic pollutants from the molecular connectivity model. *Z. Gesamte Hyg.* **1987**, *33*, 493–496.
262. 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.
263. 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.
264. 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.
265. 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.
266. OSPAR Commission. *Certain Brominated Flame Retardants-Polybrominated diphenyl ethers, polybrominated biphenyls, hexabromo cyclododecane*; OSPAR Priority Substances; OSPAR Commission: London, 2001.
267. OSPAR Commission. *Tetrabromobisphenol-A-Update*; OSPAR Priority Substances; OSPAR Commission: London, 2005.
268. Mikula, P.; Svobodová, Z. Brominated flame retardants in the environment: Their sources and effects (a review). *Acta Vet. Brno* **2006**, *75*, 587–599.
269. UNEP Recommendations of the Persistent Organic Pollutants Review Committee of the Stockholm Convention to amend Annexes A, B or C of the Convention 2009.
270. De Wit, C. A. An overview of brominated flame retardants in the environment. *Chemosphere* **2002**, *46*, 583–624.
271. McDonald, T. A. A perspective on the potential health risks of PBDEs. *Chemosphere* **2002**, *46*, 745–755.
272. 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.
273. 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.
274. 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.
275. Pavan, M.; Worth, A. P. Review of estimation models for biodegradation. *Qsar Comb. Sci.* **2008**, *27*, 32–40.
276. EPISuite v. 4.0. **2010**.
277. L. CATALOGIC; Laboratory of Mathematical Chemistry: Burgas, Bulgaria.
278. Accelrys *TOPKAT toxicity suite*; Accelrys.
279. 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).
280. Multicase Ecotoxicity <http://www.multicase.com/products/prod098.htm> (accessed May 22, 2013).
281. 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.
282. Maggiora, G. M. On Outliers and Activity Cliffs Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.
283. 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.
284. Guha, R.; Van Drie, J. H. Structure-activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.
285. Hu, Y.; Bajorath, J. Molecular scaffolds with high propensity to form multi-target activity cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 500–510.
286. 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.
287. Namasivayam, V.; Bajorath, J. Searching for Coordinated Activity Cliffs Using Particle Swarm Optimization. *J. Chem. Inf. Model.* **2012**, *52*, 927–934.
288. 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.
289. 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.
290. 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.
291. 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.
292. 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.
293. 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.
294. 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.
295. 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.
296. 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.
297. Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
298. 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.