Simona Kovarich

QSAR models for the (eco-)toxicological characterization and prioritization of emerging pollutants: case studies and potential applications within REACH

PhD in Chemical Sciences (XXV Cycle)

Supervisors: Dr. Ester Papa, PhD

Prof. Paola Gramatica

UNIVERSITY OF INSUBRIA

Varese-Como, 2013

SUMMARY

Under the European REACH regulation (Registration, Evaluation, Authorisation and Restriction of Chemical substances - (EC) No 1907/2006), there is an urgent need to acquire a large amount of information necessary to assess and manage the potential risk of thousands of industrial chemicals. Meanwhile, REACH aims at reducing animal testing by promoting the intelligent and integrated use of alternative methods, such as *in vitro* testing and *in silico* techniques. Among these methods, models based on quantitative structure-activity relationships (QSAR) are useful tools to fill data gaps and to support the hazard and risk assessment of chemicals.

The present thesis was performed in the context of the CADASTER Project (CAse studies on the Development and Application of *in-Silico* Techniques for Environmental hazard and Risk assessment), which aims to integrate *in-silico* models (e.g. QSARs) in risk assessment procedures, by showing how to increase the use of non-testing information for regulatory decision-making under REACH. The aim of this thesis was the development of QSAR/QSPR models for the characterization of the (eco-)toxicological profile and environmental behaviour of chemical substances of emerging concern. The attention was focused on four classes of compounds studied within the CADASTER project, i.e. brominated flame retardants (BFRs), fragrances, prefluorinated compounds (PFCs) and (benzo)-triazoles (B-TAZs), for which limited amount of experimental data is currently available, especially for the basic endpoints required in regulation for the hazard and risk assessment.

Through several case-studies, the present thesis showed how QSAR models can be applied for the optimization of experimental testing as well as to provide useful information for the safety assessment of chemicals and support decision-making.

In the first case-study, simple multiple linear regression (MLR) and classification models were developed *ad hoc* for BFRs and PFCs to predict specific endpoints related to endocrine disrupting (ED) potential (e.g. dioxin-like activity, estrogenic and androgenic receptor binding, interference with thyroxin transport and estradiol metabolism). The analysis of modelling molecular descriptors allowed to highlight some structural features and important structural alerts responsible for increasing specific ED activities. The developed models were applied to screen over 200 BFRs and 33 PFCs without experimental data, and to prioritize the most hazardous chemicals (on the basis of ED potency profile), which have been then suggested to other CADASTER partners in order to focus the experimental testing.

In the second case-study, MLR models have been developed, specifically for B-TAZs, for the prediction of three key endpoints required in regulation to assess aquatic toxicity, i.e. acute toxicity in algae (EC_{50} 72h *Pseudokirchneriella subcapitata*), daphnids (EC_{50} 48h *Daphnia magna*) and fish (LC_{50} 96h *Onchorynchus mykiss*). Also in this case, the developed QSARs were applied for screening purposes. Among over 350 B-TAZs lacking experimental data, 20 compounds, which were predicted

as toxic ($EC(LC)_{50}$ ≤ 10 mg/L) or very toxic ($EC(LC)_{50}$ ≤ 1 mg/L) to the three aquatic species, were prioritized for further experimental testing.

Finally, in the third case-study, classification QSPR models were developed for the prediction of ready biodegradability of fragrance materials. Ready biodegradation is among the basic endpoints required for the assessment of environmental persistence of chemicals. When compared with some existing models commonly used for predicting biodegradation, the here proposed QSPRs showed higher classification accuracy toward fragrance materials. This comparison highlighted the importance of using local models when dealing with specific classes of chemicals.

All the proposed QSARs have been developed on the basis of the OECD principles for QSAR acceptability for regulatory purposes, paying particular attention to the external validation procedure and to the statistical definition of the applicability domain of the models. QSAR models based on molecular descriptors generated by both commercial (DRAGON) and freely-available (PaDEL-Descriptor, QSPR-Thesaurus) software have been proposed. The use of free tool allows for a wider applicability of the here proposed QSAR models.

Concluding, the QSAR models developed within this thesis are useful tools to support hazard and risk assessment of specific classes of emerging pollutants, and show how non-testing information can be used for regulatory decisions, thus minimizing costs, time and saving animal lives.

Beyond their use for regulatory purposes, the here proposed QSARs can find application in the rational design of new safer compounds that are potentially less hazardous for human health and environment.

PUBLICATIONS

This thesis is based on the following publications, which are referred to with Roman numerals I-VII in the text. Some unpublished results are also cited.

I. Papa E., Kovarich S., Gramatica P. (2010) QSAR modeling and prediction of the endocrine disrupting potencies of brominated flame retardants. *Chem. Res. Toxicol*. 23, 946-954.

II. Kovarich S., Papa E., Gramatica P. (2011) QSAR classification models for the prediction of endocrine disrupting activity of brominated flame retardants. *J. Hazard. Mater*. 190, 106-112.

III. Kovarich S., Papa E., Li J., Gramatica P. (2012) QSAR classification models for the screening of the Endocrine Disrupting activity of perfluorinated compounds. *SAR and QSAR in Environmental Research* (proceedings of CMTPI 2011). 23, 207-220.

IV. Papa E., Kovarich S., Gramatica P. QSAR prediction of the competitive interaction of emerging halogenated pollutants with human transthyretin. *SAR and QSAR in Environmental Research.* Accepted August 2012 (in press).

V. Gramatica P., Cassani S., Roy P.P., Kovarich S., Yap C.W., Papa E. (2012) QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae. *Molecular Informatics* 31, 817-835.

VI. Cassani S., Kovarich S., Papa E., Roy P.P., van der Wal L., Gramatica P. Daphnia and fish toxicity of (benzo)triazoles: validated QSAR models, and interspecies quantitative activity-activity modelling. Submitted to *Journal of Hazardous Material* (October 2012).

VII. Cassani S., Kovarich S., Papa E., Roy P.P., Rahmberg M., Nilsson S., Sahlin U., Jeliazkova N., Kochev N., Pukalov O., Tetko I., Brandmaier S., Durjava M., Kolar B., Peijnenburg W., Gramatica P. Evaluation of CADASTER QSAR models for aquatic toxicity of (benzo-)triazoles and prioritization by consensus. Submitted to *Alternatives to Laboratory Animals - ATLA* (October 2012).

Additional papers published (or submitted for publication) within this thesis are following:

- Papa E., Kovarich S., Gramatica P. (2009) Development, Validation and Inspection of the Applicability Domain of QSPR Models for physico-chemical properties of Polybrominated Diphenyl Ethers. *QSAR and Combinatorial Science* 28, 790-796.
- Bhhatarai B., Teez W., Liu T., Öberg T., Jeliazkova N., Kochev N., Pukalov O., Tetko I.V., Kovarich S., Papa E., Gramatica P. (2011) CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Molecular Informatics* 30, 189-203.
- Papa E., Kovarich S., Gramatica P. (2011) On the use of local and global QSARs for the prediction of Physico-Chemical Properties of Polybrominated Diphenyl Ethers. *Molecular Informatics* 30, 232–240.
- Roy P.P., Kovarich S., Gramatica P. (2011) QSAR model reproducibility and applicability: a case study of rate constants of hydroxy radical reaction models applied to Polybrominated Diphenyl Ethers and (Benzo-)Triazoles. *Journal of Computational Chemistry* 32, 2386-2396.
- Iqbal M. S., Golsteijn L., Öberg T., Sahlin U., Papa E., Kovarich S., Huijbregts M.A.J. The influence of uncertainty in quantitative structure-property relationships on persistence and long-range transport potential: the case of polybrominated diphenyl ethers (PBDEs). *Environmental Toxicology and Chemistry* Accepted January 2013 (in press).

- Golsteijn L., Iqbal M.S., Cassani S., Hendriks H.W.M., Kovarich S., Papa E., Rorije E., Sahlin U., and Huijbregts M.A.J. The Role of Uncertain Substance Property Predictions in the Chemical Hazard Ranking of Triazoles. Submitted to *Environmental Science & Technology* (November 2012).
- Durjava M.K., Kolar B., Arnus L., Papa E., Kovarich S., Sahlin U., Peijnenburg W. Experimental assessment of the environmental fate and effects of triazoles and benzotriazoles. Submitted to *Alternatives to Laboratory Animals - ATLA* (October 2012).

CONTENTS

APPENDICES

Chapter 1

Introduction

1.1. Background

1.1.1. Emerging pollutants

Defining the ever-expanding universe of chemicals that surrounds, sustains and constitutes our life is extremely complex. Industrial and technological development of the modern society has inevitably increased the amount of chemicals produced and released into the environment. As a consequence of this, living systems have been continuously exposed to multitudes of chemicals with potential serious adverse effects over the years. Great efforts have been made by the scientific community and regulators in order to understand, prevent, control and mitigate the environmental pollution by chemicals, but this is still an open issue.

Until now, over 70 million organic and inorganic substances have been indexed in the Chemical Abstract Service register (C.A.S.) by the American Chemical Society[1], and this number increases on a daily basis. In contrast, less than 300,000 substances have been inventoried or regulated by government bodies worldwide[2], and for these chemicals, very limited information is currently available on intrinsic properties, environmental behaviour and health effects.

Since the 1970s, the attention of authorities and scientific research has mainly been focused on "conventional" pollutants, such as pesticides, HPVCs (High Production Volume Chemicals) and POPs (Persistent Organic Pollutants), and many of them are already regulated or inserted in priority lists (e.g., Stockholm Convention on Persistent Organic Pollutants[3]). However these chemicals represent just a small piece of the universe of existing chemicals that have the potential to cause adverse effects to the exposed organisms. In the last decades, concerns raised over new groups of substances, referred to as "emerging pollutants", whose presence in environment and/or adverse effects had not been detected before, or simply were not investigated since they were assumed to be innocuous. The term "emerging" refers to i) chemicals which have been newly introduced on the market and therefore in the environment, ii) substances that have been present in the environment for a long time but whose presence and effects are currently being elucidated, iii) chemicals naturally present in the environment but whose adverse effects to human health and the environment were not known in the past (e.g. natural hormones, phytoestrogens and algal toxins), and iv) a new concern raised for an old pollutant related to new aspects of their occurrence, fate and effects (e.g. the production of acrylamide during the cooking of many starchy foods) (Daughton, 2005). Examples of emerging pollutants are pharmaceuticals

[1] http://www.cas.org/content/chemical-substances (Accessed 12 December 2012)
[2] http://www.cas.org/content/regulated-chemicals (Accessed 12 December 2012)
[3] http://chm.pops.int/Convention/tabid/54/Default.aspx (Accessed 12 December 2012)

and personal care products (referred to as PPCPs), industrial and household chemicals (e.g. brominated flame retardants and perfluorinated chemicals), and nanomaterials. These chemicals, assumed to be innocuous and widely used to improve our life styles, have been gradually and continuously introduced into the environment at low doses leading to a chronic exposure of living systems.

Emerging pollutants may enter the environment in several ways, depending on their use pattern. Compounds like industrial by-products can be directly released in water bodies and air through the industrial discharges (either authorized or abusive). Industrial processes of synthesis or molecule's destruction, such as incineration, are also responsible for the accidental formation of dangerous by-products, e.g. polychloro- and polybromodibenzo dioxins and furans (PCDDs, PBDDs, PCDFs and PBDFs). Household products, pharmaceuticals and personal care products enter the environment indirectly, and continuously, through domestic discharges; in fact, many of these chemicals are not efficiently degraded in wastewater treatment plants (WWTPs) and may reach receiving water bodies and groundwater aquifers. The same may apply to drugs and diagnostic agents derived from treated/untreated hospital waste. Veterinary drugs, which are often added to animal feed, can be released to the environment with animal waste through land application (e.g., soil amendment/fertilization) or leakage from storage structures (Kolpin *et al*., 2002).

Emerging pollutants are detected increasingly throughout the environment as well as in every level of trophic chains (e.g. fish, amphibians, terrestrial animals and man). This is currently possible because of the improvement of the analytical techniques, which allow to sensibly reduce the detection limits (low-nanogram and even picogram per liter) and which are now capable of detecting compounds (previously not possible) at very low concentrations (Richardson, 2007).

A monitoring study by Kolpin and co-workers (Kolpin *et al.*, 2002), in which water samples were collected from a network of 139 U.S. streams, demonstrated the presence of a wide variety of organic wastewater contaminants (OWCs), including pharmaceuticals, hormones, steroids, insect repellents, detergent metabolites, disinfectants, plasticizers, flame retardants, antibiotics and fragrances (Figure 1.1).

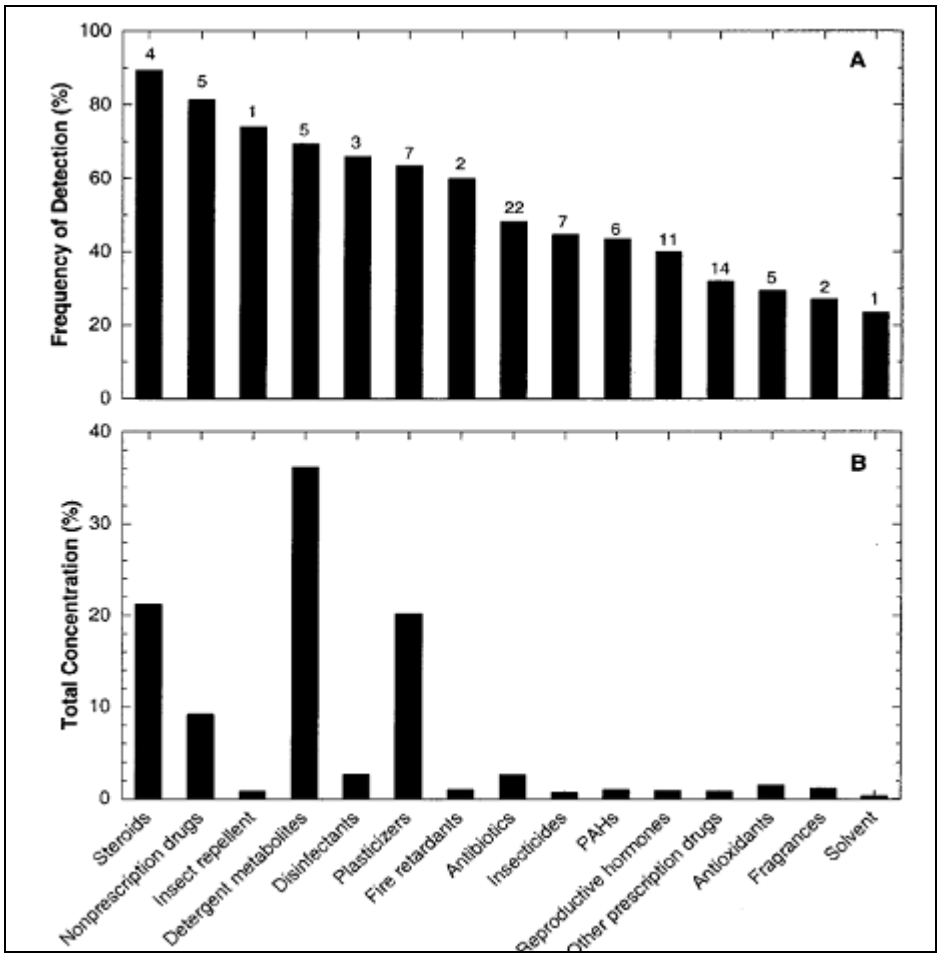


Figure 1.1. Frequency of detection (A) and percentage of total measured concentration (B) of OWCs by general use category. Number of compounds in each category is shown above the bars (Figure from Kolpin et al., 2002).

This study by Kolpin is just an example of the large amount of literature reporting the ubiquitary presence of emerging pollutants in water bodies all around the world (e.g., Hirsch *et al.*, 1999; Weigel *et al.*, 2002; Carballa *et al.*, 2004; Le and Munekage, 2004), and highlights the problem of the ineffectiveness of wastewater treatment plants to remove certain of these contaminants before discharge into receiving waters. Despite that the detected concentrations (usually in the order of ng-µg/L) are often below the threshold limits defined in regulation or in water quality criteria (e.g., Directive 67/548/EEC, Water Framework Directive[4]) for acute toxicity, the continuous exposure of aquatic species to emerging pollutants raised concerns on potential long-term and mixtures effects (Wollenberger *et al.*, 2000).

Emerging pollutants represent a matter of concern also for the terrestrial environment. A review by Thiele-Bruhn (2003) exposed the problem of the presence of antibiotic compounds in soils, whose

[4] EU Water Framework Directice (WFD): Directive 2000/60/EC of the European Parliament and of the Council establishing a framework for the Community action in the field of water policy.

measured concentrations range from a few µg/kg up to g/kg, corresponding to those found for pesticides. The main problem posed by the presence of antibiotics in soil is the high biologic activity of these chemicals to the soil fauna, especially on soil microorganisms. Effects related to dose and persistence may change the composition of the microbial community and cause resistance phenomena in soil microorganisms.

The ubiquitary presence of emerging pollutants throughout the environment raised questions, which have not been solved yet, about i) the effects that these chemicals may have on humans, wildlife and ecosystems, ii) the ability to understand their environmental fate and behaviour (which require the characterization of physico-chemical and degradation properties, and metabolism), and iii) the adequacy of regulatory measures currently used to handle their introduction into the environment.

1.1.2. Chemical Risk Assessment

A central theme in the control of chemicals consists of the Chemical Risk Assessment (CRA), which is a systematic procedure finalized at characterizing the risk derived by the exposure of humans (human health risk assessment, HRA) and the environment (environmental risk assessment, ERA) to chemical substances. Risk assessment provides information based on the analysis of scientific data which describe the likelihood of harm to humans or the environment. CRA consists of four main steps, which include:

(1) *hazard identification*, i.e. the identification of the adverse effects that a substance is inherently able to cause. This first step involves gathering and evaluating data on the health effects (to humans and wildlife) that may be produced by the exposure to a chemical, as well as the characterization of the behaviour of a chemical within the body and in the environment.

(2) *exposure assessment*, i.e. the estimation of the concentrations/doses of a chemical to which human populations or environmental compartments are or may be exposed. Exposure can be assessed by measuring exposure concentrations or by applying multimedia exposure models, which estimate the emissions, pathways and rates of movement of a substance and its transformation or degradation. When performing ERA, the output of this step consists in the estimation of a predicted environmental concentration (PEC) for single environmental compartments.

(3) *effect assessment*, i.e. the estimation of the relationship between dose or level of exposure to a substance, and the incidence and severity of an effect (dose-response assessment). Data can be obtained from experimental plant and animal laboratory studies (i.e. *in-vivo* tests on standard species, preferably performed according to standardised guidelines), from field and epidemiologic studies, but also from alternative methodologies, including *in-vitro* testing and *in-silico* estimations such as quantitative

structure-activity relationships (QSARs) and read-across. The output of this step is the identification of a safe level under which adverse effects are not expected to occur. Considering the uncertainty derived by the extrapolation from model systems to humans or ecosystems (inter- and intra-species extrapolation), for most chemicals data generated from dose-response curves (e.g. EC_{50}, LC_{50}, LD_{50}, etc…) are converted into predicted no effect concentrations/levels (PNECs, PNELs) by applying specific assessment factors. Assessment factors are numbers, usually in the range of 10-10000, which reflect the estimated degree of uncertainty in the available data.

(4) *risk characterization*, i.e. the estimation of the likelihood of the adverse effects to occur in a human population or environmental compartment due to actual or predicted exposure to a substance. In ERA, a quantitative estimation of the risk is commonly expressed as PEC/PNEC ratios (i.e., risk quotients). The likelihood of adverse effects increases as the exposure/effect level ratios increase (van Leeuwen and Vermeire, 2007).

In the complex framework of chemical control and decision-making, risk assessment (RA), which is mainly a scientific task, is followed by the risk management (RM) process. Risk management is mainly a political process, which integrate information from RA with legal, political, socio-economic, ethical and technical considerations to develop, analyse and compare regulatory options and select the appropriate regulatory response to a potential health or environmental hazard (van Leeuwen and Vermeire, 2007). Figure 1.2 summarizes the main steps involved in the RA and RM procedures.

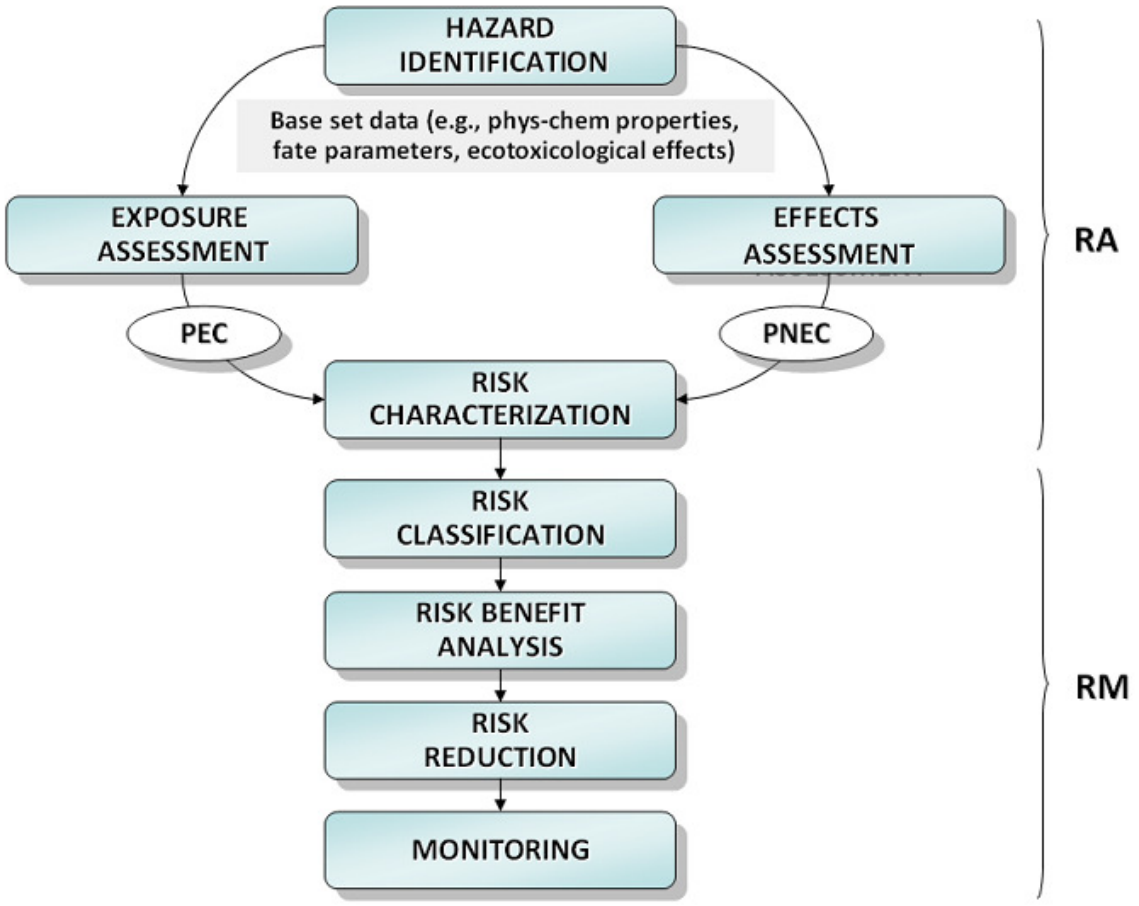


Figure 1.2. Main steps involved in risk assessment (RA) and risk management (RM) processes.

1.1.3. European regulation for chemical assessment

The European legislative framework in the context of chemical control and risk management evolved in the last decades. In the late 1960s, the increased environmental awareness highlighted the need to establish legally binding actions for the control of chemicals. The first legislative instrument adopted by the European Commission was the 1967 Council Directive on the classification, packaging and labelling of Dangerous Substances (Directive 67/548/EEC), which focused on harmonizing trade of industrial chemicals and protecting workers and consumers from exposure. In 1979, an important amendment to the Directive 67/548/EEC established a distinction between "new" chemicals (introduced onto the market after 1981) and "existing" chemicals (placed on the marked before 1981). While new substances had to be tested and assessed for their risks before they were placed on the market, existing chemicals, which counted more than 100,000 substances and over 99% of commercial chemicals in terms of tonnage, were not subject to the same requirements (Orellana, 2006), thus preventing a complete assessment of these chemicals.

In subsequent years EU legislation introduced separate legislations, which are dealing with water and air pollution, and are primarily aimed at the protection of human health. Separate legislations were also developed for specific categories of substances, e.g. plant protection products (Directive 91/414/EEC), biocides (Directive 98/8/EC), veterinary and human drugs (Directive 2004/28/EC and Directive 2004/27/EC, respectively), and cosmetics (Directive 2003/15/EC).

Two main conceptual shifts characterize the evolution of the European chemical legislation. First, the initial focus on the *hazards* of chemicals moved to the assessment of *risks*, thus promoting the development of exposure assessment methodologies and risk assessment models. Secondly, it was gradually recognized that legislation should not only protect human health but also seek to protect the environment, taking into account the entire life cycle of a chemical (van Leeuwen and Vermeire, 2007).

In the first years of 2000, it was recognized that the situation of the EU's legal framework on chemicals consisted of a patchwork of Directives and Regulations, which demonstrated a scarce efficiency in the assessment and management of chemicals and that were inadequate to secure a healthy environment for present and future generations. In fact, in three decades the EU had managed to fully assess approximately 140 chemicals out of the 2750 HPVCs[5], most of which (~85%) were lacking the basic information on physico-chemical properties, environmental fate and (eco-)toxicological effects.

The urgent need for a comprehensive and effective regulatory system for the management and marketing of chemicals led to the development of a new legal framework, i.e. the REACH regulation (EC) No 1907/2006, which was proposed for the first time in 2001 by the EU Commission in the form of a White Paper, and entered into force on 1st June 2007.

The aim of REACH (Registration, Evaluation, Authorisation and Restriction of Chemical substances) is to ensure a high level of protection of human health and the environment and, at the same time, to enhance innovation and competitiveness of the EU chemicals industry. Additional important objectives of REACH are the reduction of animal-testing, by promoting the use of alternative methods (e.g. *in vitro* testing and *in silico* techniques), and the progressive substitution of the most dangerous chemicals when suitable alternatives have been identified.

REACH introduces a single system for the regulation of both new and existing industrial chemicals, replacing about 40 existing EU Directives and Regulations.

The REACH Regulation moves the burden of risk identification and management from national authorities to the industry. All companies manufacturing or importing chemicals into EU in quantities of one tonne or more per year are required to gather information on the properties of chemicals, in order

[5] HPVCs (High Production Volume Chemicals), chemicals produced or imported in the EU in quantities of 1000 tonnes or more per year.

to ensure that they manufacture, place on the market or use substances that do not adversely affect human health or the environment. It is also mandatory for industries to register the information on chemicals in a central database that is managed by the newly established European Chemicals Agency (ECHA), placed in Helsinki, Finland.

REACH Regulation applies to all substances, preparations and articles produced or imported in quantities equal or above one tonne per year, while it doesn't include medicinal products for human or veterinary use, PPP (plant protection products) and biocides, cosmetic products, medical devices, food additives and feeding stuffs, non isolated intermediates, polymers and substances used for national defence. Special provisions have been approved for substances used in R&D (research and development).

The central requirements of REACH are defined by its acronym, i.e. Registration, Evaluation, and Authorization of chemicals, which define the supporting pillars of the new approach to chemical safety assessment. More in detail, REACH requires:

1) Registration: all the substances manufactured or imported in the EU in quantities above 1 tonne/year have to be registered by producers/importers to ECHA. The timing and amount of information required for registration depends partly on the volume produced/imported. For substances above one ton, a technical dossier (containing information on the properties, uses, classification, and guidance on safe use) must be submitted to the authorities. For substances above ten tons, a chemical safety assessment ("CSA") is required (and properly documented in a Chemical Safety Report – CSR), which includes the hazard classification of a substance and the assessment as to whether the substance is persistent, bioaccumulative and toxic ("PBT") or very persistent and very bioaccumulative ("vPvB"). Further, the CSR also describes exposure scenarios, including appropriate risk management measures, for all identified uses of dangerous, PBT, and vPvB substances. To reduce testing on vertebrate animals, data sharing is required for studies on such animals. New tests are only required when it is not possible to provide the information in any other permitted way, in order to minimize animal testing. In these situations, the manufacturer/importer would submit proposals for testing, which will be scrutinized by ECHA in the evaluation process before the tests are performed. According to the production/importation volumes and the potential hazard of chemicals (e.g. substances classified as carcinogenic, mutagenic and toxic to reproduction - CMRs), specific deadlines, between 2008 and 2018, have been fixed for the registration process.

2) Evaluation: this process consists of an examination of the data contained in the registration dossiers provided by industry, which undergo a double evaluation: dossier evaluation and substance evaluation. The dossier evaluation includes a check of the compliance with the registration requirements (if not,

industry is asked to provide further information), and a check of testing proposals, in order to prevent unnecessary testing with vertebrate animals (e.g. repetition of existing tests and poor quality tests). The substance evaluation includes the investigation of chemicals with potential risks to human health or the environment in order to identify substances of higher concern.

3) Authorization: an authorization is required for the use and marketing of "substances of very high concern" (SVHCs), which include substances classified as carcinogenic, mutagenic and toxic to reproduction (CMRs), PBT and vPvB substances, and substances of equivalent concern (e.g. endocrine disrupters). ECHA will grant an authorization for the use of such substances only if the industry demonstrates that either the risks associated to the use of the substance is adequately controlled, or that the socio-economic benefits outweigh the risks, taking into account alternative substitutes.

4) Restriction: restrictions, e.g. prohibition or specific conditions for the manufacture, trade or use, can be applied for certain dangerous substances. This procedure provides a safety net to manage risks that have not been adequately addressed by another part of the REACH system.

In addition to REACH, a new legislative tool for the classification, labelling and packaging of substances and mixtures, i.e. CLP Regulation (EC) No 1272/2008, was adopted by the EU and entered into force in January 2009. The CLP Regulation amends and replaces two previous directives related to the classification, packaging and labelling of dangerous substances (Directive 67/548/EEC) and preparations (Directive 1999/45/EC). With the purpose to facilitate worldwide trade and, at the same time, protect human health and the environment, the present Regulation incorporates and harmonises criteria for classification and labelling agreed at United Nations level, resulting in the Globally Harmonised System of Classification and Labelling of Chemicals (GHS). The GHS provides a harmonized system for globally uniform environmental, health and safety information on hazardous chemicals (substances and mixtures), aiming to protect workers, consumers and the environment by means of labelling which reflects possible hazardous effects of dangerous substances. Taking into account earlier EU legislations, the CLP Regulation introduces new classification criteria, hazard symbols (pictograms) and labelling phrases (risk and safety phrases, "R" and "S" respectively). The classification and labelling of chemicals is based on the intrinsic properties of chemicals.
Under CLP Regulation, companies are required to classify, label and package their hazardous chemicals appropriately before placing them on the market.

The implementation process of the new EU regulations requires that a large amount of data on the exposure and effects has to become available for thousands of chemicals, estimated to be around 30000,

in a short period of time (11 years). It is evident that to reach this challenging objective, a change in mind set was needed among regulatory authorities, industry and other stakeholder in order to accelerate the risk assessment and management processes (Schaafsma *et al*., 2009). To meet these needs, intensive efforts have been made to elaborate Intelligent or Integrated Testing Strategies (ITS), which are aimed at speeding up RA and RM procedures through a reduction of animal testing and an optimized intelligent use of the available information. ITS are integrated approaches comprising both testing and non-testing methods, such as *in vitro* testing, computational methods (i.e., (Q)SARs and kinetic models), chemical categories and read-across, Exposure-Based Waiving (EBW)[6], and optimized *in vivo* tests (Schaafsma *et al*., 2009). The combined use of these approaches allows to make the best use of the available information (on exposure and hazard) and to derive a *Weight of Evidence* (WoE) decision. In order to be effectively applied in regulation, results obtained from ITS approaches (as standalone methods or in combination) should be equivalent to those generated by standard testing and adequate to draw up an overall assessment, e.g. assessment of persistent, bioaccumulative and toxic chemicals (PBT assessment), or conclusions for classification and labelling (Ahlers et al., 2008).

The next section is dedicated to one of the ITS approaches mentioned above, i.e. quantitative structure-activity relationships (QSARs), to define its basic principles, regulatory requirements, and potential applications.

[6] Exposure-Based Waiving is based on the concept that toxicological testing is not necessary in case of "no relevant/significant", "limited" or "negligible" exposure (Schaafsma *et al*., 2009).

1.2. Quantitative Structure-Activity Relationship (QSAR)

The basic concept behind QSAR (Quantitative Structure-Activity Relationship) is to identify a quantitative relationship between the structure of a chemical and its biological activity, or, in case of QSPRs (Quantitative Structure-Property Relationship), a specific physico-chemical property.

In the last decades, QSAR modelling has become an important tool in different fields, e.g. chemistry, biochemistry and toxicology, because of its useful applications in environmental toxicology and chemistry as well as in drug discovery. The fundamental role of QSAR in the control of chemicals has been recognized by chemical industry and regulators (within the EU and elsewhere), especially after the REACH Regulation entered into force, requiring a large amount of data for thousands of chemicals in a relatively short period of time.

Current QSAR methodologies find their foundations in the pioneering works, in the mid-1960s, of Hansch and co-workers (Hansch C. *et al*., 1962; Hansch and Fujita, 1964), and Free and Wilson (Free and Wilson, 1964). According to the Hansch approach, hydrophobic, electronic and steric properties of molecules are combined to derive the following QSAR equation:

$$\text{Biological Activity} = a + b(\log P) + c(E) + d(S)$$

where the biological activity of a chemical is represented as a function of some physico-chemical and structural properties, which are responsible for the transport of the chemical into the cell and the binding to specific target. *LogP* (i.e. partition coefficient between n-octanol and water) is the term describing hydrophobicity, and represents the probability of a chemical to cross the cell membrane (driven by hydrophobic interactions) and reach the target site. The terms *E* and *S* encode for the electronic and steric properties of the chemical respectively, and represent the possibility of a chemical to interact with the target and be active. The toxicity of many chemicals has been modelled and predicted by equations of this kind. The limitation of the Hansch approach is that this equation is applicable only to chemicals “very similar” to, or congener with, those used to obtain the equation itself (i.e. “congenericity principle”).

The Hansch method initiated the beginning of the modern QSAR, which has made important progresses in the last 50 years toward more complex modelling approaches, supported by the continuous increase of the computing power and the introduction of several new methods (e.g. Partial Least Square Regression (PLS), Artificial Neural Networks (ANN), Bayesian approaches). An example is the 3D-QSAR analysis which starts from the 3D conformation of chemicals and correlates biological activities with 3D-property fields. A step further is the integration of 3D modelling with molecular docking and molecular

dynamics (MD) simulations, which allows to take into account inter-molecular interactions (e.g. ligand-receptor interactions) and to predict the behaviour of the whole complex. These approaches are widely applied for drug design.

The implementation of the REACH Regulation explicitly asks for the use of alternative techniques to animal testing, including QSARs, to assist in the assessment of hazardous properties of a substance, and which can, in certain cases, replace results from animal testing. The need to guarantee the scientific validity of the QSAR estimations to be used for regulatory purposes and to promote the mutual acceptance of QSAR models, led to the development of a set of general and internationally recognized principles for QSAR validation. Several principles for assessing the validity of QSARs were first proposed in 2002 at an international workshop held in Setubal (Portugal), which are known as the "Setubal Principles" (Jaworska *et al.*, 2003). These principles were then modified in 2004 by the OECD Work Programme on QSARs, and are now referred to as the OECD Principles for the validation of (Quantitative) Structure-Activity Relationship models, for regulatory purposes (OECD, 2004). The agreed principles provide member countries with a scientific basis for evaluating regulatory applicability of (Q)SAR models[7].

To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should fulfil the following requirements and have:

1) a defined endpoint

A (Q)SAR should be associated with a "defined endpoint", where endpoint refers to any physicochemical, biological or environmental effect that can be measured and therefore modelled. The intent of this principle is to ensure transparency in the endpoint being predicted by a given model, since a given endpoint could be determined by different experimental protocols and under different experimental conditions. Ideally, (Q)SARs should be developed from homogeneous datasets in which the experimental data have been generated by a single protocol. However, this is rarely feasible in practice, and data produced by different protocols are often combined.

2) an unambiguous algorithm

The intent of this principle is to ensure transparency in the model algorithm that generates predictions of an endpoint from information on chemical structure and/or physicochemical properties. Therefore a clear description of the dataset, molecular descriptors generation, modelling procedure and statistical

[7] This notation refers both to quantitative (QSAR) and qualitative (SAR) structure-activity relationship models.

methods and parameters used for validation is required. This principle also includes the need for reproducible predictions.

3) a defined domain of applicability

A (Q)SAR should be associated with a defined domain of applicability (AD), in which the model makes estimates with a defined level of accuracy (reliability). When applied to chemicals within its applicability domain, the model is considered to give *reliable results*. There aren't unique measures to define a model's AD or unique criteria to asses model reliability. Model reliability should be regarded as a relative concept, depending on the context in which the model is applied.

4) appropriate measures of goodness-of-fit, robustness and predictivity

This principle expresses the need to provide two types of information: a) the internal performance of a model (as represented by goodness-of-fit and robustness), which is determined by using a training set; and b) the predictivity of a model, determined by using an appropriate test set. It is important to note that there is no absolute measure of predictivity that is suitable for all purposes, since predictivity can vary according to the statistical methods and parameters used in the assessment.

5) a mechanistic interpretation, if possible

A (Q)SAR should be associated with a "mechanistic interpretation", wherever such an interpretation can be made. Clearly, it is not always possible to provide a mechanistic interpretation of a given (Q)SAR. The intent of this principle is therefore to ensure that there is an assessment of the mechanistic associations between the descriptors used in a model and the endpoint being predicted, and that any association is documented (OECD, 2007).

OECD principles constitute a conceptual framework to guide the development and validation of (Q)SARs, and represent a reference for the assessment of the scientific *validity* of the (Q)SAR models as well as of the *reliability* of QSAR predictions (Principle 3).

According to REACH, in order to consider a QSAR prediction for a given regulatory purpose, the *adequacy* of the QSAR result should be demonstrated, which means:

i. the estimate should be generated by a *valid* model,

ii. the model should be applicable to the chemical of interest with the necessary level of *reliability*, and

iii. the model endpoint should be *relevant* for the regulatory purpose (ECHA, 2008).

These conditions are illustrated in Figure 1.3.

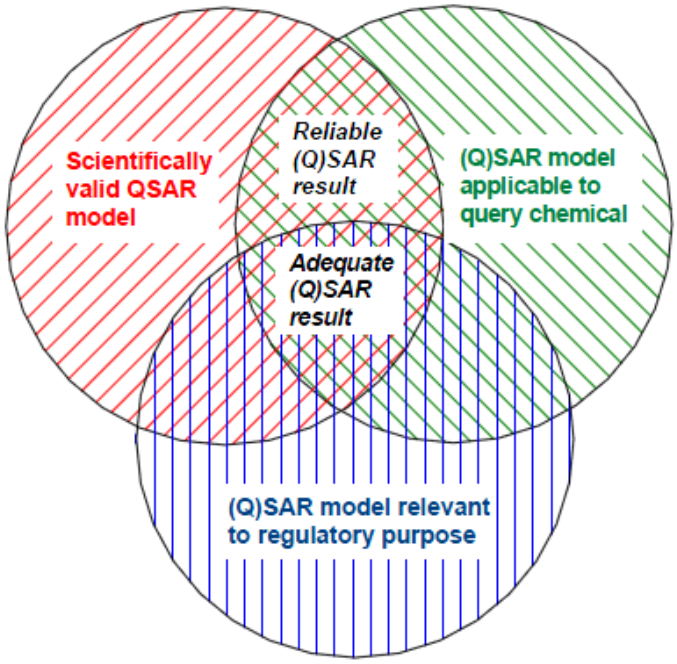


Figure 1.3. Interrelated concepts for considering a (Q)SAR prediction adequate for regulatory purposes within REACH (Figure from ECHA, 2008a).

Complete information regarding the scientific validity of QSAR models and adequacy of the generated predictions can be documented by using the appropriate QSAR Model reporting Format (QMRF) and QSAR Prediction reporting Format (QPRF). These reporting formats are required by REACH regulation (Annex XI) for the acceptance of QSAR predictions instead of test data.

1.3. The CADASTER Project

As described in section 1.1, one of the main goals of the REACH regulation is to gather the necessary amount of information to adequately assess and manage the risk of thousands of chemicals on the market, and, meanwhile, to reduce animal testing by optimized use of alternative approaches, e.g. *in silico* and *in vitro* information.

In the last decades, intensive efforts have been made by the scientific community in order to develop and improve valid instruments to meet that goal, i.e. alternatives to animal-testing, which can be used to support risk assessment and decision-making. These procedures, which are known as Intelligent Testing Strategies (ITS), include *in vitro* testing, (Q)SARs models, read-across and chemical grouping, and exposure-based waiving. To provide a practical support for their application in regulation (REACH and other regulations), several projects have been started under the EU sixth and seventh Framework Programmes for Research (e.g., OSIRIS, CAESAR, OpenTox[8], etc…). All these projects aimed at promoting the use of non-testing information for regulatory decision-making and, thus, minimizing the need for animal testing. Among these projects, CADASTER[9] (CAse studies on the Development and Application of *in-Silico* Techniques for Environmental hazard and Risk assessment), which was started in 2009 and will be concluded in 2012, aims at providing a practical guidance for the integration of alternative *in-silico* techniques, like QSAR and read-across, in the procedures of hazard- and risk assessments. The basic idea was to obtain the information needed for carrying out hazard and risk assessments for large numbers of substances by integrating multiple methods and approaches, with the aim to minimize testing, costs, and time. Chemicals belonging to four classes of emerging pollutants have been evaluated within the project, i.e. brominated flame retardants (BFRs), poly- and perfluorinated chemicals (PFCs), triazoles and benzo-triazoles (B-TAZs), as well as fragrances.

Within the CADASTER project, operational procedures for experimental design, QSAR development and validation, risk assessment and economic evaluations have been developed, tested, and disseminated in order to guide a transparent evaluation of the above mentioned four classes of emerging chemicals, taking into account variability and uncertainty in experimental data and in models.

The CADASTER Project involved nine partners from different European universities, public institutes and research centres, covering expertise in the fields of experimental testing, data mining, *in-silico* approaches, risk assessment and IT knowledge (Table 1.1).

[8] OSIRIS: Optimized Strategies for Risk Assessment of Industrial Chemicals through Integration of Non-Test and Test Information. CAESAR: Computer Assisted Evaluation of industrial chemical Substanced According to Regualtions. OpenTox: An open source predictive toxicology Framework.

[9] http://www.cadaster.eu/

Table 1.1. List of participants involved in the CADASTER Project.

Beneficiary name	Abbreviation	Country
Rijksinstituut voor Volksgezondheid en Milieu	RIVM	Netherlands
Public Health Institute Maribor	PHI	Slovenia
University of Insubria (Varese)	UI	Italy
IVL Swedish Environmental Research Institute	IVL	Sweden
Linnaeus University	LnU	Sweden
Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH)	HMGU	Germany
Ideaconsult Ltd.	IDEA	Bulgaria
Radboud University Nijmegen	RUN	Netherlands
Mike Comber Consulting	MCC	Belgium

In line with the general overview described above, CADASTER had 4 main objectives, which, apart from a work package on coordination (WP1), have been operationalized within 4 work packages (WPs).

1) Collection of data and models (WP 2)

This WP included the collection of existing experimental data and (Q)SAR models on the most common regulatory endpoints considered in the Screening Initial Data Set Dossier (SIDS[10] - internationally agreed data on the intrinsic hazards of a chemical) for the four classes of chemicals selected. Within WP2 new experimental data have been generated on endpoints and chemicals for which, as identified in WP 3, insufficient data were available for model validation and proper hazard/risk assessment.

2) Development and validation of QSAR models (WP 3)

WP3 activities were mainly focused on the development and validation of QSAR models for the prediction of physico-chemical properties and (eco-)toxicity for the four classes of chemicals selected. The models were developed in agreement with the OECD principles for the validation, for regulatory purpose, of (Q)SAR models (OECD, 2004). A preliminary evaluation of existing QSARs for the chemical classes studied, and their quality (i.e. compliance with OECD principles), was performed in order to identify the information gaps where new models were needed. An additional WP3 activity consisted of the application of similarity analysis and multivariate ranking methods for the identification of priority chemicals in the four selected classes to orient the experimental tests in WP2.

[10] http://www.oecd.org/env/chemicalsafetyandbiosafety/testingofchemicals/49944183.pdf

3) Integration of QSARs within hazard and risk assessment (WP 4)

The central issue of WP4 has been the integration of QSAR models into a probabilistic risk assessment framework. This activity included the quantitative characterization of uncertainty of QSAR predictions, sensitivity analyses of individual models with regard to their contributions in the overall risk assessment framework, and QSAR modelling of species sensitivity distributions (SSDs). Additional activities within this WP included i) the evaluation of methods and criteria for the establishment of scientific validity and applicability domains for QSAR models, ii) the evaluation of the ECETOC TRA screening risk assessment tool, in order to establish the ability of the tool for identifying chemicals of concern, at varying levels of data/information, and iii) the evaluation of the economic impacts of the substitution of chemicals from the four chemical classes, thus, fulfilling the aim of REACH of achieving a proper balance between social, economic and environmental objectives.

4) Development of website, newsletters/workshops and standalone tools for dissemination of project results (WP 5)

The core of WP5 has been the development of a on-line and standalone Decision Support System (DSS) for development, publishing and use of QSAR/QSPR models for REACH. This implied the development of a database storing all the collected experimental data and (Q)SAR models developed within the CADASTER project (QSPR-Thesaurus database[11]), which can be freely applied to generate predictions for new chemicals. In addition to (Q)SAR development and application, the QSPR-Thesaurus database allows the users to calculate molecular descriptors, and to perform experimental design, fate assessment and species sensitivity distribution (SSD).

A brief description of the four classes of chemicals studied within the CADASTER project is provided in the following paragraphs.

1.3.1 Brominated Flame Retardants (BFRs)

Brominated flame retardants (BFRs) are a class of hydrophobic chemicals that are incorporated in a variety of consumer products (e.g. electronic devices, building materials, textiles, etc..) to increase their fire resistance. Among the large number of commercially available brominated flame retardants, the three most marketed products are tetrabromobisphenol A (TBBPA), hexabromocyclododecane (HBCD), and polybrominated diphenyl ethers (PBDEs) (Figure 1.4) (Alaee *et al.,* 2003).

[11] http://qspr-thesaurus.eu/login/show.do?render-mode=full

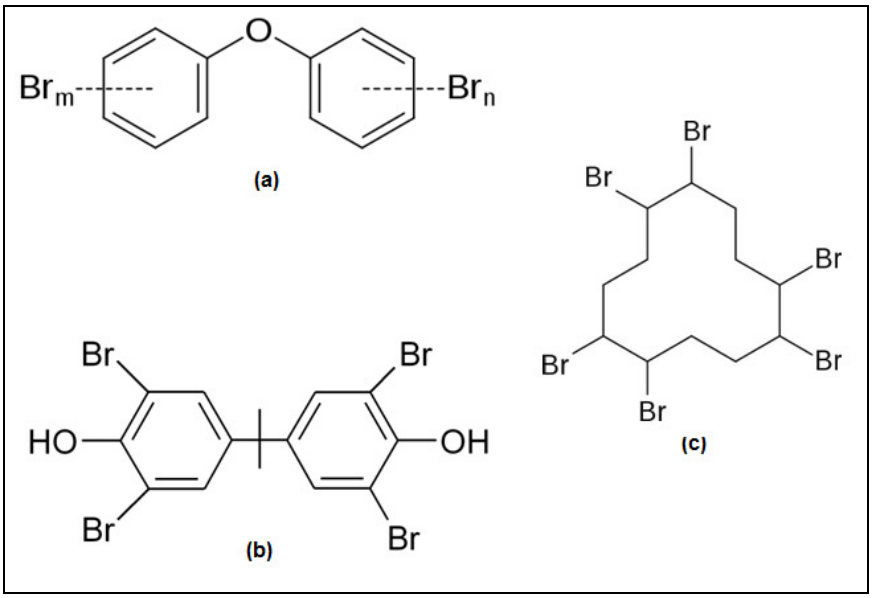


Figure 1.4. Chemical structures of PBDEs (a), TBBPA (b) and HBCD (c).

PBDEs potentially include 209 congeners divided into 10 congeneric groups (mono- to decabromodiphenyl ethers) which are characterized by different number and position of bromine substituents.

The widespread production and use of BFRs in the last 40 years has caused them to disperse throughout the environment (de Wit, 2002; Hale *et al*., 2006; Law *et al*., 2003; Law *et al*., 2006). BFRs are being found in almost all the environmental compartments (Watanabe *et al.*, 1992) as well as the indoor environment (e.g. because of their release from electronic devices), raising concerns for human exposure (Bergman, 1997; Sjödin *et al*., 2001; Sjödin *et al*., 2003; Harrad *et al*., 2004). Geographic trends show that BFRs may be subject to long range transport (LRT), thus contaminating also remote areas like polar regions (Wania and Dugani, 2003; Ikonomou *et al*., 2002; Stern and Ikonomou, 2000).

BFRs are very lipophilic compounds (logKow>5, Palm *et al.*, 2002), characterized by high environmental persistence and bioaccumulation tendency (Tomy *et al.,* 2004; Streets *et al.,* 2006; Burreau *et al.,* 2006). These properties make them likely to accumulate into organisms and in the food chains, thus explaining the high concentrations measured in biota, especially at the top levels of the trophic chains, such as birds, whales, polar bears, and even in humans (e.g. liver, adipose tissue and breast milk) (Burreau *et al.,* 2006; Muir *et al*., 2006; Sjödin *et al*., 2003; Meironyté *et al*., 2001; Norén and Meironyté, 1998).

The structural similarity of many BFRs to other classes of organohalogenated compounds, such as polychlorinated biphenyls (PCBs) and dioxins, suggests a potential dioxin-like toxicity. Several *in vivo* and *in vitro* studies highlighted for some groups of BFRs potential carcinogenic and teratogenic effects

(Helleday *et al.,* 1999), effects on neurobehavioural development (Eriksson *et al.,* 2001) and evidence of endocrine disrupting potency (Meerts *et al.,* 2001; Hamers *et al.,* 2006).

1.3.2 Fragrances

Fragrance materials are a heterogeneous group of chemicals widely used in many consumer products, including cosmetics, toiletries, household and laundry products, products used to scent the air (e.g., air fresheners and fragranced candles), as well as additives for food and drinks.

Despite that they share a common use pattern, it is not correct to consider fragrances as a single class of compounds. More than 2000 different organic chemicals have been identified as fragrance ingredients, exhibiting a wide range of physico-chemical properties. As can be observed in Table 1.2, 23 major structural classes have been identified on the basis of their chemical structure (Bickers *et al.,* 2003).

Table 1.2. Classification of fragrance ingredients based on chemical structure (Bickers *et al.,* 2003).

Structural group	**No. of chemicals**	**Structural group**	**No. of chemicals**
Esters	707	Pyrans	27
Alcohols	302	Miscellaneous	27
Ketones	259	Schiff's bases	26
Aldehydes	207	Heterocyclics	25
Ethers	100	Epoxides	25
Hydrocarbons	82	Sulfur containing	24
Acetals	63	Pyrazines	22
Lactones	61	Amines/amides	18
Carboxylic acids	42	Quinolines	14
Phenols	40	Musks	10
Nitriles	39	Coumarins	4
Dioxanes	31		
		Total	**2155**

Examples of compounds used as fragrance ingredients are musks (polycyclic, macrocyclic, alicyclic and nitromusks), salicylates, cinnamates and other esters with fragrances behaviour, substituted benzophenones, and terpene derivatives (Figure 1.5).

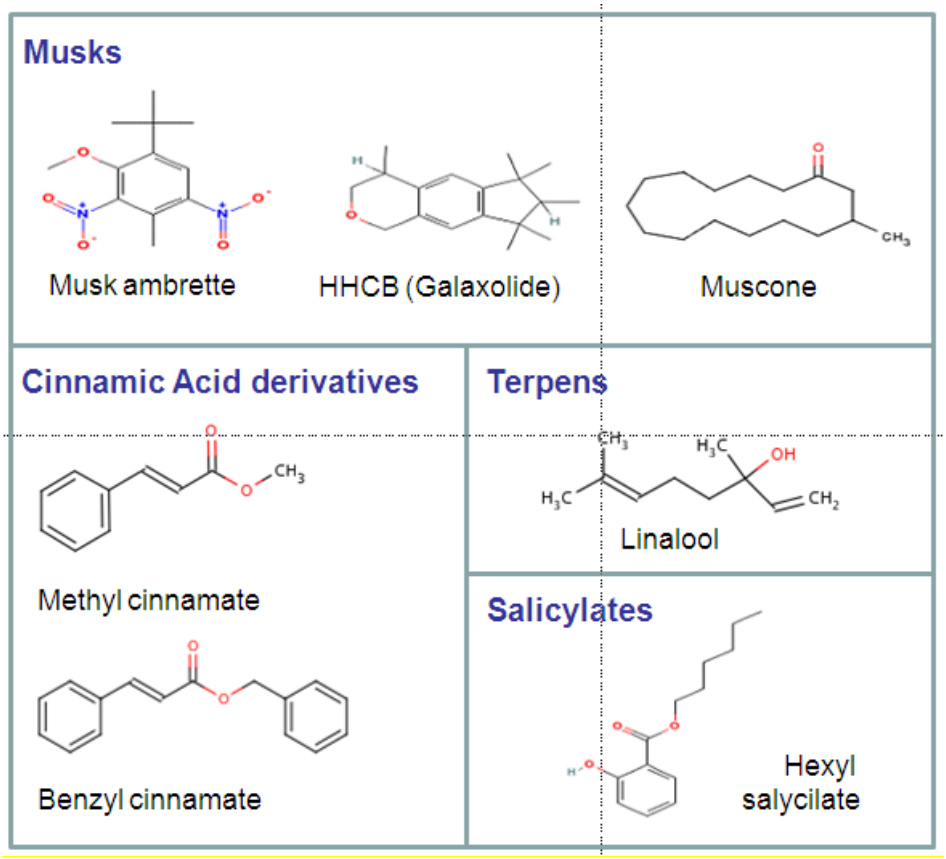


Figure 1.5. Examples of chemical structures of some of the most known fragrance materials.

In view of their typical use pattern, fragrance materials enter in the environment mainly through the wastewater treatment plants. Hence, it is the efficiency of WWTPs in removing these chemicals (through sorption, biodegradation or transformation mechanisms) that determines their presence in receiving waters and soils (through the use of WWTP sludge as a soil amendment in agriculture) (Salvito *et al.*, 2004).

Considering the wide use and potential exposure, there is limited information available related to human health and environmental effects of fragrances. Fragrance materials have been long recognized as skin allergens and irritants. However some of these chemicals, according to their physico-chemical and partition properties, are able to enter the body through numerous routes (e.g. skin absorption, inhalation, ingestion and the olfactory pathways) and impact any organ or system, thus posing concerns for systemic effects. Evidence of carcinogenicity, effects on development and reproduction, as well as endocrine disrupting potency have been reported for several fragrances (e.g. safrole, coumarin, methyleugenol, musk xylene, etc…) (Bridges, 2002 and references therein). Health concerns are also related to the potential of fragrances to induce adverse effects, or worsen existing problems (e.g. asthma) for the respiratory system (Bridges, 2002 and references therein).

The olfactory pathways represent a direct point of entry for chemicals into the brain (e.g. volatile fragrances contained in scented products), thus affecting the nervous system. Neurotoxic properties as well as physiological and drug-like effects (e.g. addiction to fragrances) have been also demonstrated for some fragrances (Aoshima and Hamamoto, 1999; Jirovetz *et al*., 1991; Hastings *et al*., 1991; Spencer *et al*., 1978; Spencer *et al*., 1984).

1.3.3 Perfluorinated Chemicals (PFCs)

Per- and polyfluorinated chemicals (PFCs) are a family of straight or branched long carbon chain analogs predominantly substituted by fluorine. The majority of PFCs are characterized by an hydrophobic fluorinated alkyl chain and a polar terminal group, mainly consisting of sulfonates or carboxylates (Figure 1.6).

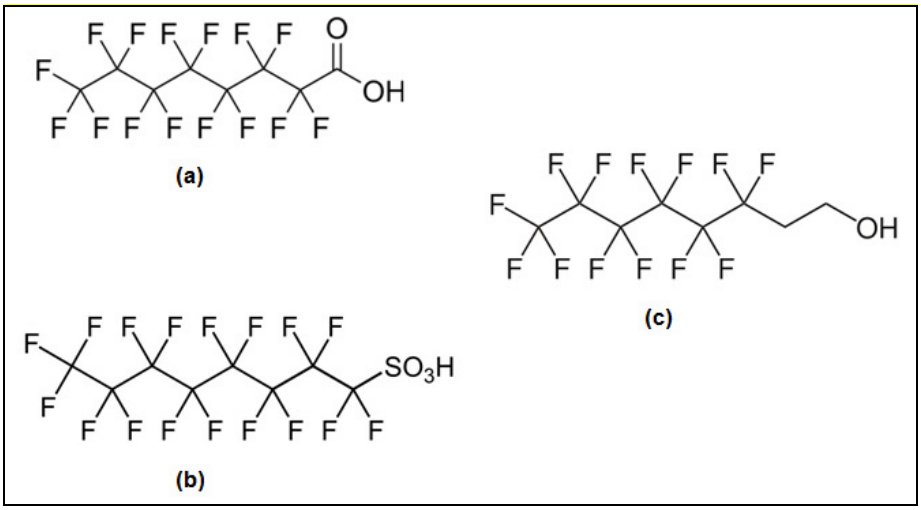


Figure 1.6. Examples of chemical structures of PFCs: (a) perfluooctanoic acid (PFOA), (b) perfluorooctane sulfonic acid (PFOS), (c) 6:2 fluorotelomer alcohol (6:2 FTOF).

Their amphiphylic nature makes PFCs suitable as surfactants. Because of their great thermal stability and stress resistance, these compounds are extensively used in a variety of household and industrial products (Laue *et al*., 2007; Prevedouros *et al*., 2006; Holzapfel, 1966), including waterproof fabrics, food packaging, non-adhesives, fire-fighting foams, cleansers, paints, etc…

The global production of some PFCs, such as perfluorooctanoic acid (PFOA), perfluorooctane sulfonate (PFOS) and perfluorooctane sulfonyl fluoride (PFOSF) (a precursor of PFOS), has been estimated to exceed 1000 tons/year (Harada and Koizumi, 2009).

The widespread production and use of PFCs for decades, combined with their high persistence, led to their global distribution, from urban to remote areas, contaminating almost all the environmental media and biota (Laue *et al*., 2007; Prevedouros *et al*., 2006; Dreyer *et al*., 2009; Giesy and Kannan, 2001;

Houde *et al*., 2006). Of particular concern is their occurrence in human blood, breast milk and tissue samples, due to the daily exposure to PFC-contaminated outdoor and indoor air, house dust, food and drinking water (Fromme *et al*., 2009).

The polar hydrophobic nature of PFCs determines a high affinity for physiological proteins (Biffinger *et al*., 2004). Once entered into the body, PFCs tend to accumulate in blood and in organs such as liver, kidneys, spleen, testicles, and brain. Among the possible toxic effects reported in literature, evidence of endocrine disruption (ED) has been found both in vivo and in vitro, and includes alteration of serum hormones levels (e.g. testosterone, estradiol, thyroxine) and expression of many genes involved in cholesterol transport and steroidogenesis (Jensen and Leffers, 2008; Shi *et al*., 2009; Weiss *et al*., 2009; Liu *et al.*, 2010).

1.3.4 Triazoles and Benzotriazoles (B-TAZs)

Triazoles and benzotriazoles (B-TAZs) are a class of synthetic molecules, which are structurally highly heterogeneous and characterized by the presence of a simple or condensed aromatic heterocyclic ring (2C + 3N atoms) (Figure 1.7).

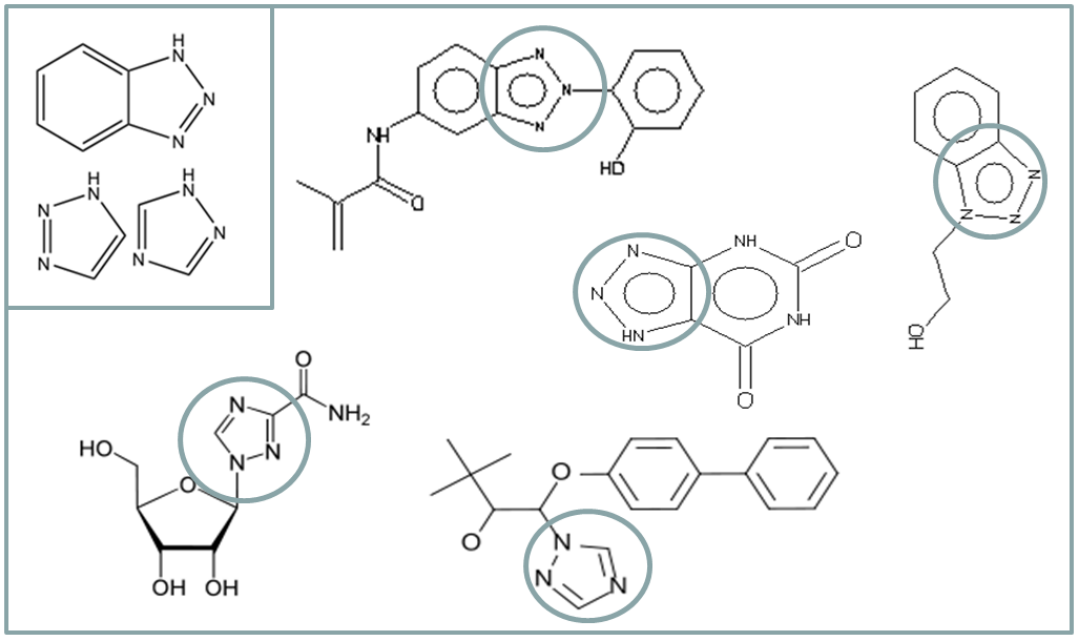


Figure 1.7. Chemical structures of triazoles and benzotriazoles.

B-TAZs are produced in large amounts (HPV chemicals) and are broadly used in various industrial processes, in agriculture as well as in households. Applications include their use as pesticides, pharmaceuticals (e.g. antimycotics, antidepressants), anti-corrosives, UV-light stabilizers for plastics, dishwashing additives, and as components of runway and aircraft de-icing fluids (ADFs).

B-TAZs, in particular benzotriazoles, may persist in the environment for a very long time due to their resistance to oxidation under environmental conditions, UV stability (Wu *et al.*, 1998), and resistance to biodegradation (Hem et al., 2003). B-TAZs are known to be fairly well soluble in water and to have a limited sorption tendency.

Releases to the environment, due to the wide production and multiple applications, caused their contamination of different environmental media. Environmental occurrence of B-TAZs was first detected in areas surrounding airports (Cancilla *et al.,* 1998; Breedveld *et al*., 2003; McNeill and Cancilla, 2008), probably due to their abundant use in ADFs. Because of the high water solubility and environmental persistence, B-TAZs have become ubiquitous contaminants of the aquatic compartment (Wolschke *et al.*, 2011, Giger *et al.,* 2006), thus raising concerns on the potential effects on aquatic organisms. Endocrine disrupting effects, such as pro-feminization or alteration of the thyroid system, and developmental toxicity have been observed both in aquatic and terrestrial organisms (Taxvig *et al.,* 2007, Kadar *et al.,* 2010).

1.4. Aim of the Thesis

In the context of the CADASTER project, the main topic of the present PhD thesis was the development of QSAR/QSPR models for the characterization of the (eco-)toxicological profile and environmental behaviour of chemical substances of emerging concern. The attention was focused on the four classes of compounds studied within the CADASTER project, i.e. brominated flame retardants (BFRs), fragrances, prefluorinated compounds (PFCs) and (benzo)-triazoles (B-TAZs). An important objective of this thesis was to make the best use of the available information to derive valid QSAR models that could be applied for the screening and prioritization of a large number of chemicals without experimental data. This was particularly relevant considering the limited amount of experimental data available for these emerging pollutants, especially for the basic endpoints required in regulation (e.g. REACH) for the hazard and risk assessment of chemicals.

Through several case-studies performed within the CADASTER project, the present thesis shows how QSAR models can be applied for the optimization of experimental testing as well as to provide useful information for the safety assessment of chemicals and support decision-making.

In the first case-study, the QSAR approach was applied for the characterization of endocrine disrupting properties (e.g. dioxin-like activity, estrogenic and androgenic receptor binding, interference with thyroxin transport and estradiol metabolism) of BFRs and PFCs (Chapter 3). An important issue stressed within this study was the dual role of QSAR models, i.e. i) their application for screening purposes ("predictive QSARs") and ii) their application for mechanistic investigation of a specific activity ("descriptive QSARs").

The second case-study presents the QSAR models developed for the prediction of three key endpoints required in regulation for the assessment of aquatic toxicity (i.e. acute toxicity in algae, daphnids and fish) for B-TAZs (Chapter 4), thus showing the potential use of QSAR estimations to support the overall assessment on chemicals' toxicity.

Finally, the third case-study addresses the development of QSPR models for the classification of ready biodegradability of fragrance materials, which is among the basic endpoints required for the assessment of the environmental persistence of chemicals (Chapter 5). The possibility to predict the potential biodegradability of chemicals on the basis of their structure, even before their synthesis, highlights the utility of QSAR/QSPR approaches in the rational design of new alternative compounds that are less persistent in the environment.

Chapter 2

Material and Methods

2.1 QSAR Procedure

As was stated in the previous section, QSA(P)R models are based on the definition of a quantitative relationship between the structure of a chemical and its biological activity (QSAR) or a specific physico-chemical property (QSPR). Figure 2.1 represents the theoretical scheme of the classic QSAR approach. For a given set of compounds, experimental data of physico-chemical properties or biological activities are determined by the measures *M* and *A* respectively. The structural information of the chemicals is described by molecular descriptors, which can be theoretically calculated by the procedure *D*. The relationships R1 and R2 are the mathematical functions of the QSARs and QSPRs. In some cases also mathematical function to quantify the Property-Activity relationships (QPAR) can be defined (R3) (Gramatica P., 2001).

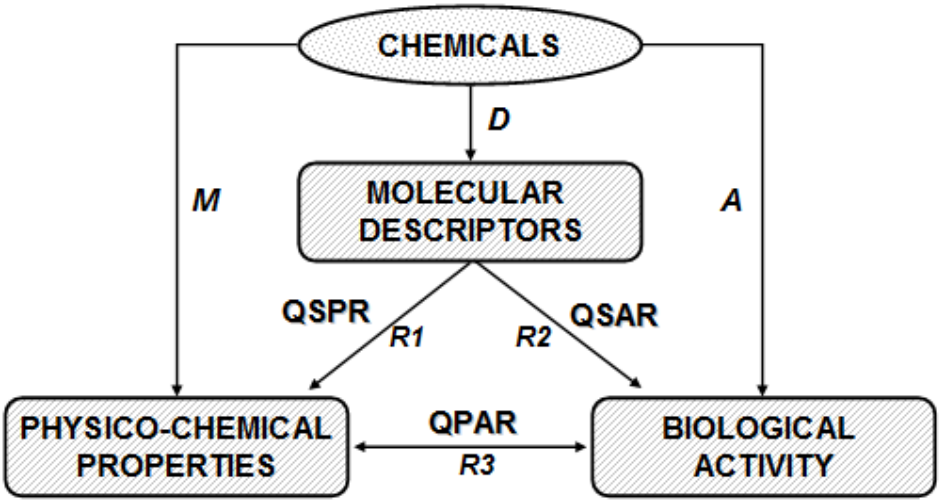


Figure 2.1. General scheme of the QSA(P)R approach.

Fundamental prerequisites to obtain good quality QSARs are the availability of input experimental data of high quality, an exhaustive representation of the chemical structure (e.g. through molecular descriptors, structural fragments), and the use of valid (i.e. statistically robust and predictive) and adequate (i.e. applicable to the compounds of interest) quantitative methods.

The following paragraphs will go through the principal steps involved in the development of QSAR models, describing the theoretical and methodological basis as well as the current challenging aspects.

2.2. Molecular representation and descriptors

A molecule is a complex structural system that can be represented through several different *molecular representations*, each constituting a different conceptual model and including different information related to chemical structure (e.g. 2D or 3D information). Examples of molecular representations commonly used for QSAR modelling, and used in the present thesis, include:

- the Simplified Molecular Input Line Entry Specification (SMILES), which is a line notation for molecules including information on connectivity among atoms, but not encoding for 2D or 3D coordinates.
- MOL files, for holding information about the atoms, bonds, connectivity and coordinates of a molecule.
- HIN files, which represent the minimum energy molecular geometries optimized by the HYPERCHEM software and encode for the mono-, bi- and tri-dimensional information of the molecules.

Chemical information from the different molecule representations is then extracted through the calculation of *molecular descriptors*, which are numerical variables quantifying the structural information of a chemical. "*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 a useful number or the result of some standardized experiment*." Many molecular descriptors have been proposed and derived from different theories and approaches, with the aim of predicting biological and physico-chemical properties of molecules (Todeschini and Consonni, 2000).

The information content of a molecular descriptor depends on the kind of molecular representation that is used and on the defined algorithm applied for its calculation. All the molecular descriptors must contain, to varying extents, the chemical information, must satisfy some basic invariance properties and general requirements, and must be derived from well-established procedures, which enable molecular descriptors to be calculated for any set of molecules.

Molecular descriptors are divided in two main classes: those derived by experimental measures (such as logP), and theoretical molecular descriptors, derived from a symbolic representation of the molecule.

LogP (or logKow) represents the octanol-water partition coefficient and is one of the most used parameter in QSAR modelling, especially in the past (it was already included in the Hansch equation). LogP is a property that can be either measured experimentally or calculated by freely and commercial packages (such as C-logP, DRAGON, and KowWIN from the EPIsuite software). However it is known that LogP values can be characterized by high variability, which may affect the quality of models based on logP (Renner, 2002; Papa *et al.,* 2005; Benfenati *et al.,* 2003).

The classical theoretical molecular descriptors can be derived in different ways:

- counting some atom-types or structural fragments in the molecule (0D and 1D- Descriptors). Some examples are the molecular weight (MW), the number of atoms included in the structure (nA), as well as list of fragments, functional groups or substituents present in the molecule.
- describing how the atoms are connected on the basis of the two dimensional representation, i.e. defining the connectivity of atoms in the molecule in terms of the presence and the nature of the chemical bonds (topological representation). Molecular descriptors derived from algorithms applied to a topological representation are called 2D-descriptors.
- representing the molecule as a rigid geometrical object in a three-dimensional space. This view allows a representation not only of the nature and connectivity of the atoms, but also the overall spatial configuration of the molecule. This representation of a molecule is called geometrical representation and the derived molecular descriptors are called 3D-descriptors

In the present thesis, three different software (one commercial and two freely available) have been used for the calculation of molecular descriptors:

- DRAGON software (ver. 5.5, 2007 – commercial software) was used to calculate mono-, bi- and tri-dimensional descriptors starting from the (x,y,z) coordinates of the chemical structure (HIN files).

  A list of the DRAGON descriptors calculated in this thesis is reported in Table 2.1.

Table 2.1. List of descriptors blocks included in the DRAGON package.

	Descriptor Blocks	
0D, 1D and 2D descriptors	Constitutional Atom centered Fragments Molecular Properties Walk and Path counts Information indices Burden eigenvalues Eigenvalues based indices 2D frequency fingerprints	Functional group counts Charge descriptors Topological descriptors Connectivity indices 2-D autocorrelations Topological charge indices 2D binary fingerprints
3D descriptors	Randic Molecular Profiles Whim descriptors	Geometrical descriptors GETAWAY descriptors

- PaDEL-Descriptor software (ver. 2.12, 2011 - freely available) was used to calculate mono- and bi-dimensional molecular descriptors and fingerprints starting from MOL files.

  A list of the PaDEL descriptors calculated in this thesis is reported in Table 2.2.

Table 2.2. Groups of descriptors included in the PaDEL-Descriptor software.

	Descriptor type	
1D-2D descriptors	Acidic group count	Fragment complexity
	ALOGP	Hbond acceptor count
	APol	Hbond donor count
	Aromatic atoms count	HybridizationRatioDescriptor
	Aromatic bonds count	Kappa shape indices
	Atom count	Largest chain
	Autocorrelation (charge)	Largest Pi system
	Autocorrelation (mass)	Longest aliphatic chain
	Autocorrelation (polarizability)	Mannhold LogP
	Basic group count	McGowan volume
	BCUT	Molecular distance edge
	Bond count	Molecular linear free energy relation
	BPol	Petitjean number
	Carbon types	Ring count
	Chi chain	Rotatable bonds count
	Chi cluster	Rule of five
	Chi path cluster	Topological polar surface area
	Chi path	Van der Waals volume
	Crippen logP and MR	Vertex adjacency information (magnitude)
	Eccentric connectivity index	Weight
	Atom type electrotopological state	Weighted path
	Extended topochemical atom	Wiener numbers
	FMFDescriptor	XLogP
		Zagreb index
Fingerprints	CDK fingerprints	EStateFingerprinter
	SubstructureFingerprinter	KlekotaRothFingerprinter

- QSPR-Thesaurus (on-line platform developed within the CADASTER Project[1] - freely available) was used to calculate mono- and bi-dimensional molecular descriptors and fingerprints starting from SMILES strings.

  A list of the molecular descriptors calculated using the QSPR-Thesaurus database is reported in Table 2.3.

  Table 2.3. Groups of descriptors calculated in the QSPR-Thesaurus database.

	Descriptor type	
1D-2D descriptors	E-state	AMBIT Descriptors
	ALogPS	MolPrint
	GSFragment	ISIDA fragments

[1] http://www.qspr-thesaurus.eu/home/show.do

In this thesis the 3D structure of each studied compound was drawn in the software HYPERCHEM (ver. 7.03 for Windows, 2002) or imported as SMILES string, and converted into 3-D structures. The energy optimisation was performed by AM1 semi-empirical method in the HYPERCHEM software.

2.3. Exploratory data analysis

Measurements related to biological activities and phisico-chemical properties, as well as structural properties (i.e. molecular descriptors) of compounds can be represented by complex matrices of multivariate data. To analyse these kind of data, chemometrical tools can be applied in order to extract information and select the most appropriate method to handle them. Therefore, exploratory analysis of the dataset, both in terms of structural representation and response domain, is an important step preceding model development.

In this thesis different exploratory techniques have been applied for different purposes, and are here briefly described.

2.3.1. Principal Component Analysis

Principal Component Analysis (PCA) is one of the best known procedures in multivariate statistics, which find application in many fields (e.g., chemistry, biology, economics, etc…). PCA allows the examination of the correlation pattern among variables and an evaluation of their relevance, the visualization of the elements by analyzing their inter-co-relationships (e.g., outliers, clusters), the synthesis of data description discarding noise, the reduction of data dimensionality by discarding unnecessary variables, and the finding of principal properties in multivariate systems. From a mathematical point of view, the aim of PCA is to transform *p*-correlated variables into a set of orthogonal variables which reproduce the original variance/covariance structure of the data. This means rotating a *p*-th dimensional space to achieve independence between variables. The new variables, called principal components (PCs), are linear combinations of the original variables along the direction of maximum variance in the multivariate space, and each linear combination explains a part of the total variance in the data. Starting from the original dataset containing *p* variables, a maximum of *p* principal axes can be derived. Additional details on the mathematical basis of PCA can be found elsewhere (e.g., Pearson, 1901; Todeschini, 1998; Jolliffe, 1986).

Because of their properties, PCA analysis is used to summarize, in few dimensions, most of the variability of a dispersion matrix of a large number of variables, providing a measure of the amount of variance explained by a few independent principal axes. The first two principal components define a plane, which represents the largest amount of variance. The elements are projected in this plane according to their *score*[2] values, in such a way as to preserve, as much as possible, the relative

[2] Scores, the transformed variable values corresponding to a particular data point.

Euclidean distances they have in the multidimensional space. The *loadings*[3] show the contributions of the variables to each component. The score and loading plots allow to overview the relation between the objects and the variables, respectively. The relation between the objects and the variables is derived by comparing (or plotting together) the scores and the loadings calculated for M principal components.

2.3.2. Experimental Design by Factorial Analysis

Statistical experimental design is a very useful method for the selection of chemicals to include in a training or an experimental set, as a fraction of a larger dataset. Statistical experimental design introduces systematic variation of several parameters simultaneously, in order to obtain as much information as possible from as few experiments or observations as possible (or, often, to optimize the number of experiments that need to be performed in order to acquire sufficient information to meet specific aims within acceptable time and/or cost constraints) (Box *et al.,* 1978).

In a factorial design observations are selected at high (+) and low (-) levels of each variable and in all the possible combinations. Center points are also included to detect and describe all the statistical variation of the data. Center points are located at the mean of the high and low settings for each variable and usually 3-4 observations are selected.

In this study four variables, represented by the score values extracted from the PCA analysis performed on molecular descriptors, were used to generate a factorial experimental design (2^4 observations) to select fragrances to be included in the training set of ready biodegradation (Chapter 5).

2.4. Modelling methods

2.4.1. Multiple Linear Regression (MLR)

The purpose of regression analysis is to describe the relationship between a dependent variable Y (quantitative response) and a set of independent variables (or predictor), in order to predict the values of Y from given values of the independent variables x_1, x_2, x_3, …, x_p. The regression model is the mathematical equation used to describe the relationship among response and predictor variables. Regression modelling may be used for descriptive or predictive purposes. The multiple linear regression model is described in algebraic form as:

$$y_i = b_0 + \sum b_i x + e_i \quad \text{or} \quad y = X \cdot b + e$$

where *x* denotes the predictor variable(s), *y* the response variable, , and *e* is the random error, also called model error or residual. In the alternative expression of the MLR equation, *y* and *b* are the vectors of the responses and estimated regression coefficients, respectively; the matrix *X* is usually

[3] Loadings, the weight by which each standardized original variable should be multiplied to get the component score.

called model matrix, whose columns are the independent variables used in the regression model (Todeschini R., 1995; software MOBY-DIGS, 2004).

In this thesis, the method used for the development of regression models was the Ordinary Least Squares (OLS) regression.

2.4.1.1 OLS method

OLS regression calculates a parametric linear model for a single response, and calculates unbiased, least squares coefficients. OLS should be used if the data set is well determined, i.e. if there are more observations than predictors and the predictors are not too highly correlated. This model assumes that the response is a linear function of the predictors, and that the errors are identically and independently distributed. Using OLS regression, the vector of regression parameters is computed minimizing the sum of squares of the differences between observed values (*y*) and values calculated using the regression equation ($\hat{y}$). Thus, to obtain a least squares best fit, each member of the matrix equation $y = Xb$ is multiplied by the transpose of matrix X (X'), i.e. $X'y = X'Xb$.

In this way, the rectangular matrix X produces a square matrix $X'X$, which can be inverted. The values of the coefficients of the OLS regression equation (b_{OLS}) are computed inverting the square matrix [$X'X$]:

$$b_{OLS} = (X'X)^{-1}X'y$$

Once the regression coefficient vector b has been estimated, the calculated responses are obtained from $\hat{y} = Xb_{OLS}$ and the estimated error vector e from $e = \hat{y} - y$.

2.4.2. Classification models

The purpose of classification analysis is to identify the relationship between a dependent categorical variable C (qualitative response) and a set of independent variables (or predictor), in order to predict the category (or class) C from given values of the independent variables x_1, x_2, x_3, …, x_p.

Classification consists in finding a mathematical model able to recognize the membership of each object to its proper class (G). Once obtained a classification model, the membership of new objects to one of the defined classes can be predicted.

Classification methods based on the calculation of distances between each *i*-th object and the class centroids[4] are among the most simple and popular approaches to classification. There are several distance measures that can be used, such as the Euclidean distance, which is one of the most commonly used, Manhattan distance, Canberra distance, Minkowski distance, Lagrange distance, Mahalanobis distance, etc… (Todeschini, 1998).

[4] The centroid of the g-th class is the p-dimensional vector of average values of the p variables calculated on the objects belonging to that class.

The class assignment of each object is based on the minimum distance between the object and each class centroid.

Classification results are commonly summarized in the *confusion matrix*, which is a square matrix GxG, where the rows represent the known (true) object assignments and the columns represent the assignments as provided by a classifier. An example of confusion matrix is given in Table 2.4 for 3 classes (A, B, C), where A, B, C and A', B', and C' represent true and assigned classes, respectively.

Table 2.4. Example of the confusion matrix.

	A'	B'	C'	Row sums
A	14	1	0	$n_A = 15$
B	3	10	2	$n_B = 15$
C	0	0	20	$n_C = 20$
Column sums	17	11	22	n = 50

The diagonal elements are the number of objects belonging to the *g*-th class correctly classified.

The confusion matrix is the main tool to estimate the classification parameters (section 2.5.3.2.).

Several classification methods have been defined (Frank and Friedman, 1989) which are based on different mathematical approaches. In this thesis the *k*-Nearest Neighbours (*k*-NN) method has been applied for classification purposes.

2.4.2.1 *k*-NN Classification method

k-NN is a not parametric (free from assumptions on the data distribution) classification method, which is performed on the basis of local information around each object and can be applied also for modelling datasets characterized by a nonlinear separation among classes (Sharaf *et al.,* 1986; Zheng and Tropsha, 2000). The *k*-NN method is based on similarity of objects (chemicals), where each object is assigned to the class most represented in its *k* nearest neighbours (i.e., *k* most similar compounds) and the *k* value represents the number of considered nearest neighbours of each objects. In this thesis, the similarity was measured by calculating the Euclidean distances between the descriptor vectors.

The *k*-NN final model is not a function but is given by an assembly of the selected distance measure, the best *k* value and all the training set objects.

2.5. Modelling approach

Figure 2.2. summarizes the modelling approach that was applied within this thesis.

As was anticipated in section 2.1, an important prerequisite for the development of good quality QSARs is the availability of input experimental data of high quality. Data curation represents a fundamental step preceding QSAR modelling (Gramatica *et al*., 2012; Tropsha, 2010; Fourches *et al*.,

2010; Muehlbacher *et al.,* 2011; Porcelli *et al.,* 2008; Li and Gramatica, 2010). Data curation do not only involves the careful selection of high quality data, but also the verification of chemical structures used as input for molecular descriptors generation. It has been demonstrated that even a small error in a chemical structure can result in significant differences in the prediction of the accuracy of a model, not only for chemicals with erroneous structural information, but also in the prediction of other chemicals using models that contain such errors (Young *et al.,* 2008). This issue is also well described in the recent publication by our research unit (Gramatica *et al.,* 2012), where practical examples are provided to highlight the importance of manual verification of structures generated from different sources and different molecular representation files (e.g., SMILES codes, MOL files, HIN files), as well as their influence on descriptors calculation.

The following paragraphs describe the conceptual scheme and principal steps characterizing the development and validation of the QSARs proposed in this thesis.

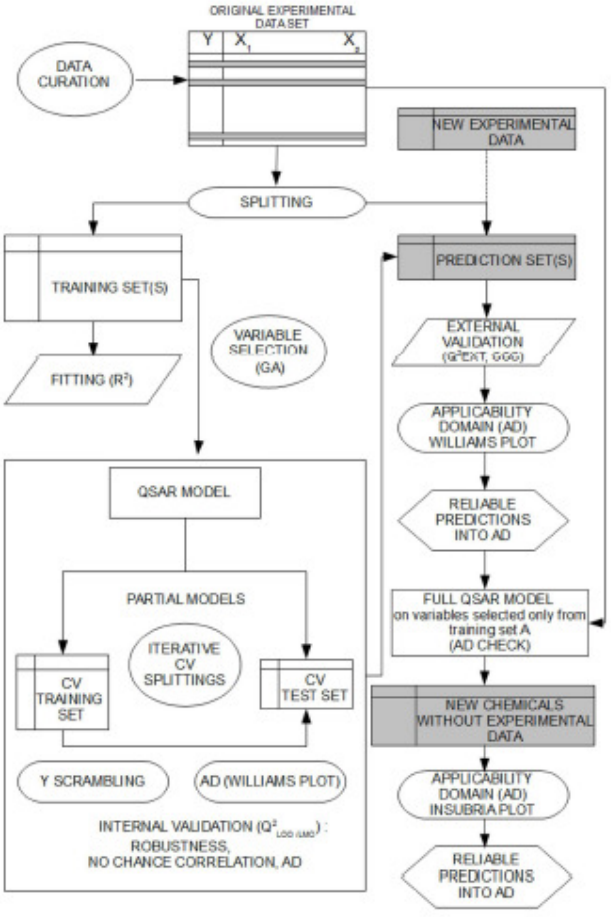


Figure 2.2. Scheme for the modelling approach (Figure from **Publication V**).

2.5.1. Dataset splitting

When a sufficient number of data was available (at least 15-20 compounds with experimental response), the dataset was divided to form a training set (used to build the model) and a prediction set, on which the external predictivity of the model was verified (section 2.5.3).

The splitting in training and prediction set should be balanced so that each of them is adequately representative of the structural and response domain of the original dataset.

To obtain this result, in most of cases two different splitting techniques were applied: a) by sorted response, and b) by structural similarity using Kohonen Artifical Neural Networks (K-ANN) (Gasteiger and Zupan, 1993).

In the first case, i.e. splitting by response, chemicals are ordered according to their increasing activity, and one out of every two (or three) chemicals are put in the prediction set. The resulting prediction set includes 50% (or 30%) of chemicals of the original dataset. The most and the least active compounds of the dataset are normally kept in the training set. In case of categorical responses, 50% (or 30%) of prediction set chemicals are randomly sampled within each class.

This splitting guarantees that both training and prediction sets cover the entire range of the experimental response and are numerically representative of the dataset. However, such splitting does not guarantee that the two sets represent the entire structural space of the original dataset. Therefore, it is possible that some compounds in the prediction set are outside the structural domain of the training set and could be wrongly predicted.

The splitting of the data set based on structural similarity, realized by K-ANN method, takes advantage of the clustering capabilities of K-ANN, allowing the selection of a structurally meaningful training set and an equally representative prediction set. Chemicals are grouped into clusters of similar compounds in a Kohonen map. Structural information is provided by molecular descriptors (in this thesis, the first three principal components of each block of DRAGON descriptors were used). As an output of the K-ANN, structurally similar chemicals are grouped within the same cluster (i.e., *neurons* of the top map). From each cluster, a selected percentage of compounds (30-50%) are randomly selected to be included in the prediction set. This splitting techniques allows to generate training and prediction sets that are structurally balanced.

2.5.2. Variable selection methods and model development

Variable selection is a necessary step to find simple and predictive QSARs, which should be based on the least number of low correlated descriptors (following Ockham's Razor philosophy). This procedure is particularly important since is known that some molecular descriptors provide only different views of the same molecular aspect. This implies the presence of highly correlated descriptors within a dataset. A selection of significant descriptors highly correlated to the studied

response is performed by applying unbiased mathematical tools, and starting from a large amount of molecular descriptors that can be calculated for the molecules included in a dataset.

The first step consists in a pre-reduction procedure that allows to remove all the constant and near-constant descriptors, as well as those characterized by a high pair-wise correlation (in general over 0.9, 0.95 or 0.98). The selection of the descriptors highly related to the response is then performed. Several selection methods are currently available (e.g. stepwise regressions, forward selection, backward elimination, simulated annealing, evolutionary and genetic algorithms, etc.) (Todeschini, 1998). In the present thesis, the Genetic Algorithms (GAs) were used for the variable selection in both regression and classification models. GAs are useful methods that are widely and successfully applied in many QSAR approaches (Leardi R., 1992).

The GA strategy for variable subset selection (GA-VSS) is based on the evolution of a population of models, i.e. a set of ranked models according to some objective function. In genetic algorithm terminology, each population individual is called chromosome and is a binary vector, where each position (a gene) corresponds to a variable (1 if included in the model, 0 otherwise). Each chromosome represents a model given by a subset of variables.

The Genetic Algorithm (GA) works, based on three main steps:

- **Random initialisation of the population.** The model population is built initially by random models with a number of variables between 1 and L. The value of the selected objective function of each model is calculated in a process called evaluation. The models are then ordered with respect to the selected objective function – model quality - (the best model is in first place in the population, the worst at position P);
- **Crossover.** From the actual population, pairs of models are selected (randomly or with a probability function of their quality). Then, from each pair of selected models (parents), a new model is generated, preserving the common characteristics of the parents (i.e. variables excluded in both models remain excluded, variables included in both models remain included) and mixing the opposite characteristics according to the crossover probability. If the generated son coincides with one of the individuals already present in the actual population, it is rejected; otherwise, it is evaluated. If the objective function value is better than the worst value in the population, the model is included in the population, in the place corresponding to its rank; otherwise, it is no longer considered. This procedure is repeated for several pairs;
- **Mutation.** After a number of crossover iterations, the population proceeds through the mutation process. This means that for each individual of the population every gene is randomly changed into its opposite or left unchanged. Mutated individuals are evaluated and included in the population if their quality is acceptable. This process is controlled by

mutation probability which is commonly set at low values, thus allowing only a few mutations and new individuals not too far away from the generating individual.

An important characteristic of the GA-VSS method is that does not provide a single model but a population of acceptable models. Within this population, there could be various models with similar predictive power, but based on different molecular descriptors.
In fact different descriptors are alternative viewpoints to represent the structural features, whose combination lead to, not equivalent, but similar results for the studied end-point. Thus, there could be many possible "best" models.

In the context of this thesis, different software were used for variable selection and model development. The MobyDigs software was used for the development of the regression QSAR models for the prediction of endocrine disrupting potency of BFRs and PFCs (Chapter 3). An "in house" application was used for the selection of the classification models (Chapter 3 and Chapter 5). Finally, the QSARINS software (2012), which was recently proposed by our research unit, was used to develop QSAR models for aquatic toxicity of B-TAZs (Chapter 4).

2.5.3. Validation of QSARs for their goodness-of-fit, robustness and predictivity

To guarantee scientific validity and reliability, QSAR models should always be verified for their ability to reproduce data used for training the model (i.e., goodness-of-fit), for their internal robustness and, when possible, for their ability to predict new data (i.e., external predictivity). External predictivity is of particular importance when QSAR models are proposed for screening, ranking and prioritization purposes.
The importance of QSAR validation is also stressed in the fourth principle established by OECD for the acceptability of QSARs for regulatory purposes (OECD, 2004).

2.5.3.1. Validation techniques and parameters used for regression models

There are many statistical indices useful to evaluate the performance of the developed regression models. A first group of them are devoted to evaluate model's fitting ability, providing a measure of how well the regression model accounts for the variance of the response variable. Several fitness functions have been proposed and are here summarised.
- Residual Sum of Squares (RSS): sum of squared differences between the observed (y) and estimated response ($\hat{y}$)

$$RSS = \sum_{i-1}^{n} (\hat{y}_i - y)^2$$

being *n* the number of training objects. This quantity is minimized by the least square estimator.

- Model Sum of Squares (MSS): sum of the squared differences between the estimated responses ($\hat{y}$) and the average response ($\bar{y}$)

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

This is a part of the total variance explained by the regression model as opposed to the residual sum of squares RSS.

- Total Sum of Squares (TSS): sum of the squared differences between the experimental responses and the average response

$$TSS = \sum_{i-1}^{n} (y_i - \bar{y})^2$$

This is the total variance that a regression model has to explain and is used as a no-model reference quantity to calculate standard quality parameters such as the coefficient of determination.

- Coefficient of determination (R^2): total variance of the response explained by a regression model. It can be calculated from the model sum of squares MSS or from the residual sum of squares RSS

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

A value of one indicates perfect fit, i.e. a model with zero error term.

- Root Mean Square of Errors (**$RMSE_{TR}$,** if calculated for the training set; **$RMSE_P$** if calculated for the prediction set): sum of the overall error of the model. It is calculated as the root square of the sum of the squared errors in predictions divided by their total number:

$$RMSE(P) = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{n}}$$

Differently from R^2, whose values vary between 0 and 1, RMSE values are influenced by the range of the response. Therefore, this parameter is useful if applied to compare different models based on the same (or very similar) training set, and developed for the same range of experimental response.

A second group of regression parameters are devoted to evaluate the goodness of prediction, i.e. the model capability to estimate future data, providing a measure of how well the regression model estimates the response variable given a set of values for predictor variables. These quantities are

obtained using validation techniques and are also used as criteria for model selection. A number of statistical techniques have been proposed to simulate the predictive ability of a model:

- Leave-One-Out and Leave-many-Out

The simplest and most general cross-validation procedures are the leave-one-out (LOO) and leave many (LMO) techniques, where a number of objects (one or more than one at time) are excluded during the model development. For each reduced data set, the model is calculated and responses for the excluded objects are predicted from the model. The squared differences between the true response and the predicted response for the object left out are added to PRESS (predictive residual sum of squares). From the final PRESS, the Q^2 value is normally calculated (Cruciani *et al.,* 1992).

$$Q^2 = 1 - \frac{PRESS_{CV}}{TSS}$$

LOO and LMO, widely applied as internal validation (or cross validation "CV") techniques, give in several cases a too optimistic predictive ability, particularly considering the LOO that introduces just a small perturbation in the dataset. However, even if LMO gives a more realistic idea of the internal predictivity than LMO, it can't be considered as a representative parameter for the real predictivity (external predictivity) of a model (Tropsha A. *et al.,* 2003).

- Y-scrambling

This validation technique is adopted to check models with chance correlation, i.e. models where the independent variables are randomly correlated to the response variables. The test is performed by calculating (several hundred of times) the quality of the model (usually R^2 and Q^2) randomly modifying the sequence of the response vector *y*, i.e. by assigning to each object a response randomly selected from the true responses (Lindgren *et al.,* 1996). Low values of the averaged R^2 and Q^2 scrambled (R^2_{YS}, Q^2_{YS}) are indicative of a well founded (not by chance) model.

- QUIK rule

The Quik rule (Todeschini *et al.,* 1999) is normally applied in order to select only models where the correlation between the block of the modeling descriptors and the response (K_{XY}) is higher than the correlation among the descriptors (K_{XX}), i.e. $K_{XY} > K_{XX}$.

- External validation

External validation techniques verify model predictivity toward compounds never used for model development. This set of "external compounds" can either be obtained *by a priori* splitting of the data set (*prediction set*), according to the procedures explained in paragraph 2.5.1, or be represented by new data became available, or produced, after model development. Several validation parameters are currently available to evaluate external predictive ability of QSARs and are summarized in Table 2.5.

Table 2.5. Statistical parameters for external validation

Statistic	Definition	Equations and terms
Q^2_{F1} [Shi *et al.*, 2001]	Variance explained in external prediction	$Q^2_{F1} = 1 - \frac{PRESS_{EXT}}{SS_{EXT}(\bar{y}_{TR})}$ $SS_{EXT}(\bar{y}_{TR}) = \sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{TR})^2$ $\bar{y}_{TR}$ = average of training observed responses
Q^2_{F2} [Schüürmann *et al*., 2008]	Variance explained in external prediction	$Q^2_{F2} = 1 - \frac{PRESS_{EXT}}{SS_{EXT}(\bar{y}_{EXT})}$ $SS_{EXT}(\bar{y}_{EXT}) = \sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2$ $\bar{y}_{TR}$ = average of external observed responses
Q^2_{F3} [Consonni *et al.*, 2009]	Variance explained in external prediction	$Q^2_{F3} = 1 - \frac{\left(\frac{PRESS_{EXT}}{n_{EXT}}\right)}{\left(\frac{TSS}{n_{TR}}\right)}$ n_{EXT} = number of external objects n_{TR} = number of training objects
CCC [Chirico and Gramatica, 2011; Chirico and Gramatica, 2012]	Concordance correlation coefficient	$\hat{\rho}_c = \frac{2\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(y_i - \bar{y})^2 + n(\bar{x} - \bar{y})^2}$ x_i = external response observed for the *i-th* object y_i = external response predicted using the model $\bar{x}$ = average of observed responses $\bar{y}$ = average of responses predicted by the model

2.5.3.2. Validation techniques and parameters used for classification models

Validation techniques applied for classification models are the same explained for the regression models, and include cross validation techniques (by LOO and LMO) and external validation. Several validation parameters are used to evaluate model classification accuracy, i.e. ability of the model to assign chemicals to the correct (real) class.

- Non-Error Rate (NER) or Overall Accuracy (OA): percentage of objects correctly classified

$$NER\% = \frac{\sum_g c_{gg}}{n} \times 100$$

where c_{gg} are the diagonal elements of the confusion matrix.

It represents the ability of a classifier to correctly assign all the objects of the classes.

The same quantity can be defined for each class separately (Class Non-Error Rate, NER_g):

$$NER_g\,\% = \frac{c_{gg}}{n_g} \times 100$$

When dealing with binary classification models, i.e. models classifying objects into two groups on the basis of whether they have some property or not (Table 2.6) (such as carcinogen/not carcinogen, active/inactive), two statistical parameters are commonly used to evaluate classification accuracy, i.e.:

- **Sensitivity (Sn)**: proportion of actual positives which are correctly identified

$$Sn = TP / (TP + FN)$$

- Specificity (Sp): proportion of negatives which are correctly identified

$$Sp = TN / (TN + FP)$$

where TP (true positive) is the number of compounds correctly classified as active, TN (true negative) is the number of compounds correctly classified as inactive, FN (false negative) is the number of active compounds classified as inactive, and FP (false positive) is the number of inactive compounds classified as active (Cooper *et al.,* 1979).

Table 2.6. Confusion matrix in binary classification.

	Assigned class	
Real class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

2.5.4. Applicability Domain Analysis

As stressed in the third OECD principle (OECD, 2004), any QSAR model is characterized by a specific Applicability Domain (AD), i.e. a theoretical spatial region in which the model is expected to provide predictions with a given reliability. Ideally, model AD should express the structural, physico-chemical and response space of the model, which is mainly dependent on the nature of the chemicals used to develop the model.

Several methods are currently available for the evaluation and definition of the applicability domain of QSAR models (Netzeva *et al.,* 2005; Sushko *et al.,* 2010; Sahigara *et al.,* 2012).

In the present thesis, the approach used to define model AD was based on the theoretical structural space defined by the descriptors used in the models (i.e. modelling descriptors). In most of cases, the AD was quantitatively defined by the leverage approach (Tropsha *et al.,* 2003; Gramatica, 2007; Eriksson *et al.,* 2003). The leverage (*h*) is a measure used to quantify the distance of a compound from the structural space of a model. Leverage values are the diagonal elements of hat matrix (Atkinson, 1985) and are calculated by:

$$h_i = x_i^T (X^T X)^{-1} x_i \quad (i = 1, \ldots, m)$$

where x_i is the descriptor row-vector of the query compound *i*, *m* is the number of query compounds and *X* is the *n* x *p* matrix of the training set (*n* is the number of training set samples and *p* the number of model descriptors).

The boundaries of model domain are defined by the leverage cut-off (h^*), which is set as 3(p+1)/n. Leverage values greater than h^* mean that the query compound is outside of the model AD; predictions generated for these chemicals are extrapolations of the model and could be not reliable. The leverage approach takes into account the multivariate combination of the modelling descriptors.

Model AD can be graphically visualized using the *Williams plot* (Figure 2.3), which is the plot of leverage values versus standardized residuals (std residuals between experimental and predicted responses). This plot allows to verify the presence in the training and prediction sets of structural outliers (i.e. compounds with h_{ii} greater than h^*) and response outliers (i.e., compounds with standardized residuals greater than 2.5 standard deviation units). It is particularly useful to identify structural outliers within the training set since these chemicals are structurally highly influential on the selection of the modelling variables.

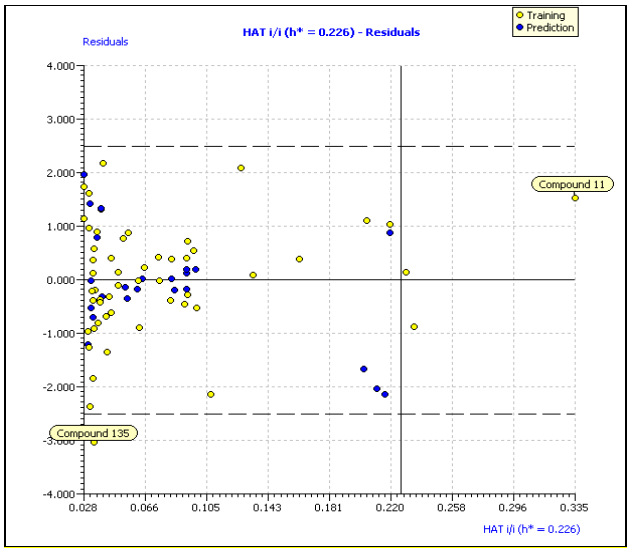


Figure 2.3 Example of Williams plot.

The leverage approach can be also applied to evaluate the degree of extrapolation of the predictions obtained for chemicals lacking experimental data. The chemicals with h_{ii} greater than h^* fall outside the AD of the model. Predictions generated for these chemicals are extrapolated by the model and should be considered as not reliable. Graphically, the plot of hat diagonal values versus predicted values was proposed by our research unit (and referred to as *Insubria Graph*) to visualize interpolated and extrapolated predictions for chemicals without experimental data (Figure 2.4).

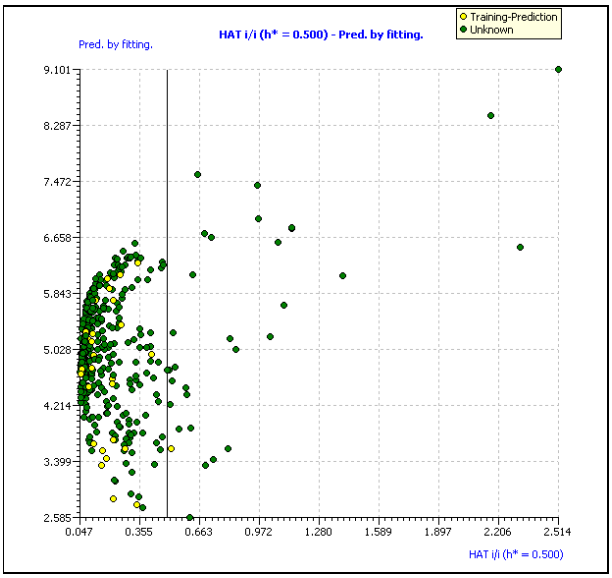


Figure 2.4 Example of Insubria graph.

Chapter 3

QSAR Modeling of Endocrine Disrupting Potency: BFRs and PFCs

3.1 Introduction

An important issue in chemical risk assessment concerns the potential risk to humans and wildlife posed by exposure to both natural and man-made chemicals that are capable of modulating or disrupting the endocrine system. A comprehensive definition of endocrine-disrupting chemicals (EDCs) have been proposed by Kavlock *et al.* (1996), who defined EDCs as "exogenous agents that interfere with the synthesis, secretion, transport, binding, action, or elimination of natural hormones in the body which are responsible for the maintenance of homeostasis, reproduction, development and behavior". Adverse effects caused by endocrine disruptors not only affect the exposed organisms, but also their progeny and populations.

There is a fair agreement in considering endocrine disruption as a mode or mechanism of action, rather than an adverse health effect, which can potentially lead to other outcomes, such as carcinogenic, reproductive, or developmental effects, routinely considered for regulatory decision-making. Evidence of endocrine disruption alone can influence priority setting for further testing and the assessment of the results of this testing could lead to regulatory action if adverse effect are shown to occur (EPA, 2011[1]; van Leeuwen and Vermeire, 2007).

Endocrine disruption is a complex area to address since there are many possible modes of action to take into account, and it is difficult to establish causal links between exposure to suspected EDCs and any measured effects (Vos *et al.,* 2003).

Different mechanisms have been recognized through which EDCs may interfere with the endocrine system: (i) agonistic effect, by binding to the cellular receptor of a hormone and activating normal cell response at the wrong time or to an excessive extent; (ii) antagonistic effect by binding to the receptor, preventing natural hormonal binding and activation of the receptor; (iii) alteration of hormonal blood levels by binding to hormone transport proteins; (iv) interference with metabolic processes by affecting the synthesis, or the elimination rate, of hormones[2].

The universe of endocrine disrupting chemicals is wide and not yet fully defined. Chemicals with ED activity include natural and synthetic hormones (e.g. contraceptive pills), which are released to the environment, mainly through sewage treatment plants, and can exert hormonal actions on other animals, and man-made chemicals intended for other purposed but that may interfere with the endocrine system of living organisms. Man-made chemicals that are associated with endocrine disruption in wildlife and in human individuals include some pesticides, such as DDT (dichlorodiphenyltrichloroethane), its metabolites and other chlorinated compounds, and a number of industrial chemicals, industrial by-products and chemicals used in consumer products, such as

[1] http://www.epa.gov/endo/pubs/edsparchive/2-3attac.htm
[2] http://ec.europa.eu/environment/endocrine/definitions/endodis_en.htm

dioxins, polychlorinated biphenyls (PCBs), polybrominated diphenyl ethers (PBDEs), and respective polychlorinated/polybrominated dibenzo-p-dioxins (PCDDs/PBDDs) and dibenzofurans (PCDFs/PBDFs). Despite experimental evidences show that the hormonal activity of the majority of these chemicals is many times weaker than the physiological hormones (e.g., Moore *et al.,* 1997; Hamers *et al.,* 2006), endocrine-mediated adverse effects on human and wildlife due to exposure to EDCs, which are ubiquitous in the environment, have been demonstrated (Harrison *et al* 1995; Colborn *et al.,* 1995 and references therein).

In the EU REACH regulation, endocrine disrupting chemicals are included in Title VII (Article 57-f), which deals with the authorization of substances of very high concern (SVHC). SVHCs include: i) substances classified as carcinogenic, mutagenic or toxic for reproduction (CMR) category 1 or 2, in accordance with Directive 67/548/EEC, ii) substances which are persistent, bioaccumulative and toxic (PBT) or very persistent and very bioaccumulative (vPvB) in accordance with the criteria set out in Annex XIII of REACH, and iii) substances having endocrine disrupting properties (or those having PBT, vPvB properties without fulfilling the criteria for PBT, vPvB) for which there is scientific evidence of probable serious effects to human health or to the environment which give rise to an equivalent level of concern to those of CMR, PBT or vPvB substances. Always according to article 57, endocrine disrupters are identified on a case-by-case basis. In fact, while clear guidance exists for CMR and PBT substances, internationally agreed methodologies or criteria are not available at the moment for the assessment of endocrine disrupting properties. Therefore, decision for inclusion of EDCs in Annex XIV (i.e. authorization list) will be based on the available information, e.g. from several independent sources or newly developed test methods leading to the assumption/conclusion that a substance has or has not ED property, and a weight of evidence approach will be used.

In this context, information derived from (Q)SAR predictions is useful to identify chemicals with endocrine disrupting potency, avoiding time-consuming and expansive testing, and to support decision-making.

In the present thesis, endocrine disrupting potential of two classes of halogenated pollutants studied within the CADASTER Project, i.e. brominated flame retardants (BFRs) and perfluorinated chemicals (PFCs), was investigated by means of non-animal testing methods. I*n vitro* data measuring specific endocrine modulating effects (e.g. estrogenic and androgenic receptor binding, interference with thyroxin transport and estradiol metabolism) were used to develop quantitative structure-activity relationships.

This chapter is organised in two sections. The first section summarizes the QSAR models that have been developed for different endpoints related to ED activity of BFRs, and highlights the potential use of *in silico* techniques, like QSARs, for the screening and prioritization of chemicals, starting from a limited amount of experimental information. The second section is focused on a specific

mechanism of endocrine disruption, i.e. the interference with thyroxin transport by T4-TTR competition, for which experimental evidences have been reported for both BFRs and PFCs. In this case the main issue stressed is the role of QSAR models to identify and quantify the most important structural features involved in a specific mechanism of action.

Section I. Endocrine disrupting potency of BFRs

This section deals with the QSAR models that have been developed for the characterization of the toxicological profile of brominated flame retardants (BFRs), which was based on their potential ability to interfere with the endocrine system. Several pathways and mechanisms of endocrine disruption were investigated, and are here briefly described.

Ahryl hydrocarbon Receptor (AhR) mediated pathway

The ahryl hydrocarbon receptor (AhR), also known as dioxin receptor (DR), is a cytosolic transcription factor, which is a member of the family of bHLH/PAS (basic Helix-Loop-Helix/Per-Arnt-Sim) proteins.

The physiological ligands of this receptor are unknown, but it binds several exogenous ligands such as synthetic aromatic hydrocarbons, halogenated aromatic compounds (HACs), including dioxins and dioxin-like compounds, and polycyclic aromatic hydrocarbons (PAHs).

AhR is normally inactive, bound to several chaperones[3]. The binding of the intracellular AhR with such exogenous ligands implies the dissociation of the chaperones, resulting in AhR translocation into the nucleus and heterodimerization with the nuclear protein ARNT (AhR nuclear translocator). The complex AhR-ligand-ARNT recognises specific regulatory sequences in DNA, called dioxin-responsive enhancer (DRE), leading to changes in gene transcription. These changes include the induction of the cytochrome P-450 isozyme CYP 1A1 (Okey *et al.,* 1994; Whitlock, 1993; Hu and Bunce, 1999), which is a member of the cytochrome P450 superfamily of enzymes (monooxygenases), and is involved in the metabolic activation of aromatic hydrocarbons[4]. CYP 1A1 induction can be assayed as 7-ethoxyresorufin-O-deethylase (EROD) activity.

The activation of AhR-mediated pathway is a critical toxicological mechanism for many HACs, including PBDEs. Particularly high binding affinity has been found for the 2,3,7,8-tetrachlorodibenzo-

[3] Chaperones, proteins assisting the folding of other proteins and with additional functions. The AhR cytosolic complex consists of a dimer of Hsp90 and a single molecule of XAP2 (immunophilin-like protein hepatitis B virus X-associated protein 2), which are responsible for the protection of the receptor from proteolysis, maintenance of the correct conformation for ligand binding and prevention of the premature translocation into the nucleus and binding to ARNT.

[4] http://www.ncbi.nlm.nih.gov/gene/1543.

p-dioxin (TCDD), which is commonly used as reference compound when analysing AhR binding and AhR-mediated signal transduction activation.

Estrogenic, androgenic and progestagenic activity

Estrogens, androgens and progestogens are steroid hormones responsible for the regulation of a number of functions related to reproduction, sexual development and behaviour, maintenance of sexual characteristics, as well as other physiological functions. Androgens (e.g. testosterone) are commonly considered as "male sex hormones" in vertebrates since they control the development and maintenance of male characteristics (i.e. male sex organs and male secondary sex characteristics), spermatogenesis and sperm production, as well as other functions, including inhibition of fat deposition and increasing of muscle mass (Singh *et al.,* 2006; Sinha-Hikim *et al.,* 2004). Estrogens (e.g. estradiol - E_2) and progestogens (i.e. progesterone) are considered as "female sex hormones", since they promote the development, maturation and maintenance of female primary and secondary sexual characteristics (e.g. uterus, ovaries and breasts), regulate reproductive cycles and control all the physiological and morphological changes occurring during pregnancy in order to support the gestation.

The actions of these hormones are mediated by specific receptors belonging to the hormone nuclear receptors superfamily, i.e. estrogen receptor (ER), androgen receptor (AR) and progesterone receptor (PR). The steroid hormones enter passively into the cells and bind to respective receptors. The hormone-receptor complex binds to specific DNA sequences, called hormone responsive elements, and regulate gene expression. Estrogens, androgens and progesterone can enter all cells, but their actions are dependent on the presence of the specific receptors (ER, AR and PR respectively) in the cell.

Experimental evidences show that several environmental pollutants are able to interact with these hormone receptors (HR) leading to the activation of the ER/AR/PR-mediated pathway (agonists) or preventing the binding of the natural hormones (antagonists) (Fang *et al.,* 2003; Kojima *et al.,* 2004; Colborn, 1995; Jensen *et al.,* 1995).

Several *in vitro* assays have been developed for the detection of chemicals with the potential to bind to hormone receptors, such as the reporter gene assays that measure HR binding-dependent transcriptional activity (Fang *et al.,* 2000; Murk *et al.,* 1996).

As described in the introduction (paragraph 3.1), an alternative mechanism of endocrine disruption is the interference with metabolic processes of hormones. It has been demonstrated, for example, that some HACs are able to affect the elimination rates of the hormone estradiol (E_2) by inhibiting the enzyme estradiol-sulfonyltransferase (E_2SULT), which is responsible for E_2 inactivation (Kester *et al.,* 2002).

Interference with thyroid hormones (TH)

The thyroid hormones, triiodothyronine (T3) and thyroxine (T4), are tyrosine-based hormones produced by the thyroid gland that are responsible for regulation of the energetic metabolism of the organism, including increase of the basal metabolic rate, promotion of growth and development, regulation of protein, fat, and carbohydrate metabolism.

Several *in vivo* and *in vitro* studies have shown that environmental pollutants, such as organohalogens, can interfere with the thyroid hormone system by affecting:

i) the thyroid gland function and regulation, with potential induction of adverse effects including thyroid hyperplasia, tumors or hypothyroidism;

ii) the thyroid hormone metabolism (e.g. sulfation, deiodiation and glucuronidation) and transport mechanisms (e.g. by binding to the plasma thyroid hormone transport protein transthyretin (TTR), thereby displacing the natural ligand T4);

iii) the binding of TH to respective TH receptors (TR) and activating, or preventing, the TR-mediated pathway (Brower *et al.,* 1998, Legler *et al.,* 2003).

Among the suspected EDCs, brominated flame retardants (BFRs) are ubiquitary pollutants that have been identified as potential endocrine disrupters. Experimental evidences show that BFRs are endocrine-active compounds with the potential to interfere with thyroid hormone homeostasis, as well as to interact with steroid receptors (e.g. estrogens, androgens) and aryl hydrocarbon receptors (dioxin-like activity) (Legler and Brouwer, 2003; Meerts *et al.,* 2000; Meerts *et al.,* 2001; Lilienthal *et al.,* 2006; Chen *et al.,* 2001; Peters *et al.,* 2006).

Despite of toxicological evidences and the increasing concentrations detected in both wildlife and humans, at present experimental data are available only for a limited number of BFRs, such as some PBDE congeners (e.g. BDE-47, BDE-99, BDE-100, BDE-183, BDE-209), TBBPA and HBCD. To fill the gap of experimental data, approaches based on quantitative structure-activity relationships (QSAR) can be applied to predict lacking information (e.g. for all 209 potential PBDE congeners, as well as hydroxylated and methoxylated PBDE metabolites, which are found to be ubiquitous in the environment and even more active than their parent compounds), as well as to screen and prioritize chemicals for experiments, with a consequent reduction of costs, time and of the number of tested animals.

In this thesis, regression and classification models were developed for different endpoints related to endocrine disruption potency. Results of these studies have been published and a detailed description and discussion of the proposed models can be found in **Publications I** (regression models) and **II** (classification models).

3.2 Methods

3.2.1 Modelled endpoints

Experimental data, measured *in vitro* for several endpoints related to endocrine disruption, were collected from literature (Chen *et al.,* 2001; Hamers *et al.,* 2006; Hamers *et al.,* 2008). The modelled endpoints were selected according to the number of experimental data available.

In the regression study the following endpoints were considered:

- AhR relative binding affinity (RBA), calculated as the ratio of EC_{50} measured for the reference compound TCDD and EC_{50} measured for the test compounds (Chen *et al.,* 2001);
- EROD (ethoxyresorufin-O-deethylase) induction potency ($EC_{50}EROD_{ind}$ μM) (Chen *et al.,* 2001);
- AhR agonism ($EC_{50}DR_{ag}$ μM) (Hamers *et al.,* 2006);
- Estrogen receptor (ER) agonism ($EC_{50}ER_{ag}$ μM) (Hamers *et al.,* 2006);
- Progesterone receptor (PR) antagonism ($IC_{50}PR_{ant}$ μM) (Hamers *et al.,* 2006);
- Thyroxine-transthyretin (T4-TTR) relative competing potency (T4-REP), calculated as the ratio of IC_{50} measured for the reference compound (T4) μM and IC_{50} of the test compound μM (Hamers *et al.,* 2006; Hamers *et al.,* 2008);
- Estradiol sulfotransferase (E_2SULT) relative inhibition (E_2SULT-REP), calculated as the IC_{50} measured for the reference compound PCP (pentachlorophenol) μM and IC_{50} for the test compound μM (Hamers *et al.,* 2006; Hamers *et al.,* 2008).

Only data with a specific value of EC_{50}, IC_{50} or relative activity were used for QSAR modelling. All the responses were converted into logarithmic units and, to obtain increasing trends of toxicity, the experimental values for the responses $EC_{50}EROD_{ind}$, $EC_{50}DR_{ag}$, $EC_{50}ER_{ag}$ and $IC_{50}PR_{ant}$ were transformed into the logarithm of the inverse μM concentrations.

Experimental data measured by Hamers and collaborators (Hamers *et al.,* 2006 and Hamers *et al.,* 2008) were additionally used for the development of classification models. Models were developed for the following endpoints:

- Aryl hydrocarbon (dioxin) Receptor agonism (DR_{ag}) and antagonism (DR_{ant});
- Estrogen Receptor agonism (ER_{ag}) and antagonism (ER_{ant});
- Androgen Receptor antagonism (AR_{ant});
- Progesterone Receptor antagonism (PR_{ant});
- T4-TTR Competing Potency ($T4\text{-}TTR_{comp}$);
- E_2SULT Inhibiting Potency (E_2SULT_{inh}).

$E(I)C_{50}$ values were converted into specific classes of ED potency. In this case, the information included in results reported as "no response" could be used for model development since these data were assigned to the class of "no ED potency".

The definition of the classes of activity was based on the classification criteria proposed by Hamers and collaborators (Hamers *et al.,* 2006) and are reported in Table 3.1.

Table 3.1. Classification criteria for ED potency of BFRs (proposed by Hamers *et al.,* 2006) and classes modelled for each end-point.

Hamers class	**Criteria**	**DR_{ag}, DR_{ant}, ER_{ag}, ER_{ant}, AR_{ant}, PR_{ant}**	**T4-TTR_{comp}, E_2SULT_{inh}**
No potency	Response <20% of control at 10 µM	Inactive (Class 1)	Inactive (Class 1)
Low potency	$E(I)C_{50}$ >10 µM & response >20% of control	Active (Class 2)	Moderately active (Class 2)
Moderate potency	1.0 µM < $E(I)C_{50}$ <10 µM	Active (Class 2)	Moderately active (Class 2)
High potency	0.1 µM < $E(I)C_{50}$ < 1.0 µM	Active (Class 2)	Very active (Class 3)
Very high potency	0.01 µM < $E(I)C_{50}$ < 0.1 µM	Active (Class 2)	Very active (Class 3)

As can be observed in the table reported above, for the endpoints DR_{ag}, DR_{ant}, ER_{ag}, ER_{ant}, AR_{ant} and PR_{ant}, whose experimental data were available for 24 BFRs, chemicals were assigned into two classes of ED potency: Class 1 ("inactive", i.e. no response was observed), and Class 2 ("active", i.e. any evidence of ED potency, from low to very high, was measured). For these endpoints, binary classification models were developed.

Three classes of ED potency were modelled for the endpoints T4-TTR_{comp} and E_2SULT_{inh}, for which a higher number of experimental data (n_{obj} = 29) were available: Class 1 ("inactive", i.e. no response was observed), Class 2 ("moderately active", i.e. low or moderate ED potency), and Class 3 ("very active", i.e. high to very high ED potency).

3.2.2 Datasets

The experimental datasets used for the development of the regression and classification models are the result of an extended literature search specifically focused on ED properties of PBDEs and BFRs. Taking into account the complexity of the endpoints considered in this study, it was decided to use only experimental data measured by one research group, in order to guarantee a better quality and homogeneity of the input data, which is a fundamental requirement for defining a robust structure-activity relationship. It is known that the use of heterogeneous experimental data from different sources and laboratories can affect the quality of models, by increasing the noise in the modelled response.

Experimental data were collected for several PBDE congeners, a few hydroxy-BDE (OH-PBDEs) and methoxy-BDE (CH_3O-PBDEs) metabolites, tetrabromobisphenol-A (TBBPA), tetrabromobisphenol-A-

bis(2,3)dibromopropyl ether (TBBPA-DBPE), 2,4,6-tribromophenol (246-TBP), and HBCD (Chen *et al.,* 2001; Hamers *et al.,* 2006; Hamers *et al.,* 2008).

Different datasets were used to develop the models depending on the experimental data available for each specific endpoint (**Appendix I** - Table A-1).

Hundreds of BFRs without experimental data were also considered in these studies for screening purposes. The complete studied dataset was finally composed of 243 chemicals, including all the 209 PBDE congeners, several hydroxylated and methoxylated PBDE metabolites (OH-PBDEs and CH_3O-PBDEs), brominated phenols, brominated bisphenol A compounds (TBBPA analogs) and other BFRs on the market, including three compounds used as alternatives to decaBDE, i.e. decabromodiphenylethane (DBDE), ethylene bistetrabromo phthalimide (EBTPI), 1,2-bis(2,4,6-tribromophenoxy) ethane (TBE).

The list of 243 studied BFRs is reported in **Appendix I** (Table A-2).

3.2.2.1 Training and Prediction sets

In order to perform the external validation of the models, the original data sets were preliminarily split into training sets, which were used to develop the model, and prediction sets, whose data were not involved in any phase of model development and were used only later for the validation of models. Dataset splitting was performed only when a sufficient number of experimental data were available ($n_{obj} > 16$). The splitting was carried out by random selection of prediction set objects (section 2.5.1). In particular:

a) regression models (quantitative responses): data were sorted in ascending order of activity and 1 every 2 chemicals were included in the prediction set;

b) classification models (qualitative data): 30% of prediction set chemicals were randomly sampled within each class of activity.

3.2.3 Molecular structures and descriptors

The 3D structures of 243 BFRs were drawn and minimized to their lowest energy conformation using the semi-empirical method AM1 in the HYPERCHEM program (ver. 7.03 for Windows, 2002), and were then used as input files for descriptor calculations.

The theoretical molecular descriptors, which encode for the information on the mono-, bi- and tri-dimensional structure of the chemicals, were computed by the software DRAGON (ver. 5.5). In a preliminary step, constant or near-constant values and descriptors with a high pair-wise correlation (>0.98) were excluded to reduce redundant and non useful information. At the end of this procedure a final set of nearly 700 molecular descriptors was used as input variables in the model development.

3.2.4 QSAR modelling and applicability domain

3.2.4.1 Regression models

Multiple linear regression models were developed by the ordinary least square (OLS) regression method using the software MOBY DIGS. All the possible combinations of variables (up to 2 molecular descriptors) were explored by applying the *All Subset Selection* method. A population consisting of the best 100 models was generated by maximising the cross-validated Q_{LOO}^2. Several statistical parameters were used to compare and validate the models for their goodness-of-fit, robustness and external predictive ability (e.g. R^2, Q^2_{LOO}, R^2_{ys}, Q^2_{EXT-F1}, RMSE). Details regarding QSAR development and validation are provided in **Publication I**.

For the selection of the best models, the analysis of the applicability domain was also performed. This included the identification of response outliers (compounds with cross-validated standardized residuals greater than 2.5 std) and of chemicals that were structurally very influential in determining the parameters of the models ($h > h^*$). The structural domain defined by the *hat* cut-off value (h^*) was used also to assess the reliability of predictions generated by the models for chemicals without experimental data. Predictions obtained for high leverage chemicals ($h > h^*$) were considered as model extrapolations and assessed as not reliable.

3.2.4.2 Classification models

Classification models were built by applying the *k*-nearest neighbour (*k*-NN) method (Sharaf *et al.,* 1986; Zheng and Tropsha, 2000). The *k*-NN method was applied to auto-scaled data, and the *a priori* probability of belonging to a class was set as proportional to the number of chemicals in the defined classes of ED potency. The predictive power of the model was checked for *k* values between 1 and 10.

Due to the small dimensions of the training sets, only models based on a maximum of two variables were considered. All the possible combinations of molecular descriptors were explored by maximizing the overall percentage of correct assignments (OA%) and the population of the best 100 models was analysed for each modelled endpoint.

Moreover, parameters sensitivity (Sn) and specificity (Sp) were calculated for all the studied endpoints. For the endpoints T4-TTR_{comp} and E_2SULT_{inh}, for which 3 classes of activity were defined according to Table 3.1, Sn was calculated after grouping into a single class of "active" chemicals belonging to Class 2 (moderately active) and 3 (very active).

In a precautionary approach, it was preferred to select models that minimized the number of false negative (i.e. active compounds classified as inactive).

To define the applicability domain of the classification models two approaches were combined. The first approach was based on the range of modelling descriptors defined by the training set chemicals. The second method was based on the calculation of Euclidean distance (from the structural similarity) and was performed by the software ToxMatch (ver. 1.06). For each class, compounds having a Euclidean distance higher than the training set were considered as structural outliers (beyond the AD of the model).

Predictions of compounds lying outside the structural domains of the proposed models were considered as model extrapolations.

3.3 QSAR models for ED potency of BFRs

Equations, modelling descriptors and statistical performance of the proposed models are summarised in Table A-3 (regression models) and Table A-4 (classification models) in **Appendix I**.

Despite the limited amount of experimental data available (datasets composed of 8 up to 29 chemicals), both regression and classification models developed in this study are characterized by high fitting power, internal robustness and external predictivity. As anticipated in section 3.2.2, all the models were applied to predict unknown activity of over 200 BFRs without experimental data for screening and prioritization purposes. The analysis of the applicability domain of the models, which is particularly relevant when dealing with small datasets, highlighted that the majority of the considered BFRs were included in the AD of the models, with a percentage of reliable predictions ranging from 75 to 100%.

Specific comments and discussion concerning mechanistic interpretation of modelling descriptors and applicability domain of the proposed models (i.e. interpolated and extrapolated predictions) are reported in **Publications I** and **II**.

Some general remarks resulting from the analysis of both regression and classification results are here reported.

Higher binding affinity with and consequent activation of the dioxin receptor AhR was predicted for low and medium brominated PBDE congeners (mono-penta PBDEs) with few or without *ortho* substituents. The presence of bromine atoms in *ortho* positions is highly relevant in 3D conformation of the diphenylethers since they induce a rotation (~90°) of one phenyl ring around the C-O-C central plane (Zhao *et al.,* 2008; Hu *et al.,* 2005). Conversely, PBDEs without Br substituents at the *ortho*-positions are more similar to dioxins and coplanar PCBs, for which high AhR binding affinity have been already demonstrated (Chen *et al.,* 2001).

Classification models developed for AR/PR antagonism highlighted a low to high activity for all BFRs, except those having all *meta*- and *para*-positions substituted (i.e. PBDEs with [3,3',4,4',5,5'] substitution pattern, TBBPA-derivates and deca-BDE alternatives). The fact that *meta* and *para* Br

substitutions are unfavorable for AR antagonism by PBDEs was also confirmed in the molecular docking study of Yang and co-workers (Yang *et al.,* 2009).

QSAR predictions for ER agonism and antagonisms are also in line with experimental observation (Meerts *et al.,* 2001; Hamers *et al.,* 2006). In particular, estrogenic activity through ER binding was predicted for lower-brominated PBDEs, while higher-brominated PBDEs were predicted as ER antagonists. Antiestrogenic activity was also predicted for OH-PBDEs. Similar results were obtained *in vitro* for OH-PCBs (Moore *et al.,* 1997). In this thesis estrogenic activity of BFRs was also investigated by modelling E_2SULT inhibiting potency. E_2SULT inhibition comparable to the reference compound PCP (pentachlorophenol) was predicted for OH-PBDEs with hydroxyl groups in *meta*- and *para*-positions, bromophenols and bisphenol-A analogues. These QSAR predictions are in fair agreement with experimental results from Kester *et al.* (2002), who suggested that E_2SULT inhibition could be favoured by planar hydroxylated polyhalogenated derivatives of PCBs, polychlorinated dibenzodioxins and furans (PCDD/Fs), PBDEs, tetrabromo- and tetrachlorobisphenol-A. Moderate E_2SULT inhibiting potency have been predicted for the majority of PBDEs. In particular, a trend of increasing activity from higher to lower brominated congeners was found.

Finally, the modelling of T4-TTR competing potency predicted an affinity toward TTR higher or comparable than the natural ligand T4 (IC_{50} = 0.055 µM (Hamers *et al.,* 2006)) for all the BFRs containing an aromatic hydroxyl group (i.e. OH-PBDEs, brominated phenols and bisphenols A compounds). In particular, the regression model highlighted higher T4-TTR competing potency for OH-BDEs with hydroxyl groups in *meta*- and *para*-positions in comparison to *ortho*-OH-PBDEs. The structural resemblance of many BFRs, and in particular OH-PBDE metabolites, to thyroxin is evident (Figure 3.1).

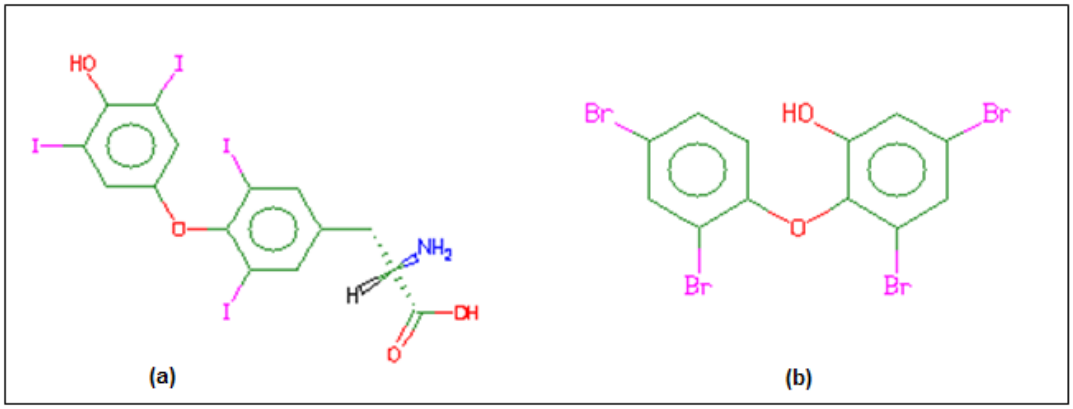


Figure 3.1. Structures of Thyroxine T4 (a) and 6-OH-BDE-47 (b).

This result is in agreement with experimental and *in-silico* observations reported in the literature (Hamers *et al.,* 2008; Meerts *et al.,* 2000; Harju *et al.,* 2007) that the phenolic group is required to increase TTR binding affinity because the natural ligand T4 is also hydroxylated.

QSAR predictions, AD information and values of modelling descriptors are available as supporting information in **Publication I** and **II**.

3.4 Screening and prioritization of PBDEs

As part of the CADASTER project activities, a prioritization of BFRs, and in particular of PBDEs, was required in order to optimize the experimental testing to perform within WP2.

For this purpose, a characterization of the toxicological profile of BFRs was performed within this thesis and combined to structural similarity analysis. This activity was finalized to suggest a limited subset of potentially most hazardous BFRs that, at the same time, covered a broad variation in the structural chemical domain.

The 243 BFRs considered in this study were screened for the potential hazard on the basis of the available information on their endocrine disruption potency, which included both experimental data and QSAR predictions generated by the here proposed regression and classification models. Only reliable predictions (i.e. prediction for chemicals included in the AD of all the models) were used for screening purposes. It is to note that all the proposed models were characterized by wide applicability domains, providing reliable (interpolated) predictions for over 85% of the studied BFRs for almost all the endpoints.

As it was commented in the previous paragraph, the modelled endpoints covered a wide range of ED mechanisms, each involving specific interactions between chemicals and the target protein complex (receptor, transporter or enzyme), and this prevented us from performing a ranking of BFRs according to their "general" endocrine disrupting activity. However, an attempt to prioritize BFRs for ED potency was done by analysing separately two different types of ED responses: i) dioxin-like activity, which included the endpoints AhR RBA, EROD induction and AhR agonism, and ii) other ED pathways, considering ER agonism, PR antagonism, T4-TTR competing potency and E_2SULT inhibition.

Within the two groups of ED activity, chemicals were ordered according to increasing activity (or separated by classes of ED potency). Chemicals showing the highest potency for all the considered endpoints were selected as potentially more hazardous.

Two separate lists of the most active BFRs for dioxin-like activity and for other ED pathways were then provided to the CADASTER partners responsible for the experimental testing (Table 3.2).

Table 3.2. Prioritized BFRs on the bases of potential ED potency (experimental and predicted by QSARs).

Dioxin-like activity					*Other ED pathways*			
BDE-002	BDE-037	BDE-082	BDE-170	BDE-196	BDE-001	BDE-049	BDE-140	BDE-200
BDE-003	BDE-039	BDE-085	BDE-171	BDE-198	BDE-002	BDE-050	BDE-142	BDE-201
BDE-011	BDE-041	BDE-096	BDE-173	BDE-199	BDE-004	BDE-084	BDE-145	BDE-203
BDE-012	BDE-042	BDE-108	BDE-177	BDE-206	BDE-005	BDE-086	BDE-170	BDE-204
BDE-013	BDE-054	BDE-119	BDE-178	4OH-BDE-42	BDE-008	BDE-088	BDE-171	BDE-207
BDE-015	BDE-055	BDE-122	BDE-181	2'OH-BDE-66	BDE-009	BDE-089	BDE-172	6OH-BDE-47
BDE-022	BDE-056	BDE-126	BDE-185	4-BP	BDE-018	BDE-091	BDE-174	6OH-BDE-99
BDE-028	BDE-060	BDE-127	BDE-190		BDE-040	BDE-102	BDE-180	6'OH-BDE-49
BDE-033	BDE-066	BDE-134	BDE-192		BDE-041	BDE-104	BDE-190	6OH-BDE-90
BDE-035	BDE-077	BDE-136	BDE-193		BDE-044	BDE-115	BDE-197	
BDE-036	BDE-079	BDE-142	BDE-195		BDE-048	BDE-133	BDE-198	

The prioritization based on potential hazard was then combined with a multivariate structural characterization, which was performed by the CADASTER Partner LnU.

Principal component analysis (PCA) was applied to theoretical molecular descriptors calculated for 209 PBDEs. In this case a subset of 30 PBDEs was first identified according to the environmental relevance (i.e., PBDE most commonly found in the environment, such as BDE-28, -47, -99, -100, -153, -154, -183, -209), and availability of chemicals for the experimental tests (as indicated by Partner PHI, responsible for testing). These chemicals are highlighted in red in Figure 3.2, which plots the first two principal components of the PCA analysis based on theoretical molecular descriptors.

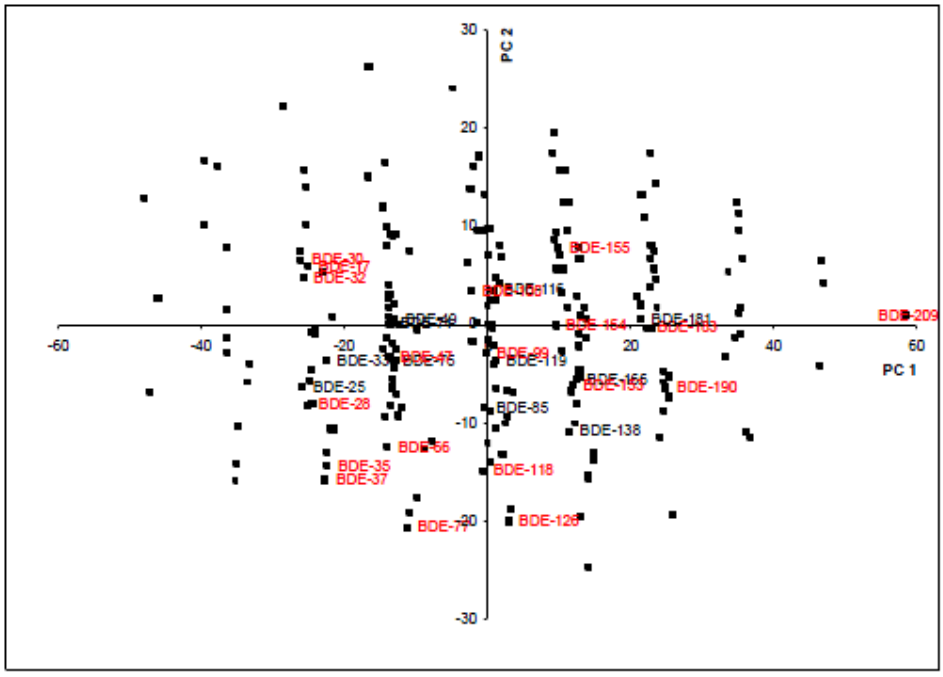


Figure 3.2. PCA analysis (plot of PC1 vs. PC2) of 209 PBDE congeners.

Among these compounds, the Partner LNU suggested 16 PBDE congeners characterized by a wider structural diversity, taking into account both the degree and the substitution pattern of bromine atoms (Table 3.3).

Table 3.3. List of priority compounds, with eventually alternatives, based on structural similarity.

No	BDE	No. Br	No. ortho
1	#17, #30 or #32	3	2
2	#28	3	1
3	#35 or #37	3	0
4	#47	4	2
5	#66	4	1
6	#77	4	0
7	#99	5	2
8	#100	5	3
9	#118	5	1
10	#126	5	0
11	#153	6	2
12	#154	6	3
13	#155	6	4
14	#183	7	3
15	#190	7	2
16	#209	10	4
17	#85	5	2
18	#138	6	2

After this experimental design based on toxicological and structural characterization, a final set of 12 PBDE congeners (Table 3.4) was selected by the Partner PHI for testing bioconcentration (BCF) and bioaccumulation (BAF) potential in sediment for the species *Tubifex tubifex*.

Table 3.4. PBDE congeners selected for experimental testing.

No	**Selected PBDEs**	
1	BDE-2	
2	BDE-26	
3	BDE-66	TBDE-71X
4	BDE-77	
5	BDE-99	
6	BDE-119	TBDE-71X
7	BDE-180	TBDE-79X
8	BDE-197	TBDE-79X
9	BDE-198	
10	BDE-203	TBDE-79X
11	BDE-204	
12	BDE-207	TBDE-79X, TBDE-83RX

Section II. T4-TTR competing potency of BFRs and PFCs

The second section of this chapter will be focused on the investigation of a specific mechanism through which some organic pollutants may interfere with the endocrine system, i.e. T4-TTR competing potency. As described in the previous section, T4-TTR competition consists in the ability of an exogenous chemical to compete with and displace the circulating thyroid hormone T4 from the binding sites of its transport protein transthyretin (TTR), with a consequent alteration of serum levels of T4.

In the cerebrospinal fluid TTR is the primary carrier of T4, whose levels are of crucial importance especially during the development of the central nervous system. Alteration of T4 levels in the sensitive phases of gestation and early life stages may seriously affect foetal and neonatal development (de Escobar *et al.,* 2004). Experimental evidences indicated that halogenated pollutants such as brominated flame retardants and perfluorinated compunds may compete with T4 for the binding to TTR (Weiss *et al.,* 2009; Meerts *et al.,* 2000; Hamers *et al.,* 2006; Gutshall *et al.,* 1989; Lau *et al.,* 2007; Chang *et al.,* 2008; Ucán-Marin *et al.,* 2010). TTR binding potency of these chemicals, in some cases even higher than the natural ligand T4, was explained by the structural resemblance of many BFRs (PBDEs and OH-BDEs in particular) to T4 (Hamers *et al.,* 2006) and by the "polar-hydrophobic" nature of PFCs, which determines high affinity for protein binding (Weiss *et al.,* 2009; Biffinger *et al.,* 2004).

In order to investigate the structural features involved in TTR binding affinity of these two classes of emerging pollutants, quantitative structure-activity relationships were developed specifically for BFRs and PFCs by applying different modelling approaches, i.e. regression and classification (**Publications I**, **II** and **III**). These "local" models were then applied to predict T4-TTR competing potency of a big set of BFRs and PFCs without experimental data.

Another interesting issue was to investigate whether structurally different chemicals like PFCs and BFRs could share the same mechanism of interaction with the carrier TTR. An experimental evidence of this assumption was found in a study by Peterson *et al.* (1998), where the authors analysed the interactions between TTR and the anti-inflammatory drug flufenamic acid (Flu) by X-ray crystallography. In Figure 3.3, the structure of Flu is reported and compared with the chemical structures of an hydroxy-PBDE (OH-BDE-47) and the perfluoroalkyl acid (PFOA).

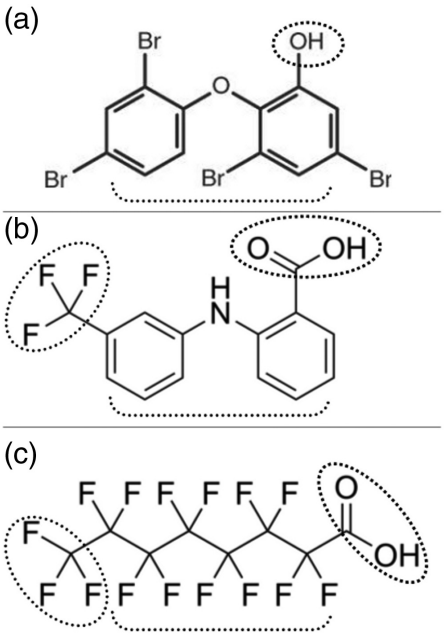


Figure 3.3. Chemical structures of an hydroxylated PBDE (a), Flu (b) and PFOA (c). The common structural features probably implied in the binding with TTR are highlighted (Figure from **Publication III**).

The study performed by Peterson and co-workers allowed to identify the specific binding sites between TTR and Flu responsible for the formation of the Flu-TTR binary complex (i.e. Flu interacts with residues of two adjacent TTR molecules). In particular, the molecular interaction involves the following structural fragments of Flu: i) the CF_3 substituent, which occupies the innermost halogen-binding pocket, interacting with Ser- 117, Thr-119, Leu-110, and Ala-108, ii) the biphenyl system, which interacts with a hydrophobic patch of the T4 binding site (between the residues Leu-17, Thr-106, Ala-108, Thr-119, and Val-121), and iii) the carboxylate group, which is placed at the entrance of the funnel-shaped binding pocket, forming electrostatic interactions with the side chains of the Lys-15 residues from opposing TTR subunits (Peterson *et al.,* 1998).

As can be observed in Figure 3.3, some of the structural features involved in Flu-TTR binding can be found both in BFRs (i.e. biphenyl system) and PFAAs (-CF_3 and -COOH as opposite terminal groups). Therefore, the Flu-TTR mechanism of binding supports the idea of a similar interaction of PFCs and BFRs with TTR. It was then decided to explore the possibility of defining a quantitative relationship between the structures of BFRs and PFCs and their T4-TTR competing potency, and to develop a single "common" QSAR model for these two classes of halogenated chemicals (**Publication IV**).

3.5 Methods

3.5.1 Datasets

Training sets used for the development of the local models for BFRs are reported in Table A-1 and consists of 17 data of T4-TTR relative competition (T4-REP values), used for the regression model, and 29 data of T4-TTR competing potency, including 12 inactive, 9 moderately active and 8 very active chemicals (Hamers *et al.,* 2006; Hamers *et al.,* 2008), which were used for the classification model.

Experimental data for T4-TTR competing potency of PFCs were collected from literature (Weiss *et al.,* 2009) and were used for the development of classification models. The experimental data set consisted of 24 PFCs with different carbon chain length, fluorination degree and functional groups, including several perfluorinated alkyl acids (PFAA), sulfonates (PFAS), sulfonamides and telomer alcohols (FTOH). The measured IC_{50} values were converted into two classes of T4-TTR competing potency according to the classification criteria proposed by Hamers and co-workers (Table 3.1), which were the same criteria used for the class assignment of BFRs. In particular, all the compounds showing any evidence of T4-TTR competing potency (from low to very high) were assigned to the class '1' (active); the remaining inactive compounds were assigned to the class '2'.
For modelling purposes, the five PFAS were converted into the respective sulfonic acids (PFAS(A)) and not included in the training set used to develop the QSAR models. The predictions obtained for these compounds in acidic form were then compared with the experimental data measured for the corresponding salts.

Finally, experimental data available for BFRs and PFCs were combined in a common dataset used for the development of regression models (32 compounds, including 15 PFCs and 17 BFRs) and classification models (53 compounds, including 24 PFCs and 29 BFRs). The endpoint modelled by regression was the logarithm of the relative binding potency toward T4 (logT4-REP), while for classification purposes three classes of T4-TTR competing potency were considered, i.e. no activity (Class 1), moderate activity (Class 2) and high activity (Class 3), according to the same classification criteria previously used (Table 3.1). The five PFAS(A) were included in the common dataset, since the results obtained from the local model developed for PFCs suggested that the salt-to-acid conversion did not influence the modelling results.

The different experimental datasets used in this study are reported in Table 3.5.

Table 3.5. Datasets used for the development of the local models for BFRs (regression and classification) and PFCs (classification), and for the common models (regression and classification).

ID	Name	Name extended	Log T4-REP	Class T4-TTR
19	BDE-019	2,2',6-triBDE	--	1
28	BDE-028	2,4,4'-triBDE	--	1
38	BDE-038	3,4,5-triBDE	-2.66	2
39	BDE-039	3,4',5-triBDE	--	1
47	BDE-047	2,2',4,4'-tetraBDE	-2.66	2
49	BDE-049	2,2',4,5'-tetraBDE	-2.66	2
79	BDE-079	3,3',4,5'-tetraBDE	--	1
99	BDE-099	2,2',4,4',5-pentaBDE	--	1
100	BDE-100	2,2',4,4',6-pentaBDE	--	1
127	BDE-127	3,3',4,5,5'-pentaBDE	-2.6	2
153	BDE-153	2,2',4,4',5,5'-hexaBDE	--	1
155	BDE-155	2,2',4,4',6,6'-hexaBDE	--	1
169	BDE-169	3,3',4,4',5,5'-hexaBDE	-2.66	2
181	BDE-181	2,2',3,4,4',5,6-heptaBDE	-2.1	2
183	BDE-183	2,2',3,4,4',5',6-heptaBDE	--	1
185	BDE-185	2,2',3,4,5,5',6-heptaBDE	-2.13	2
190	BDE-190	2,3,3',4,4',5,6-heptaBDE	-2.21	2
206	BDE-206	2,2',3,3',4,4',5,5',6-nonaBDE	--	1
209	BDE-209	2,2',3,3',4,4',5,5',6,6'-decaBDE	--	1
214	TBBPA	3,3',5,5'-tetrabromobisphenol-A	0.25	3
215	246-TBP	2,4,6-tribromophenol	1.06	3
216	6OH-BDE-47	6-OH-2,2',4,4'-tetraBDE	-0.51	3
221	HBCD	hexabromocyclododecane γ	--	1
222	TBBPA-DBPE	tetrabromobisphenol-A-bis(2,3)dibromopropyl ether	-1.98	2
224	4-OH-BDE-42	4-OH-2,2',3,4'-tetraBDE	0.54	3
225	3-OH-BDE-47	3-OH-2,2',4,4'-tetraBDE	0.6	3
226	5-OH-BDE-47	5-OH-2,2',4,4'-tetraBDE	0.48	3
227	4'-OH-BDE-49	4'-OH-2,2',4,5'-tetraBDE	0.54	3
228	2'-OH-BDE-66	2'-OH-2,3',4,4'-tetraBDE	-0.19	3
244	PFHxA	Perfluorohexanoic acid	-2.15	2
245	PFDoA	Perfluorododecanoic acid	-3	2
246	PFOA	Perfluorooctanoic acid	-1.19	3
247	PFDcA	Perfluorodecanoic acid	-2.15	2
248	PFHxS(A)*	Perfluorohexane sulfonic acid	-1.07	3
249	PFBA	Perfluorobutyric acid	--	1
250	PFBS(A)*	Nonafluorobutane sulfonic acid	-2.52	2
251	PFHpA	Perfluoroheptanoic acid	-1.41	2
252	PFNA	Perfluorononanoic acid	-1.66	2
253	PFTdA	Perfluorotetradecanoic acid	-2.7	2
254	FTOH(6:2)	2-Perfluorohexyl ethanol	--	1
255	FTOH(8:2)	2-Perfluorooctyl ethanol	--	1
256	FOSA	Perfluorooctane sulfonamide	-2	2
257	7H-PFHpA	7H-Perfluoroheptanoic acid	-2.15	2
258	N-EtFOSE	2-(N-ethylperfluoro-1-octane sulfonamido) ethanol	--	1

ID	Name	Name extended	Log T4-REP	Class T4-TTR
259	PFOS(A)*	Perfluorooctane sulfonic acid	-1.19	3
260	PFUnA	Perfluoroundecanoic acid	-2.52	2
261	N-EtFOSA	N-ethyl perfluorooctane sulfonamide	--	1
262	N-MeFOSE	2-(N-methylperfluoro-1-octane sulfonamido) ethanol	--	1
263	N-MeFOSA	N-methyl perfluorooctane sulfonamide	--	1
264	FTUA(6:2)	2H-Perfluoro-2-octenoic acid (6:2)	-2.15	2
265	N,N-Me2FOSA	N,N-dimethyl perfluorooctane sulfonamide	--	1
266	L-PFDS(A)*	Perfluorodecane sulfonic acid	--	1
267	L-PFOSi(A)*	Perfluorooctane sulfinic acid	-1.46	2
		N_{TR} (BFRs)	**17**	**29**
		N_{TR} (PFCs)	**15**	**24**
		N_{TR} (Tot)	**32**	**53**

(*) Perfluoroalkylsulfonic acids (PFAS(A)) not included in the training set of the classification model for PFCs.

An additional set of over 200 BFRs (**Appendix I,** Table A-2) and 33 PFCs (poly- and per-fluorinated compounds with different carbon chain lengths and functional groups, e.g. carboxylates, sulfonates, sulfonamides, alcohols, acrylates, etc. – **Appendix I,** Table A-5) of possible environmental concern with unknown T4-TTR competing potency was considered for screening purposes. Many of these chemicals are of interest for REACH Regulation and the evaluation of their potential toxicity has been one of the topics of the CADASTER Project.

3.5.2 Molecular structures and descriptors

The structures of the studied PFCs and BFRs were created and energetically optimized by the semi-empirical method AM1 using the HYPERCHEM program. The software DRAGON was then used for the calculation of mono-, bi- and tri-dimensional molecular descriptors. Constant, near-constant and pair-wise correlated descriptors were excluded and the remaining descriptors were used as input variables for the further variable selection procedure, which was necessary to select the best QSAR models based on a limited amount of variables (up to 2 or 3 variables).

3.5.3 QSAR modelling and applicability domain

The modelling approach and methods used for the development of regression and classification models for T4-TTR competing potency was the same described in section 3.2.4. All the datasets were first split into training and prediction sets, and once verified the predictive ability of the selected variables (Split models), the models were recalibrated using all the available experimental data (Full models).

The selection of the modelling variables for the local models of BFRs and PFCs was performed by applying the *All Subset Selection* method, which explores all the combination of molecular descriptors up to two variables. The availability of a larger training set for the development of the

common models for BFRs and PFCs (32 data for the regression model and 53 data for the classification model), allowed us to train the models up to three variables. The GA-based variable selection procedure was started after the *All Subset Selection* to explore the combinations of descriptors up to three variables in the models.

The applicability domain of both regression and classification models was defined by the leverage approach.

3.6 Local models for T4-TTR competing potency of BFRs

The local QSAR models developed for BFRs have been already described in section 3.3. They consist in a multiple linear regression based on two variables, i.e. qpmax and MATS6v, and 17 BFRs in the training set, and a *k*-NN classification model developed on a training set of 29 BFRs and based on the molecular descriptors nArOH and DISPe. Equation and statistical performances of the two models are reported in Tables A-3 and A-4.

Consistent predictions were found when the regression and classification models were applied to predict T4-TTR competing potency of 243 BFRs. Both QSARs identified as very active all the hydroxylated PBDE metabolites and hydroxylated BFRs, such as bromophenols and TBBPA analogs. The higher TTR binding affinity of hydroxylated BFRs, even exceeding that of the natural ligand T4, has already been documented in the literature (Meerts *et al.,* 2000; Hamers *et al.,* 2006; Hamers *et al.,* 2008) and can be explained by their structural resemblance to the hormone T4. Moderate T4-TTR competing potency was predicted for most of PBDEs. However, different structural information was identified as relevant for increasing T4-TTR competition by regression and classification models. In particular, the regression model predicted higher activity for tetra to octa-PBDEs with ortho-2,2',6,6'- or 2,2',6-bromines, mainly in absence of 3,3' substitutions. The majority of PBDE congeners classified as moderately active by the *k*-NN model were characterized by an asymmetric distribution of Br substituents in the phenyl rings (information encoded by the modelling descriptor DISPe).

3.7 Local models for T4-TTR competing potency of PFCs

To predict T4-TTR competing potency of PFCs, four best classification models were identified and a prediction by consensus was proposed (**Publication III**). The four models were first developed on a training set of 10 chemicals (Split models) and externally validated on the remaining chemicals included in the prediction set (9 PFCs). Despite the limited dimension of the training set, the developed QSARs showed high classification accuracy for both training and prediction set chemicals (overall accuracy > 90%). Once the predictive ability of the models was verified, the four classification models were newly calibrated using all the 19 experimental data (Full models). Molecular descriptors and classification performance of Split and Full models are reported in Table 3.6, where can be

observed that the use of all the experimental information further improved classification accuracy (100% correct assignments for all the four full models).

Table 3.6. Descriptors and parameters of the best four local classification models for PFCs.

ID	Descriptors	*k*	set	n	Sn	Sp	OA%
M1	AMW, HATS6m	1	Training set	10	1	1	100
			Prediction set	9	1	1	100
			Full dataset	19	1	1	100
M2	nH, HATS6m	1	Training set	10	1	0.75	90
			Prediction set	9	1	1	100
			Full dataset	19	1	1	100
M3	nH, F06[C-O]	1	Training set	10	1	1	100
			Prediction set	9	1	0.75	90
			Full dataset	19	1	1	100
M4	T(F..F), HATS6m	1	Training set	10	0.83	1	90
			Prediction set	9	1	1	100
			Full dataset	19	1	1	100

In this case, the proposal to apply the consensus approach was not to improve the intrinsic statistical performances of the models, rather to cover a wider structural domain of applicability, provided by the five molecular descriptors AMW, HATS6m, nH, F06[C-O], and T(F..F), and possibly to guarantee predictivity toward new compounds, after a proper check of the AD.

This assumption is supported by the fact that the five modelling descriptors encode for different structural features that have been experimentally identified to be relevant for the interaction with TTR (Weiss *et al,* 2009). In particular, AMW is a constitutional descriptors representing the average molecular weight and, therefore, is an index of the molecular dimension. The fingerprint F06[C-O] counts the frequency of C-O at topological distance 6 and is related to both the length of the carbon chain and to the number of oxygen atoms. nH counts the number of hydrogen atoms and, in the studied dataset, was able to discriminate poly- from per-fluorinated compounds as well as the chemicals with different terminal functional groups. More precisely, low nH values were calculated for PFAA and PFAS (nH = 1-2) and high values for FTOHs, sulfonamides and sulfonamido-etOH (nH = 4-10). In the study performed by Weiss and co-workers, acidic PFAA and PFAS (pKa < 3), which were dissociated in the test system (pH = 8), showed a much higher TTR binding potency than non acidic PFCs, such as FTOHs and perfluoroalkyl sulfonamides (non dissociated in the test system). Therefore, the type of functional end group, which determines the anionic state of PFCs (i.e. COO^- and SO_3^-), is a relevant structural feature influencing the affinity to TTR. This stresses the importance of the theoretical descriptor nH in modelling T4-TTR competing potency.

Another simple, but relevant descriptor is T(F..F), which summarizes the topological distances between F atoms in the molecule; T(F..F) is related to the fluorination degree, which is another factor known to change the affinity to TTR.

The only tri-dimensional descriptor selected by GA was HATS6m (leverage weighted autocorrelation of lag 6), a GETAWAY descriptor weighted by atomic masses (Consonni *et al.,* 2002). Beyond the specific structural meaning of this 3D descriptor, a simple interpretation was provided by relating its values with molecular weight (MW) and carbon chain length (nC) of the studied compounds. In particular, a parabolic-like relationship was found for PFAA and PFAS between HATS6m and MW (Figure 1 of **Publication III**). For these two groups of PFCs, it was verified that HATS6m has the largest values for compounds with a MW in the range of 300–500 g/mol and an intermediate carbon chain length (i.e. 6<nC<11 for PFAA and 5<nC<8 for PFAS). Additionally, the comparison of HATS6m values with the degree of T4-TTR competing potency (Figure 3.4) highlighted that increasing values of the descriptor corresponded to an increase of T4-TTR competing potency. Therefore, also the structural information encoded by HATS6m is crucial for classifying active and inactive chemicals, and provide an additional information concerning the degree of TTR affinity.

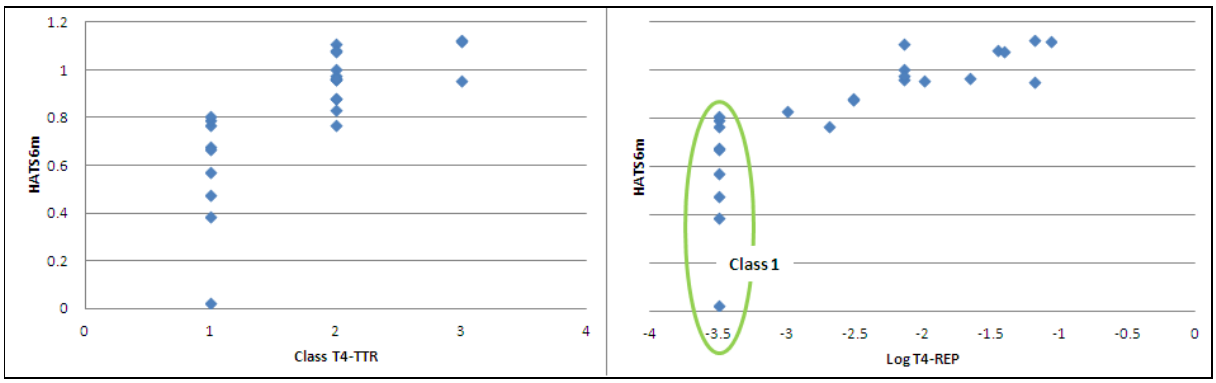


Figure 3.4. Relationship between HATS6m values and degree of T4-TTR competing potency.

Overall, the modelling molecular descriptors, selected from among a large population of input descriptors, provide integrated information on the molecular carbon chain length, fluorination degree and type of functional end group of PFCs, thus encoding exactly the same structural features that were experimentally found to be relevant for TTR binding (Weiss *et al.,* 2009).

Consensus predictions based on the four Full models were generated for the five PFAS, whose activity was measured for the respective salts, and the 33 additional PFCs lacking experimental data. The class assignment by consensus was based on the predictions obtained by the majority of models (at least three models) or, in case of opposite predictions (i.e. 2 models predicting as 'active' and the other 2 predicting as 'inactive'), the most precautionary prediction (i.e. 'active') was assigned.

All the PFAS were classified as active, probably overestimating the activity of L-PFDS(A), which is the perfluoroalkyl sulfonic acid with the longer carbon chain (nC=10), whose salt was measured as inactive.

Among the 33 PFCs screened in this study for their T4-TTR competing potency, all the PFAA with a carbon chain from 6 to 15 atoms, all the PFAS, perfluoropentane- and perfluorohexanesulfonyl fluorides, perfluorooctanamide and 1,1,2,2-tetrahydroperfluorohexyl iodide were predicted as active. the fluorotelomer alcohols were classified as inactive, as it was also reported in literature (Weiss *et al.,* 2009). While the presence of hydroxyl groups attached to an aromatic ring system play a key role in enhancing TTR affinity, as it is the case of the natural ligand T4 or other potent TTR-binding compounds (e.g. OH-PBDEs and bromo-phenols), the hydroxyl group itself is not essential for binding to TTR (Weiss *et al.,* 2009). Other compounds predicted as inactive were the perfluoroalkyl sulfonamides, polyfluorinated chemicals containing one or more aromatic rings and PFCs with terminal groups, such as cyano or acrylates.

QSAR predictions and values of the five modelling descriptors for all the 57 studied PFCs are available as Supplementary Material of **Publication III**.

3.8 Global models for T4-TTR competing potency of BFRs and PFCs

Two different modelling approaches, i.e. MLR regression and classification by *k*-NN, have been applied for the development of QSAR models able to quantitatively define the structural features responsible for the interaction of BFRs and PFCs with TTR.

The best models for the prediction of T4-TTR competing potency have been proposed in **Publication IV** and are reported in the following tables (Table 3.7 and Table 3.8).

Table 3.7. Molecular descriptors and statistical performances of the classification model proposed for T4-TTR competing potency.

Modelling Descriptors	***Model***	***set***	***n***	***k***	***Sn%***	***Sp%***	***OA%***
nArOH, F03[Br-Br], HATS6m	Split Model	Training set	37	5	91	79	82
		Prediction set	16	5	89	86	81
	Full Model	Training set	53	5	94	86	87

Table 3.8. Molecular descriptors and statistical performances of the regression model proposed for T4-REP.

Modelling Descriptors	***Model***	***set***	***n***	***R^2%***	***Q^2_{loo}%***	***RMSE***	***Q^2_{ext}%***	***CCC%***
R5u, F07[C-O], nArOH	Split Model	Training set	23	89	81	0.42	--	94
		Prediction set	9	--	--	0.34	88-93	95
	Full Model	Training set	32	89	84	--	--	94

All the models were first developed on a reduced training set (Split Models) in order to verify the predictive ability of the modelling variables on the prediction set, and then calibrated using all the available experimental data (Full Models). As it can be observed in Tables 3.7-8, the statistical parameters calculated to verify internal robustness and external predictivity (i.e. classification accuracy, Sn and Sp for classification and R^2 and Q^2 criteria for regression) have values always higher than 80%, indices of stable and predictive models. Details regarding model validation, interpretation of descriptors and applicability domains, in addition to the proposed MLR regression equation, are reported in **Publication IV**.

It is important to highlight that some of the modelling variables selected by the genetic algorithm, among hundreds of molecular descriptors, encoded for the same structural features recognized in literature as relevant for the interaction with TTR (Hamers *et al.,* 2008; Harju *et al.,* 2007; Peterson *et al.,* 1998). Among these:

i) nArOH (selected both in the classification and regression model, and already included in the local model for BFRs), which counts the number of aromatic hydroxyl groups, that are known to increase the TTR-binding capacity of BFRs;

ii) F07[C-O], which counts the frequencies of C-O at topological distance 7, and takes into account the presence of phenolic OH in BFRs as well as carboxylic and sulphonic acid groups attached to carbon chain at a specific distance (07), which characterises the most active PFCs;

iii) HATS6m, which is a 3D descriptor (already selected as relevant in the local models for PFCs) encoding for atomic heterogeneity, molecular dimension and shape. As was already commented in the previous paragraph, HATS6m is able to recognize the most active PFCs, taking into account the presence of the -COOH/SO_3H groups, the carbon chain length (i.e. 6<nC<11 for PFAA and 5<nC<8 for PFAS) and molecular weight (MW range 300-500 g/mol).

Additionally, the 3D GATAWAY descriptor R5u was particularly important in modelling T4-REP (regression model) of BFRs and PFCs, since was able to distinguish among active and inactive PFCs and BFRs in the same region of MW (330-550 g/mol).

3.9 Conclusions

In the present thesis, endocrine disruption potential of two classes of emerging halogenated pollutants, i.e. BFRs and PFCs, was investigated by quantitative structure-activity relationships.

Simple MLR regression (**Publications I**) and *k*-NN classification models (**Publications II**) have been developed specifically for BFRs for several endpoints related to ED potency, including binding affinity with Ah receptor and induction of AhR-mediated pathway (i.e. "dioxin-like potential"), antiandrogenic and antiprogestagenic properties, estrogenic activity by interaction with ER and inhibition of E2 metabolism, and interference with T4 transport by binding with TTR. The analysis of modelling molecular descriptors was useful to highlight some structural features and important structural alerts that could increase ED activity, such as the presence of Br atoms in *meta/para*-positions (with unoccupied *ortho*-positions), inducing dioxin-like activity, and the presence of aromatic hydroxyl groups, that greatly increased TTR affinity, E_2SULT inhibition as well as antiestrogenic activity. The presence of aromatic OH- group is a known structural alert for ED potency (Liu *et al.,* 2007; Roncaglioni *et al.,* 2008; Li and Gramatica, 2010). All the proposed models were applied for the prediction of ED potential for over 200 BFRs (including three alternative compounds to the banned deca-BDE) without experimental data. The screening of BFRs allowed to prioritize the most hazardous chemicals (on the basis of ED potency profile), that have been suggested to other partners involved in the CADASTER Project in order to optimize the experimental testing.

Particular attention was focused on the endpoint T4-TTR competing potency, for which experimental homogeneous data (i.e. measured by the same working group using the same *in vitro* assay) were available for both BFRs and PFCs. For this endpoint, local classification models were developed for PFCs and, also in this case, were applied to screen 33 PFCs without experimental data (**Publication III**). The best molecular descriptors selected for modelling this activity encode for the key structural features involved in TTR binding affinity: the simultaneous presence of an alkyl chain of specific length (between 6–10 atoms), and of a trifluoromethyl and a carboxylic (or sulfonic) opposite terminal groups. This finds confirmation in the experimental evidences reported in literature.

Additionally, robust QSAR models (MLR regression and *k*-NN classification) were developed including BFRs and PFCs in the training set (**Publication IV**). The development of a single QSAR for two classes of structurally different chemicals was performed in order to explore the possibility to identify the common structural features responsible for the binding of PFCs and BFRs to the same target (TTR). The molecular descriptors selected in the proposed models by GA procedure were found to be consistent with those selected in the local models for BFRs and PFCs, and encoded for the structural features involved in TTR binding highlighted by experimental studies: i) the presence of phenolic OH

in BFRs, ii) the presence of carboxylic and sulphonic acid groups attached to a carbon chain of specific length, resembling the length of the diphenyl ether (between 6-9 atoms for the most active PFCs), iii) the molecular dimension (MW up to 550 g/mol for both PFCs and BFRs), and iv) degree of bromination of non hydroxylated PBDEs.

All the proposed models were developed in agreement with the OECD principles for acceptability of QSAR predictions in regulation. Particular attention was paid to model validation and applicability domain, which are basic aspects that should be evaluated when proposing robust and valid QSARs, especially when models are based on limited amount of experimental data.

The complete list of training set chemicals, molecular descriptors, regression equations (of MLR models) and *k* values (for classification models) are provided in the publications here mentioned. This information allows for a transparent and feasible application of the proposed models. QSAR predictions for chemicals lacking experimental data as well as the information on reliability of the provided prediction (i.e. interpolated and extrapolated predictions) are also made available in the publications.

The present study demonstrates how the QSAR approach is able to extract useful information also from limited amount of experimental data. QSAR predictions were here used to screen many chemicals without experimental data, to identify those compounds with the highest concern for ED-related activity, and to perform experimental design in an optimized testing strategy.

Chapter 4

QSAR Modeling of Aquatic Toxicity of Triazoles and Benzotriazoles

4.1 Introduction

Under the current international regulatory system dealing with chemical substances (e.g., REACH, Biocides Regulation, Water Framework Directive, Cosmetics Directive), the assessment of aquatic toxicity is among the basic requirements for environmental risk assessments of chemicals.

Aquatic toxicity can be defined as the potential harm of a substance to living organisms in the aquatic environment. Defining an acceptable level of protection of the aquatic ecosystem that guarantees the protection of all the species and the functioning of the ecosystem itself is highly complex. Because of the impossibility to identify and test all the most sensitive species for all chemicals, a simplification and standardization of toxicity testing is needed. For this reason, to assess the aquatic toxicity only a very limited number of species are tested, namely an algal species, a crustacean species and a fish. These organisms cover the three key trophic levels of the aquatic ecosystem and are considered as surrogate for all aquatic organisms.

In routine toxicity testing, acute and chronic tests are performed, where the toxicity of a chemical is mainly measured through mortality, decreased growth rate and lowered reproductive capacity. In acute toxicity testing, test species are exposed to increasing concentration of a chemical for a relatively short period of time (in relation to the life cycle of the organism); the aim of these tests is to determine the concentration of the chemical that will elicit a specific response, such as mortality or any other measurable adverse effect. Mortality is expressed as the median lethal concentration (LC_{50}), which is the estimated concentration of the test material that will kill 50% of the test organisms in a predetermined period of time. Similarly, median effect concentrations (EC_{50}) can be calculated for any specified effect. In chronic toxicity tests, effects are studied over prolonged periods of exposure, often over entire life cycles and usually the endpoints are primarily sub-lethal (such as growth) or measurements of reproductive output. Sub-chronic studies are of longer duration than acute exposure but generally do not exceed a period equivalent to one-third of the time taken for a species to reach sexual maturity (van Leewuen and Vermeire, 2007).

In order to find application in regulatory field, testing methods for aquatic toxicity have been harmonized by the Organisation for Economic Cooperation and Development (OECD), which provided specific guidelines for different endpoints. A brief survey of the testing methods most commonly used for the assessment of aquatic toxicity is following and it's supported by Table 4.1 summarising the relative OECD technical guidelines.

The standard acute toxicity test in fish is 96-hour LC_{50} test (OECD 203), which is the median lethal concentration measured after a 96 hours exposure of the tested organisms. Test species are usually juveniles of various fish species, e.g. zebrafish (*Brachydanio rerio*), fathead minnow (*Pimephales promelas*) or rainbow trout (*Oncorhynchus mykiss*). Chronic fish tests are normally performed using

eggs, embryos, or juveniles and last from 7 to more than 200 days. Test endpoints include hatching success, growth, spawning success, and survival.

Acute and chronic tests are also conducted using crustacean, such as *Daphnia magna*, *Daphnia pulex* or any other suitable *Daphnia* species less than 24 h old. Acute toxicity test is based on the endpoint 48-hour EC_{50}, that is the concentration causing an immobilization in 50% of the test organisms (OECD 202). Longer term testing through maturation and reproduction is used to assess chronic toxic effects (OECD 202, OECD 211). The chronic testing endpoints include time to first brood, number of offspring produced per female, growth, and survival.

Testing procedures studying the toxic effects of chemicals on aquatic plants has centred on unicellular green algae (e.g. *Pseudokirchneriella subcapitata*), diatoms (*Navicula pelliculosa*) or cyanobacteria (*Anabaena flos-aque* or *Synechococcus leopoliensis*). The endpoints generally measured in phytotoxicity studies are photosynthesis and population growth. Because of the short life cycles, both acute and chronic endpoints can be obtained from these kinds of tests. A standardized test typically used to determine an acute EC50 is the algal growth inhibition test (OECD 201).

Table 4.1. List of some test methods, with relative OECD technical guidelines (TG), available for the assessment of aquatic toxicity.

OECD TG	**Test**
OECD 201	Algae, Growth Inhibition Test
OECD 202	Daphnia sp., Acute Immobilisation Test and Reproduction Test
OECD 211	*Daphnia magna*, Reproduction Test
OECD 203	Fish, Acute Toxicity Test
OECD 204	Fish, Prolonged Toxicity Test: 14-Day Study
OECD 210	Fish, Early-Life Stage Toxicity Test
OECD 212	Fish, Short-term Toxicity Test on Embryo and Sac-Fry Stages
OECD 215	Fish, Juvenile Growth Test

In environmental risk assessment of chemicals, acute toxicity tests may have two main applications. The first application consists in providing the basic set of acute toxicity data for the three trophic levels (i.e. algae, daphnids and fish). By applying specific assessment factors, these data can be used for the estimation of PNECs of a specific chemical (Chapter 1, section 1.1.2). Under REACH regulation, for example, acute toxicity data for algae, daphnids and fish are among the basic information required for the chemical safety assessment (CSA) and, therefore, are essential for the registration of chemicals (Chapter 1, section 1.1.3).

The second important application is for toxicological screening and classification of chemicals. International classification criteria based on aquatic toxicity tests have been described in Annex VI of Directive 67/548/EEC[1] (European Commission, 1991). These criteria, reported in Table 4.2, were applied to classify chemicals as dangerous for the aquatic environment and to assign specific risk phrases.

Table 4.2. Criteria for classification of chemicals for aquatic environmental hazard (Annex VI – Directive 67/548/EEC).

Indication of danger and risk phrases	**Classification criteria**
Very toxic to aquatic organisms (R50) and May cause long-term adverse effects in the aquatic environment (R53)	$E(L)C_{50}$[a] < 1 mg/L and the substance is not readily degradable or $\log K_{ow}$ >3
Very toxic to aquatic organisms (R50)	$E(L)C_{50}$[a] < 1 mg/L
Toxic to aquatic organisms (R51) and May cause long-term adverse effects in the aquatic environment (R53)	$E(L)C_{50}$[a] < 10 mg/L and the substance is not readily degradable or $\log K_{ow}$ >3
Harmful to aquatic organisms (R52) and May cause long-term adverse effects in the aquatic environment (R53)	$E(L)C_{50}$[a] < 100 mg/L and the substance is not readily degradable

[a] 96h LC_{50} for fish, 48h EC_{50} for daphnids and 72h EC_{50} for algae.

This Directive has been amended several times and will be definitively repealed by Regulation (EC) No 1272/2008 from 1st June 2015, by introducing the Globally Harmonised System of Classification and Labelling of Chemicals (GHS). Under this regulation, acute and chronic aquatic toxicity data, in conjunction to information on bioaccumulation and degradation potential, are used for the classification of chemicals for aquatic environmental hazard. In particular, the new classification system consists of one acute classification category ("Acute Category 1"), which is defined on the basis of acute aquatic toxicity data only (EC_{50} or LC_{50}), and three chronic classification categories, whose criteria combine acute aquatic toxicity data and environmental fate data (degradability and bioaccumulation data). The system also introduces a "safety net" classification ("Chronic Category 4") for use when the data available do not allow classification under the formal criteria but there are nevertheless some grounds for concern. Classification categories are summarised in Table 4.3.

[1] Directive 67/548/EEC of 27 June 1967 on the approximation of laws, regulations and administrative provisions relating to the classification, packaging and labelling of dangerous substances.

Table 4.3. Criteria for classification of chemicals for aquatic environmental hazard (Regulation (EC) No 1272/2008).

Category	**Classification criteria**
Acute category 1	E(L)C50 [a] < 1 mg/L
Chronic category 1	E(L)C50 [a] < 1 mg/L and the substance is not rapidly degradable and/or the experimentally determined BCF ≥ 500 (or, if absent, $logK_{ow}$≥4)
Chronic category 2	1 < $E(L)C_{50}$ [a] ≤ 10 mg/L and the substance is not rapidly degradable and/or the experimentally determined BCF ≥ 500 (or, if absent, $logK_{ow}$≥4), unless NOECs > 1 mg/L
Chronic category 3	10 < $E(L)C_{50}$ [a] ≤ 100 mg/L and the substance is not rapidly degradable and/or the experimentally determined BCF ≥ 500 (or, if absent, $logK_{ow}$≥4), unless NOECs > 1 mg/L
Chronic category 4	Data do not allow classification under the above criteria, but there are some grounds for concern (e.g. poorly soluble substances for which no acute toxicity is recorded at levels up to the water solubility, and which are not rapidly degradable and have an exp. determined BCF ≥ 500 (or log K_{ow} ≥ 4))

[a] 96h LC50 for fish, 48h EC50 for crustacea and 72h EC50 for algae.

Always in the context of chemical screening and classification, under REACH legislation aquatic toxicity data are considered, in addition to other toxicological information (e.g. carcinogenicity, mutagenicity and toxicity for reproduction), to identify substances that fulfil the toxicity criterion (T) in PBT assessment (i.e. substances that are persistent, bioaccumulative and toxic).

Despite determination of definitive criteria for T is normally based on chronic tests, in order to minimize animal testing (as expressly required by REACH), acute tests are considered at a screening level. In particular, a substance is considered to potentially meet the criteria for T when an acute $E(L)C_{50}$ value from a standard acute toxicity test is less than 0.1 mg/l (Table 4.4). If the screening criterion is met, the substance is referred to definitive T testing and chronic studies are required.

In addition to data from standard toxicity tests, or when toxicity data are not available, data obtained from alternative non-testing methods (e.g. QSARs) can be accepted at a screening level. In these cases, a complete and detailed documentation and assessment of their reliability, adequacy and relevance is required. If QSAR estimations indicate that the substance fulfils the screening criteria for T (i.e., EC_{50} or LC_{50} < 0.1 mg/l), long-term testing are necessary. However, if the predicted acute $E(L)C_{50}$ is < 0.01 mg/l, it may be decided to avoid confirmatory chronic testing on fish.

Table 4.4. Use of acute experimental data and non-testing data for T (screening) assessment (ECHA, 2008c).

Endpoint	Type of data	Criterion	Screening assignment*	Definitive assignment
Short-term aquatic toxicity	acute tests or valid QSARs	EC50 or LC50 ≥ 0.1 mg/L	presumably not T	-
Short-term aquatic toxicity	acute tests or valid QSARs	EC50 or LC50 < 0.1 mg/L	potentially T	-
Short-term aquatic toxicity	acute tests	EC50 or LC50 < 0.01 mg/L	-	T

* The screening assignments should always be considered together for P, B and T to decide if the substance may be a potential PBT/ vPvB candidate.

Due to the large number of chemicals produced worldwide lacking experimental data, industries, regulatory authorities and non-governmental organizations have a growing interest in the development of reliable QSAR models, which could facilitate the risk assessment procedure and support decision-making.

Currently, the (Q)SARs for the prediction of aquatic toxicity with the widest diffusion and application is ECOSAR[2]. The Ecological Structure Activity Relationships (ECOSAR) Class Program is an easy-to-use and freely available computer program developed and routinely applied by the US EPA (Environmental Protection Agency) for predicting aquatic toxicity to fish, aquatic invertebrates (daphnids), and green algae. In the ECOSAR tool, acute and chronic toxicity for several aquatic species are predicted by equations dependent on the logarithm of octanol-water partition coefficient (logP), developed ad-hoc for 120 chemical classes. Prediction capability of ECOSAR models have been extensively evaluated in literature (Reuschenbach *et al.,* 2008; Moore *et al.,* 2003; Tunkel *et al.,* 2005) for large sets of industrial chemicals, with varying molecular structures. However, the (Q)SARs currently available in ECOSAR for some specific classes are based on very limited training sets (composed, in some cases, of 1 or 2 data only), are not statistically robust and the applicability domain is not clearly defined.

In the present study, several QSAR models have been developed and proposed for the prediction of aquatic toxicity of a specific class of hazardous chemicals studied within the CADASTER Project, i.e. substituted triazoles and benzotriazoles (B-TAZs). The high water solubility, resistance to biodegradation and the occurrence of these emerging pollutants in water bodies (Wolschke *et al.,* 2011; Giger *et al.,* 2006) raised concerns on the potential adverse effects toward the aquatic life.
The developed QSARs aimed at predicting three key endpoints that are required in regulation for the assessment of aquatic toxicity of chemicals, i.e. the acute toxicity in algae, daphnids and fish. The

[2] http://www.epa.gov/oppt/newchems/tools/21ecosar.htm

main objective of the models was to directly predict the toxicity toward the three aquatic organisms, independently of *a priori* knowledge of the chemical mode of action (MoA), and to apply them for the screening of over 300 B-TAZs without experimental data. Many of these chemicals are included in the ECHA pre-registration list[3] and might need registration under REACH.

Part of this study was involved in a collaborative activity within the CADASTER Project. Different Partners, including LnU (Linnaeus University), HMGU (Helmholtz Zentrum Munchen), IDEA (Ideaconsult Ltd) and IVL (Swedish Environmental Research Institute), developed different QSAR models for the same endpoints (acute toxicity in algae, daphnids and fish) by applying different modelling approaches (MLR-OLS, PLSR, ANN) and theoretical molecular descriptors (DRAGON, PaDEL-Descriptor and QSPR-THESAURUS web). As task of the CADASTER Project, a Consensus model should be proposed by integrating the predictions generated by the QSAR models developed by the various Partners.

4.2 Methods

4.2.1 Modelled endpoints

In compliance with the first OECD Principle for QSAR validation, i.e. "a defined endpoint", and to limit experimental variability, each QSAR was developed using only acute toxicity data measured in a single species. The three species, representative of the three key trophic levels of the aquatic ecosystem, were selected according to the quality and amount of experimental data available.

QSAR models were developed for the following eco-toxicity endpoints:

- EC_{50} in the green algae *Pseudokirchneriella subcapitata* (formely known as *Scenedesmus capricornutum*), measured as growth inhibition within 72 hours by the OECD 201 test protocol;
- EC_{50} in the crustacean *Daphnia magna*, measured by the OECD 202 immobilization test (considering 48h of exposure);
- LC50 in the fish *Oncorhynchus mykiss*, measured within 96h of exposure, according to the OECD 203 test.

[3] The first deadline of the REACH implementation process (30th November 2008) implied the pre-registration of all substances manufactured/imported in quantities ≥ 1 ton/year and/or already on the market in EU (phase-in substances).

4.2.2 Data sets

Experimental data for acute toxicity in *Pseudokirchneriella subcapitata* (EC_{50} 72h), *Daphnia magna* (EC_{50} 48h) and *Oncorhynchus mykiss* (LC_{50} 96h) were collected from the FOOTPRINT Pesticide Properties Database (PPDB), a database of physiochemical and (eco)toxicological data on pesticides developed in the context of the EU-FP6 research project (PPDB, 2009). In the database, each data point is associated to a score related to data quality which varies between 1 (worst quality data) and 5 (best quality data). In particular, 1 stands for "estimated data with little or no verification", 2 for "unverified data of unknown source", 3 for "unverified data of known source", 4 for "verified data", and 5 for "verified data used for regulatory purposes". Due to the fundamental relevance of the quality of the input data to the performance of QSAR models, only data corresponding to the highest quality-scores were used for QSAR modeling. In particular, data with the quality-score 4 and 5 were included in "daphnids" and "algae datasets". Since limited amount of toxicity data were available in the FOOTPRINT database for the species *Oncorhynchus mykiss*, data of quality-score 3, 4 and 5 were included in the "fish dataset". The different datasets were obtained by collecting data for all the compounds containing a triazole (TAZs) or benzo-triazole (BTAZs) ring. Given the limited number of toxicity data for B-TAZs (homogeneously determined on the same species using the same protocol) we also added other azo-aromatic compounds, including diazines, triazines and similar compounds, to enlarge the response and structural domain of the studied datasets.

The final datasets used for model development were composed as following:

- *Algae dataset*: 35 compounds, including 17 B-TAZs and 18 azo-aromatic compounds;
- *Daphnids dataset*: 97 compounds, including 46 B-TAZs and 51 azo-aromatic compounds;
- *Fish dataset*: 75 compounds, including 27 B-TAZs and 48 azo-aromatic compounds.

In addition to B-TAZs included in the datasets, other compounds containing triazole or benzotriazole rings without experimental data (Bhhatarai and Gramatica, 2011; Roy *et al.,* 2011) were considered in this study to be screened and predicted for their aquatic toxicity. A final complete dataset of 386 B-TAZs (with or without experimental data) was thus studied (listed in Appendix II – Table A-1).

4.2.2.1. Validation sets

To verify the predictive capability of the models, the datasets were first split into a training set (~70% of compounds), used for model development, and a prediction set (~30% of compounds), used only later for the external validation. As described in Chapter 2 (section 2.5.1), two different splitting techniques were applied: by ordered response and by structural similarity using Kohonen Artificial Neural Networks (K-ANN).

The use of two different splittings for the external validation of the models avoids the possible bias given by only one kind of splitting (either on response or structure).

In the context of the CADASTER Project, additional data for the three endpoints became available after the development of the models and were considered as "blind" evaluation set ("EV set") for further external validation. In particular, 18 data for 96h LC_{50} in *Onchorynchus mykiss* were collected from the FOOTPRINT database, while new experimental data of acute toxicity in *Pseudokirchneriella subcapitata* (EC_{50} 72h) and *Daphnia magna* (EC_{50} 48h) were measured, respectively for 13 and 12 B-TAZs, by two CADASTER Partners (PHI and RIVM) using the same OECD procedures applied in the modelled data (Durjava *et al.,* submitted to *ATLA*). The chemicals tested for acute toxicity in algae and daphnids were specifically selected in a previous prioritization analysis (Paragraph 4.3.2).

Table 4.5 summarizes the datasets (training, prediction and evaluation sets) used for the development and validation of the QSAR models for aquatic toxicity of B-TAZs.

Table 4.5. Data sets used for modeling aquatic acute toxicity of B-TAZs.

Endpoint	**Full dataset**	**Split by response**		**Split by K-ANN**		**EV set**
		N_{TR}	**N_P**	**N_{TR}**	**N_P**	
EC_{50}-72h algae	35	24	11	22	13	13
EC_{50}-48h daphnids	97	65	32	65	32	12
LC_{50}-96h fish	75	53	23	53	23	18

A preliminary analysis performed on the structural and response domains covered by the three EV sets highlighted that the EV sets for the endpoints "EC_{50} 72h algae" and "EC_{50}-48h daphnids" were characterized by a rather constricted range in responses, most notably for algae. Since the statistical parameters normally used to perform the external validation of QSARs (e.g. Q^2_{EXT} F1, F2, F3 and CCC) are highly sensitive to the distribution of the responses in the training and validation sets, it was not reasonable to apply such parameters and, therefore, to perform a statistical external validation of the two models on the EV sets. However, EC_{50} data measured for algae and daphnids were used for a "qualitative" validation of the QSARs developed for these two endpoints. For the qualitative external validation, experimental and predicted EC_{50} data were converted into classes of aquatic toxicity, and then it was verified the agreement between experimental and predicted classes. Four classes of aquatic toxicity were used taking into account the thresholds of $E(L)C_{50}$ applied by EU for the categorization of chemicals hazardous to the aquatic environment (Directive 67/548/EEC – Table 4.3; Regulation (EC) No 1272/2008 – Table 4.4):

(1) very toxic ($EC(LC)_{50} \leq 1$ mg/L)

(2) toxic ($EC(LC)_{50} \leq 10$ mg/L)

(3) harmful ($EC(LC)_{50} \leq 100$ mg/L)

(4) not harmful ($EC(LC)_{50} > 100$ mg/L).

4.2.3 Prioritization of B-TAZs for experimental tests

A list of priority B-TAZs was suggested to the CADASTER Partners responsible for the experimental tests. The prioritization was performed on the basis of their toxicological profile and structural similarity.

Principal component analysis (PCA), based on various theoretical molecular descriptors, was performed to explore the structural similarity of the 386 B-TAZs studied in this work. The final outcome of the experimental design consisted in a list of priority B-TAZs representative of the entire structural space. This selection included also chemicals with a structure similar to the more active B-TAZs, on the basis of the available literature data on vertebrate and invertebrate toxicity (PPDB, 2009), but with no or few experimental data available. Thereupon, a further selection was applied based upon the possibility of purchasing the chemicals to be tested (Durjava et al., submitted to *ATLA*).

4.2.4. Molecular structures and descriptors

Chemical structures for 386 B-TAZs and other azo-aromatic compounds were generated from SMILES notations and verified for their correctness. Structures were drawn and minimized to their lowest energy conformation using the semi-empirical AM1 method in HYPERCHEM software. Various different theoretical molecular descriptors (mono- and bi-dimensional) were then calculated using both commercial (DRAGON ver. 5.5) and freely-available software (PaDEL-Descriptor ver. 2.13, QSPR-Thesaurus). To calculate PaDEL descriptors, a conversion of HYPERCHEM format (hin files) to MDL-MOL format (the recommended format in PaDEL-Descriptor) was performed using Open Babel (ver. 2.3.0). The SMILES notations (from PubChem[4] or generated by Open Babel from hin files) were used to generate QSPR-THESAURUS descriptors. To minimize redundant structural information, constant, near-constant and pair-wised correlated (correlation > 0.95) descriptors were excluded from the original matrix of variables. A total of 308 DRAGON descriptors, 258 PaDEL descriptors and 253 QSPR-Thesaurus descriptors were separately used as input variables for QSAR modeling.

4.2.5 QSAR modelling and Applicability Domain

The models were developed by Multiple Linear Regression (MLR) using the Ordinary Least Squares (OLS) method, and the Genetic Algorithm-Variable Subset Selection (GA-VSS) was applied for variable selection. For each endpoint, the modeling procedure for variable selection was applied to the training sets obtained from the two splittings (by ordered response and by K-ANN) separately, resulting in two parallel populations of models, based on various descriptor combinations. From each

[4] http://pubchem.ncbi.nlm.nih.gov/

population, models were compared for their robustness, predictive performances (toward their respective prediction set chemicals) and applicability domain in order to select the most adequate for the prediction of the modeled endpoints, i.e. acute toxicity in algae, daphnids and fish.

Different statistical parameters were used to validate the models for their goodness-of-fit, robustness and external predictive ability (e.g. R^2, Q^2_{LOO} and Q^2_{LMO}, R^2_{ys}, Q^2_{EXT}F1-F2-F3 and CCC). Arbitrary cut-off values were used to accept models as externally predictive, in the case of small data sets: 0.7 for Q^2_{EXT} F1, F2 and F3, and 0.82 for CCC. In addition, RMSE was used to measure and compare prediction accuracy in the training ($RMSE_{TR}$) and in the prediction ($RMSE_P$) sets.

The Williams Plot was used to verify the presence of both response outliers and structural outliers in the training set. The leverage approach was also applied to evaluate the degree of extrapolation for the predictions obtained for compounds without experimental data. The Insubria Graph was used to visualize interpolated and extrapolated predictions.

4.3 QSAR models for aquatic toxicity of B-TAZs

The main aim of the present study was to develop QSAR models, based on theoretical molecular descriptors, for the prediction of the potential aquatic toxicity of B-TAZs and to propose them as supporting tools for classification and risk assessment of chemicals.

The proposed approach is different than traditional QSARs for toxicity, which are typically based on the logarithm of the octanol/water partition coefficient (logP). LogP is an important parameter representing the hydrophobic properties of substances and is considered to be a model for the absorption of molecules into cellular membranes (which is relevant for, among others, narcosis). However, hydrophobicity is not the only factor involved in biological activity of chemicals, since also other electronic and steric effects play an important role, especially when dealing with specific acting compounds, such as B-TAZs.

The efficiency of logP to model toxicity highly depends on the chemical's mode of action (MoA) as well as on the variability of its experimental/estimated value. Many studies already highlighted the high variability (sometimes several orders of magnitude) of experimentally derived and estimated logP values when using different determination methods (Renner, 2002; Papa *et al.,* 2005). As it is explained in detail in **Publication VI**, in this study we decided to develop global models, based only on theoretical molecular descriptors, and independent of the MoA of the studied chemicals for the following reasons:

- Poor statistical performances were obtained using logP as a single descriptor in QSAR models (range of R^2: 0.49-0.69, range of Q^2: 0.47-0.67).

- large variations in R^2 and Q^2_{LOO} values, up to 20%, were found among QSAR models based on different logP, i.e. "AlogPS_logP" (QSPR-Thesaurus database),"KOWWIN" (EPI Suite), "MLOGP" (DRAGON ver. 5.5) and "XlogP" (PaDEL-Descriptors ver 2.12) (Figure 4.1).

- the dataset is composed of B-TAZs which can be associated to over 20 different mechanisms of toxic action, and the amount of experimental data available for each of these MoA is not sufficient to develop separate MoA-based QSAR models following the OECD principles.

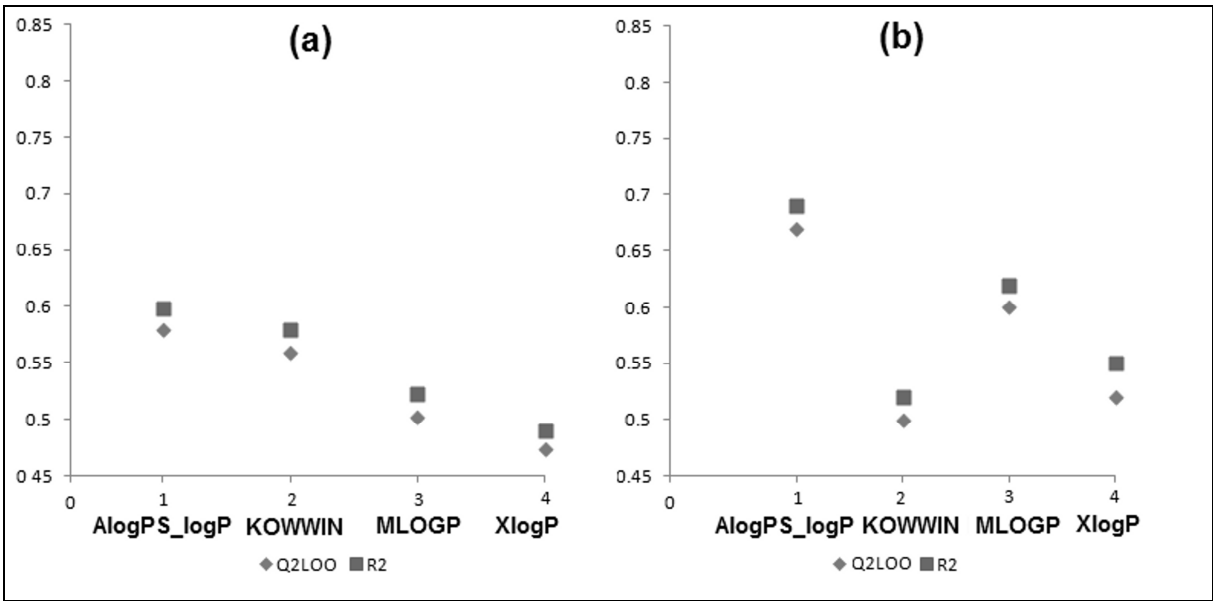


Figure 4.1. Variation in logP-based model performances (R^2 and Q^2_{LOO}) for the prediction of *Daphnia magna* (a) and *Onchorynchus mykiss* (b) toxicity (Figure from **Publication VI**).

Several MLR models, based on DRAGON, PaDEL and QSPR-Thesaurus descriptors, were developed for the three studied endpoints, i.e EC_{50} in *Pseudokirchneriella subcapitata*, EC_{50} in *Daphnia magna*, and LC_{50} in *Onchorynchus mykiss*.

MLR equations and statistical parameters of the QSAR models selected, and proposed in **Publications V** and **VI,** are reported in **Appendix II** (Table A-7). Looking at the statistical parameters reported in the table, it can be observed that the proposed models, based on different molecular descriptors, show good and similar performances in terms of goodness-of-fit, internal robustness and external predictivity, as verified by various criteria.

To support the evaluation of the predictive ability and of the applicability domain of the models, the plot of experimental versus predicted values and the Williams Plots were always analyzed. As an example, Figure 4.2 shows the plots of the model developed for EC_{50} in *Daphnia magna*.

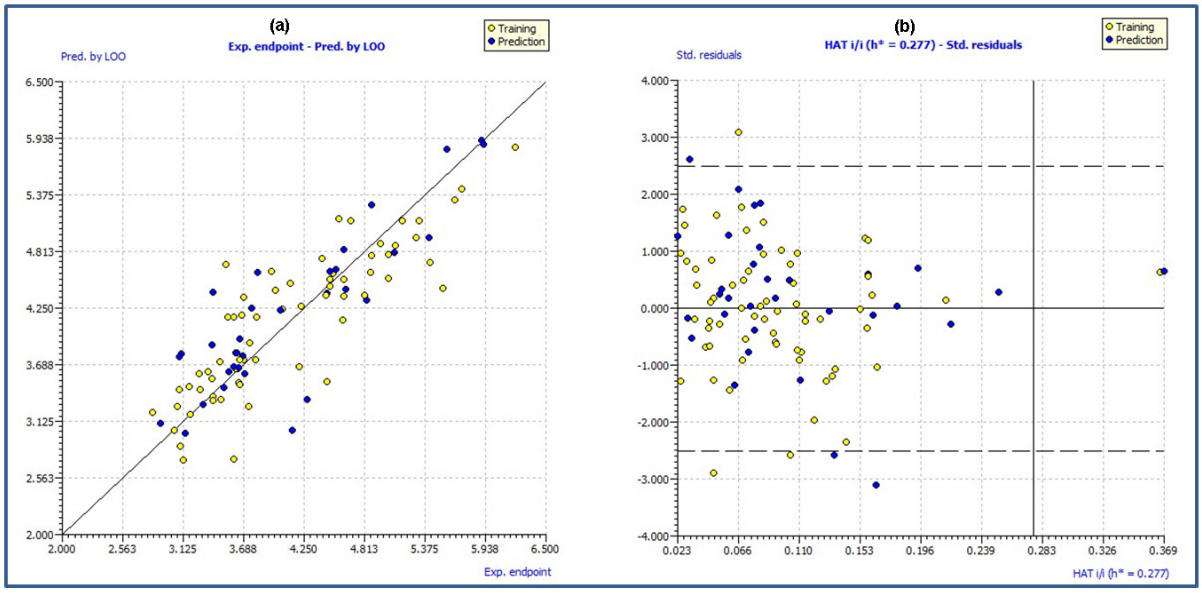


Figure 4.2. Plot of experimental vs. predicted pEC_{50} values (a) and William plot (b) of the model for EC50 in *Daphnia magna* (Split Model based on random splitting).

As was explained in section 4.2.2.1, a statistical external validation of the models developed for algae and daphnids using CADASTER experimental data as a validation set was not possible. However, we decided to assess the capability of the proposed QSARs to correctly classify the chemicals into defined categories of aquatic toxicity, as already applied in literature (Reuschenbach *et al.,* 2008; Dom *et al.,* 2010). For this purpose, EC_{50} values predicted for the 13 chemicals tested by PHI (algae EV set) and for the 12 chemicals tested by RIVM (daphnids EV set) were converted into the following toxicity classes: (1) very toxic ($EC_{50} \leq 1$ mg/L), (2) toxic ($EC_{50} \leq 10$ mg/L), (3) harmful ($EC_{50} \leq 100$ mg/L), (4) not harmful ($EC_{50} > 100$ mg/L).

Tables 4.6 and 4.7 show the experimental and predicted classes of the B-TAZs tested in the CADASTER project for acute toxicity in the algae *Pseudokirchneriella subcapitata* and in *Daphnia magna*, respectively.

Table 4.6. Experimental and predicted toxicity classes of 13 B-TAZs tested by PHI for acute toxicity in the *Pseudokirchneriella subcapitata*.

		Experimental		Pred DRAGON		Pred PaDEL		Pred QSPR-THESAURUS	
Name	**CAS No**	**pEC50**	**class**	**pEC50**	**class**	**pEC50**	**class**	**pEC50**	**class**
Triazophos	024017-47-8	4.62	2	5.58	1	4.94	2	4.99	2
Triadimefon	043121-43-3	4.59	2	4.74	2	4.86	2	4.85	2
Propiconazole	060207-90-1	4.91	2	5.06	2	5.14	2	5.24	2
Penconazole	066246-88-6	4.89	2	4.82	2	4.93	2	5.05	2
Diclobutrazol	075736-33-3	4.87	2	5.13	2	5.23	2	5.25	2
Paclobutrazol	076738-62-0	4.39	3	5.04	2	5.03	2	5.30	2
Hexaconazole	079983-71-4	4.92	2	4.97	2	5.00	2	5.33	2
Uniconazole-P	083657-17-4	4.63	2	4.91	2	5.05	2	5.08	2

		Experimental		Pred DRAGON		Pred PaDEL		Pred QSPR-THESAURUS	
Name	**CAS No**	**pEC50**	**class**	**pEC50**	**class**	**pEC50**	**class**	**pEC50**	**class**
Diniconazole	083657-24-3	5.26	2	4.99	2	5.25	2	5.02	2
Myclobutanil	088671-89-0	4.31	3	5.08	2	5.13	2	5.56	1
Cyproconazole	094361-06-5	4.52	2	4.94	2	4.92	2	5.62	1
Epoxiconazole	106325-08-0	4.58	2	5.18	2	5.25	2	5.13	2
Difenoconazole	119446-68-3	5.45	2	5.92	1	5.73	1	5.48	2

Table 4.7. Experimental and predicted toxicity classes of 12 B-TAZs tested by RIVM for acute toxicity in the *Daphnia magna*.

		Experimental		Pred DRAGON	
Name	**CAS No**	**pEC50**	**class**	**pEC50**	**class**
Benzotriazole	000095-14-7	2.88	4	3.63	3
Guanazole	001455-77-2	4.38	2	3.14	3
Ribavirin	036791-04-5	2.55	4	2.87	4
Triadimefon	043121-43-3	4.00	3	4.13	3
Diclobutrazol	075736-33-3	4.4	3	4.64	2
Paclobutrazol	076738-62-0	3.81	3	4.59	2
Hexaconazole	079983-71-4	4.81	2	4.55	2
Flusilazole	085509-19-9	5.00	2	4.81	2
Myclobutanil	088671-89-0	4.37	3	4.25	3
Cyproconazole	094361-06-5	3.97	3	4.50	2
Fenchlorazole-ethyl	103112-35-2	5.21	2	4.22	3
Triticonazole	131983-72-7	4.52	2	4.69	2

A detailed discussion on this qualitative evaluation of the models is provided in **Publication VII** (submitted to *ATLA*). Overall, good agreement was found between experimental classes and classes predicted by the model for EC_{50} in *Pseudokirchneriella subcapitata*, while several discrepancies were observed in the classification obtained by the model for EC_{50} in *Daphnia magna*. However, it is important to note that the models are more likely to overestimate, rather than underestimate, the toxicity. It can be concluded that the models provide conservative predictions.

4.3.1 Interpretation of modeling descriptors

A detailed description and interpretation of the modeling molecular descriptors is provided in **Publications V** (QSARs for algae toxicity) and **VI** (QSARs for daphnids and fish toxicity). As a general comment, the interpretation of the modeling descriptors suggests that B-TAZs' acute toxicity to the three aquatic species is governed mainly by polarizability, size and branching of the compounds. Among the most important descriptors selected in different models, those encoding for electronic distribution and polar properties of the molecules (e.g. TPSA (NO), TPSA(tot), nHDon, SHBint2, Mp) were all negatively correlated to the toxicity; this indicates that more polar and soluble chemicals

(chemicals with a higher tendency to form hydrogen bonds with water) are less toxic. On the contrary, molecular descriptors related to molecular dimension and branching (e.g. AEigm, nCar, CIC1, VP-1, AeigZ, AMR, C-C) were found to positively correlate with toxicity. These factors are known to increase hydrophobic properties of molecules and, hence, their uptake and subsequent toxicity.

It is interesting to note that the most important descriptors in each model or those singularly more correlated to the studied end-points are also inter-correlated (e.g., 86% correlation between TPSA(tot) and SHBint2, respectively DRAGON and PaDEL descriptors included in the fish model). This demonstrates that the applied variable selection procedure (GA) was able to select, from a highly different set of descriptors as input, molecular descriptors encoding for very similar structural information, even if represented by the various software in different way.

4.3.2 Applicability Domain to a Large Set of B-TAZs

The QSARs developed in this study have been applied to predict the acute toxicity of over 300 B-TAZs without experimental data. We considered as reliable only predictions for compounds falling into the structural AD of the model ($h < h^*$) and borderline compounds. As an example, the Insubria Graph of the model developed for LC_{50} in *Onchorynchus mykiss* is reported in Figure 4.3.

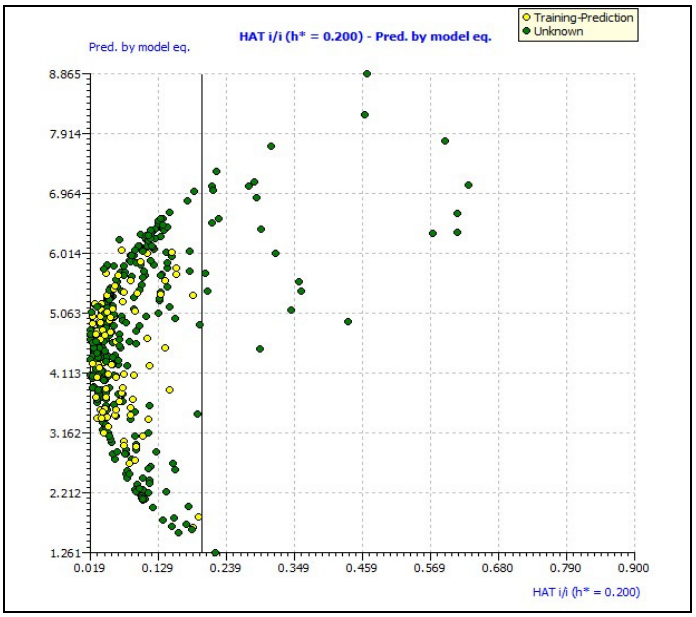


Figure 4.3. Insubria graph of the model for LC50 in *Onchorynchus mykiss* (Full Model).

This analysis highlighted that more than 90% of the studied B-TAZs were included in the structural AD of all the QSARs. Therefore we can conclude that all the proposed models are able to provide interpolated, reliable, predictions for hundreds of B-TAZs without experimental data.

4.4 Consensus models for aquatic toxicity of B-TAZs

In the context of the CADASTER Project, all the Partners involved in the development of models (i.e. LnU, HMGU, IDEA and IVL) developed additional QSARs for the prediction of aquatic toxicity of B-TAZs, using the same datasets and endpoints, but applying different modelling approaches (e.g. PLSR, ANN) and theoretical molecular descriptors.

A summary of the modeling approaches and molecular descriptors used by different Partners is provided in Table 4.8.

Table 4.8. Modelling approaches applied by WP3 Partners for the QSAR models on aquatic toxicity of B-TAZs.

	UI	**LnU**	**IVL**	**IDEA**	**HMGU**
Modelled Endpoint	EC50 algae EC50 Daphnia LC50 fish	EC50 algae EC50 Daphnia LC50 fish	EC50 algae EC50 Daphnia LC50 fish	 EC50 Daphnia LC50 fish	EC50 algae EC50 Daphnia LC50 fish
Input format of chemical structures	HIN files for DRAGON, MOL for PaDEL, SMILES for CADASTER	SMILES	HIN files	SMILES	3D SDF files prepared by Corina
Molecular Descriptors	0D-2D Dragon v.5.5, PaDEL, CADASTER	0D-3D Dragon v.6	0D-3D Dragon v.6	0D-2D Dragon v.5.5	0D-3D CADASTER
Descriptor selection	Genetic Algorithm in QSARINS [3]	Latent variables	Descriptors with VIP>1. Four latent variables	Genetic Algorithm in MobyDigs	Only highly cross-correlated (R>0.95) and almost constant descriptors (less than 3 unique values) were eliminated
Algorithm	Multiple Linear Regression (MLR) using ordinary-least-squares (OLS)	Partial least squares regression (PLSR) and Bayesian Lasso on PLS latent variables (BLASSO-PLS)	Partial least squares regression (PLSR)	Multiple Linear Regression (MLR) using ordinary-least-squares (OLS)	kNN, ASNN, FSMLR, PLS, MLRA, SVM
Applicability Domain	Leverage	Leverage distance to the model on PLS latent variables	DModX	Leverage	STD of ASNN

UI= University of Insubria (this study); LnU=Linnaeus Uniersity; IVL= IVL Swedish Environmental Research Institute; IDEA= Ideaconsult Ltd; HMGU=Helmholtz Zentrum München.

Predictions by consensus were obtained by combining predictions from different models and approaches, taking into account statistical performances and applicability domains of individual models.

For each endpoint, new QSAR models, slightly different from those discussed in section 4.3, were proposed (**Publication VII**). First, the new models proposed for the Consensus modeling are based on higher number of molecular descriptors (one or two descriptors more than the previously presented QSARs). The increase of modeling variables was needed in order to have statistical performances comparable to the other WP3 models (most of them based on hundreds of molecular descriptors) and it didn't affect the quality of the models, since the *occam's razor* principle was still respected. Secondly, in order to facilitate the upload of the Consensus models on the QSPR-Thesaurus database, it was decided that all the models developed by different WP3 Partners should be based on the same training sets. This decision implied a slight modification of the composition of our training sets, now composed of 31, 90 and 77 chemicals for algae, daphnids and fish endpoints respectively.

Table 4.9 summarizes statistical performances of the individual models (developed by different Partners) and of the consensus models developed for the three endpoints. The table provides also the number of molecular descriptors included in each model (N_{desc}) and information related to the model ADs, which were verified for the 386 B-TAZs (AD_{386}).

Table 4.9. Statistical performances of individual and consensus models developed for EC_{50} 72h in *Pseudokirchneriella subcapitata* ("EC_{50} algae"), EC_{50} 48h in *Daphnia magna* ("EC_{50} daphnids") and EC_{50} 96h in *Oncorhynchus mykiss* ("LC_{50} fish").

Endpoint	Model / Method / Descriptors	N_{Desc}	R^2	Q^2	$RMSE_{TR}$	$RMSE_{EX}$	Ext. Val.[b]	AD_{386}
EC_{50} algae	**UI - OLS DRAGON 5.5**	4	0.85	0.78[a]	0.39			88%
	UI - OLS PaDEL-Descriptor	4	0.83	0.76[a]	0.41			93%
	IVL - PLS DRAGON 6.0	375	0.96	0.90[a]	0.21			86%
	LnU – BLASSO-PLS DRAGON 6.0	242	0.7	-	0.58			85%
N_{TR} = 31	**HMGU - ASNN ADRIANA.Code**	118	0.7	0.7	0.53			96%
	Consensus	-	0.88	-	0.36			66%
EC_{50} daphnids	**UI - OLS DRAGON 5.5**	6	0.79	0.75[a]	0.38			89%
	IVL - PLS DRAGON 6.0	245	0.8	0.74[a]	0.37			57%
	LnU – BLASSO-PLS DRAGON 6.0	243	0.59	-	0.53			77%
	HMGU - ASNN ADRIANA.Code	132	0.7	0.7	0.44			91%
N_{TR} = 90	**IDEA - OLS DRAGON 5.4**	5	0.79	0.73[a]	0.38			88%
	Consensus	-	0.82	-	0.36			48%
LC_{50} fish	**UI - OLS DRAGON 5.5**	5	0.82	0.79[a]	0.47	0.43	Q^2ext>0.84	92%
	UI - OLS PaDEL-Descriptor	5	0.76	0.71[a]	0.55	0.4	Q^2ext>0.86	97%
	IVL - PLS DRAGON 6.0	503	0.89	0.75[a]	0.37	0.43	Q^2ext=0.85	73%
N_{TR} = 77	**LnU – BLASSO-PLS DRAGON 6.0**	243	0.7	-	0.62	0.74	Q^2ext>0.53	84%
	HMGU - ASNN ADRIANA.Code	123	0.6	0.6	0.73	0.62	Q^2ext=0.7	76%
N_{EV} = 18	**IDEA - OLS DRAGON 5.4**	6	0.84	0.76[a]	0.45	0.28	Q^2ext>0.93	91%
	Consensus	-	0.85	-	0.44	0.37	Q^2ext>0.88	53%

[a] Q_{LOO}^2; [b] Additional parameters are reported in **Publication VII**.

All the QSARs proposed by different WP3 Partners were checked for their goodness-of-fit, robustness, predictivity and applicability domain, in agreement with OECD principles for the validation of QSARs for regulatory purposes.

The parameter RMSE was used for the comparison of predictive ability of different models. As can be observed in Table 4.9, RMSE values of the consensus models for the three endpoints are lower than the majority of the individual models. The fact that the application of the consensus approach often lead to better predictive performances is widely documented in literature (Zhu *et al.,* 2008; Gramatica *et al.,* 2004; Bhhatarai *et al.,* 2011).

Once verified the robustness and reliability of all the individual WP3 QSARs developed for the acute toxicity in algae (EC_{50} 72h in *Pseudokirchneriella subcapitata*), daphnids (EC_{50} 48h in *Daphnia magna*) and fish (EC_{50} 96h in *Oncorhynchus mykiss*), these models were applied to the complete dataset of 386 B-TAZs, with and without experimental data, for the prediction of the acute toxicity in the three aquatic species. Consensus predictions were derived by averaging (by arithmetic mean) the predictions obtained by individual models.

Applicability domains of individual models toward the 386 B-TAZs were analyzed in order to assess the reliability of predictions. In particular, while the domains of our models as well as the models proposed by IVL, LnU and IDEA are based on the structural space covered by the modeling molecular descriptors, and thus can be considered as "structural domains", the accuracy of predictions (estimated from the standard deviations of ensemble predictions) was used as ADs of the HMGU models, and can be considered as "response domains". Only predictions for compounds included in the applicability domain of different models were considered reliable, i.e. 65.8% for algae model, 48.2% for daphnids model and 53.1% for the fish model. The fact that different approaches were used also to define the reliability of predictions adds confidence to the conclusive assessment of the consensus predictions.

The mean absolute deviation (*MAD*) was used to verify the agreement among predictions obtained by different models for single compounds. *MAD* was calculated as the arithmetic mean of the absolute difference between individual model prediction and consensus prediction:

$$\mathrm{MAD} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - \hat{y}_C\right|$$

where $\hat{y}_i$ is the prediction obtained by individual models (UI, LnU, IVL, IDEA, and HMGU), $\hat{y}_C$ is the prediction obtained by consensus, and **n** is the number of individual models considered to generate consensus predictions.

The higher the MAD value, the higher is the disagreement among model predictions. Figure 4.4 shows the trend of MAD over the 386 B-TAZs for the three modeled endpoints. MAD values

calculated for interpolated and extrapolated predictions, respectively "MAD (in)" and "MAD (out)" in the graphs, are differentiated with different colors.

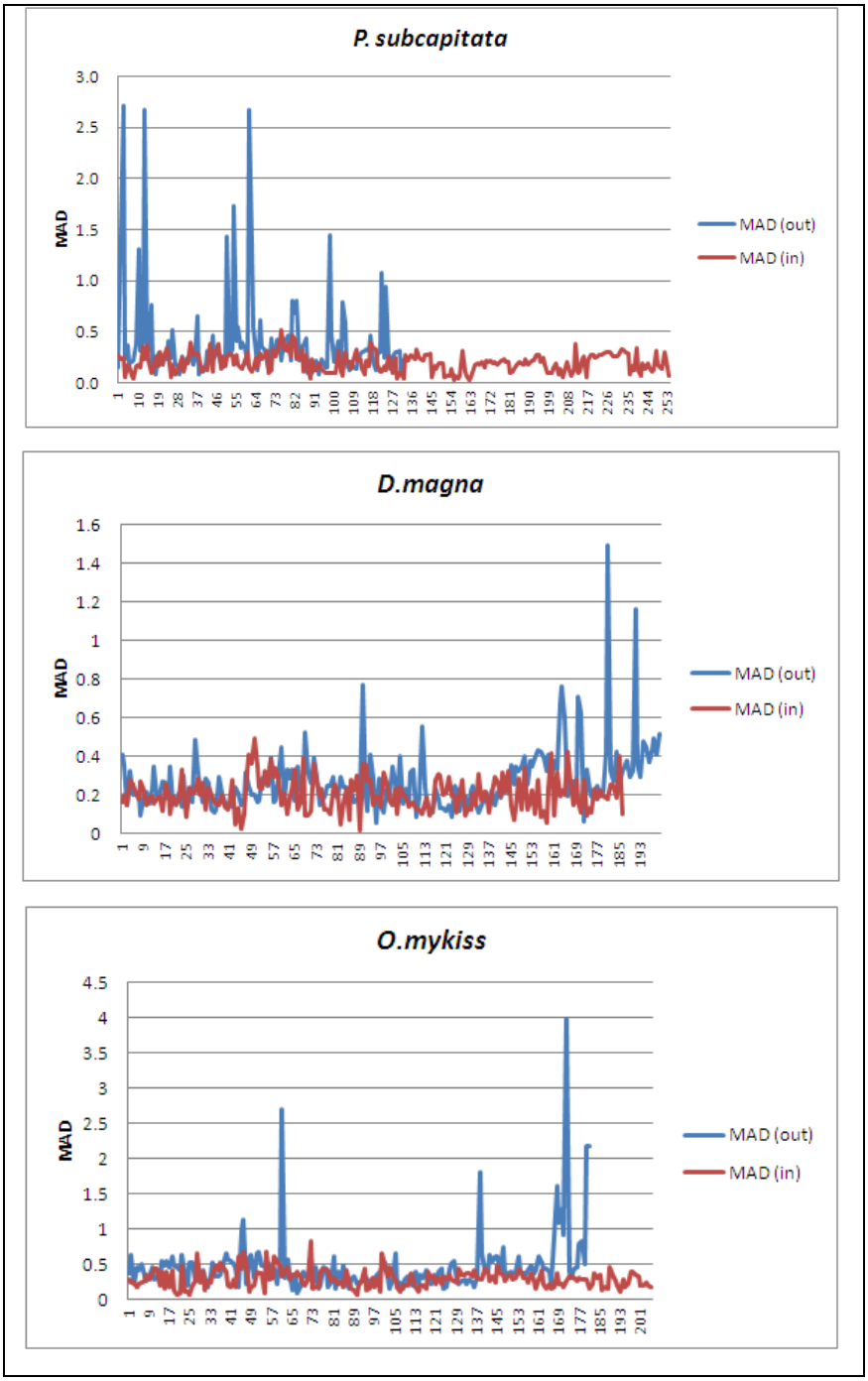


Figure 4.4. Median absolute deviations of predictions by individual models from the consensus predictions for the three modeled endpoints, i.e. 72h EC_{50} in *Pseudokirchneriella subcapitata*, 48h EC_{50} in *Daphnia magna* and 96h LC_{50} in *Oncorhynchus mykiss*.

As expected, the highest disagreement among predictions obtained by different models (MAD > 1 log unit) was found for compounds outside the AD of all the models. These QSAR predictions should be treated carefully since they are model's extrapolations and could be not reliable. On the other hand,

when only B-TAZs included in the applicability domains of all the models were considered, good agreement among predictions was reached. This is evident in Figure 4.4, where MAD values ranging from 0 and 0.5 can be observed for interpolated predictions.

This study highlights the importance of the consensus approach when dealing with chemicals without experimental data. The fact that different models, based on different modeling approaches and with different applicability domains, provide similar predictions increases the confidence in the QSAR estimations, both in terms of prediction accuracy and reliability. To rely on QSAR estimations is particularly important when these QSARs are proposed for their application in regulation or for screening purposes, as is the case of the here proposed models.

4.5 Screening of B-TAZs for acute toxicity in the aquatic environment

Consensus predictions of acute toxicity in *Pseudokirchneriella subcapitata*, *Daphnia magna* and *Oncorhynchus mykiss* obtained for the 386 B-TAZs were analyzed by PCA (Principal Component Analysis) in order to characterize the toxicological profile of B-TAZs and to identify the most active compounds in the aquatic environment. Only predictions for B-TAZs included in the AD of all the models, i.e. 128 compounds, were considered since these predictions were assessed as reliable.

Figure 4.5 shows the plot of the first two principal components (PC1 and PC2), where the 128 B-TAZs (red dots) are distributed according to the weight of each variable, i.e. acute toxicity in algae, daphnids and fish (blue lines). In particular, in the right part of the graph it is possible to see the B-TAZs that are characterized by an overall higher toxicity in the three organisms in the analyzed aquatic scenario (i.e. B-TAZs with higher acute toxicity for algae, daphnids and fish). On the opposite side of the plot, B-TAZs with relative lower acute toxicity for the three species are visualized.

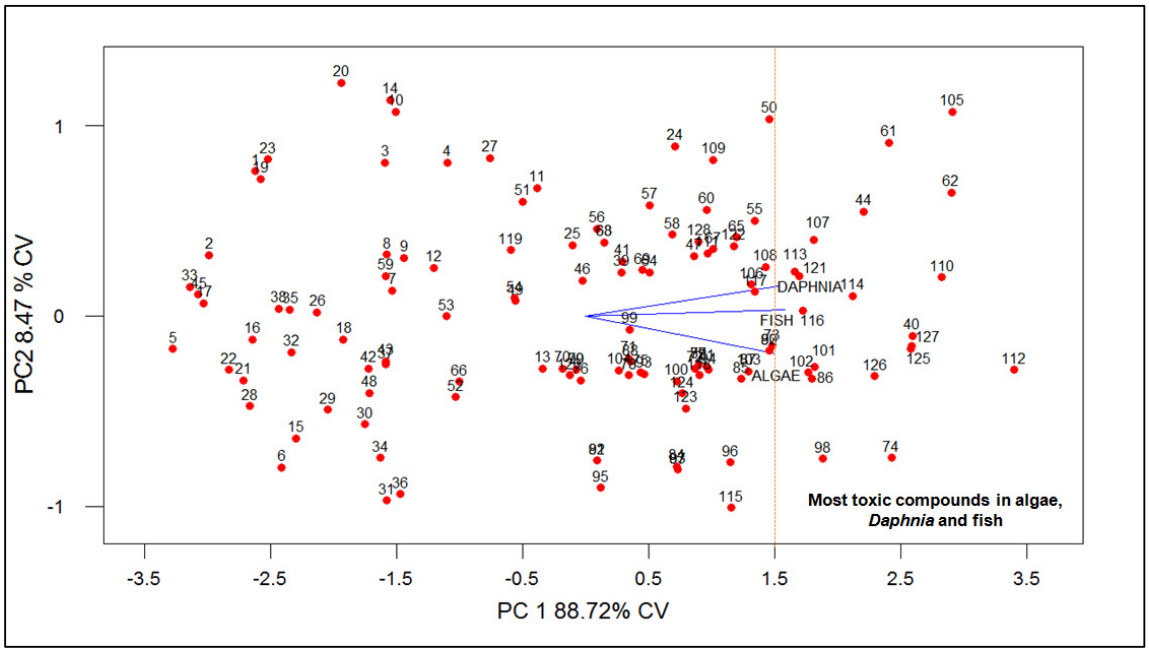


Figure 4.5. PCA analysis (PC1 vs PC2) of reliable consensus predictions obtained for algae (pEC_{50} *Pseudokirchneriella subcapitata*), daphnia (pEC_{50} *Daphnia magna*) and fish (pLC_{50} *Onchorhynchus mykiss*).

Therefore, the first principal component, which explains almost 90% of the overall information, provides a trend of aquatic toxicity, since it separates B-TAZs predicted as globally "more toxic" from less hazardous ones. It is to note that the compounds most globally active are not necessarily the most active for all the three species.

An arbitrary cut-off along the PC1 was defined and the most globally active B-TAZs (PC1 scores >1.5), among the numerous B-TAZs assessed in this study, were identified (Appendix II - Table A-8). For these chemicals, the following ranges of toxicity were predicted for the three organisms:

- *Pseudokirchneriella subcapitata:* EC_{50}(72h) 0.55 - 2.40 mg/L
- *Daphnia magna:* EC_{50}(48h) 1.76-16.99 mg/L
- *Onchorhynchus mykiss:* LC50(96h) 0.36-4.22 mg/L

According to the classification criteria for aquatic toxicity, the prioritized B-TAZs can be classified as "very toxic" ($EC(LC)_{50} \leq 1$ mg/L) or "toxic" ($EC(LC)_{50} \leq 10$ mg/L) for the three species (or at least two), and are threfore highlighted for the necessary experimental tests.

4.6 Conclusions

In the present study, we have developed new QSARs for the prediction of the aquatic toxicity of a class of hazardous environmental pollutants, i.e. triazoles and benzo-triazoles (B-TAZs), toward three key species of the aquatic ecosystem: the algae *Pseudokirchneriella subcapitata* (**Publication V**), the crustacean *Daphnia magna* (**Publication VI,** submitted to *J. Hazard Mater.*), and the fish *Onchorynchus mykiss* (**Publication VI,** submitted to *J. Hazard Mater*). The selected endpoints are among the basic toxicological information that are required to perform risk assessment of chemicals.

The proposed models have been developed and validated on the basis of the OECD principles for QSAR acceptance in regulation (OECD, 2004), are characterized by high external predictivity and wide applicability domain.

We have investigated the possibility to propose models which are more easily applicable by regulators. To this end, models have been developed using molecular descriptors calculated with both commercial (DRAGON) and freely available software (PadEL-Descriptor and QSPR-Thesaurus platform). For all the considered endpoints, valid models based on descriptors generated by the free tools (PaDEL-Descriptor and/ or QSPR-Thesaurus platform) have been proposed. Additionally, to guarantee an easier applicability of the models, they will be freely available, together with the relative documentation (QMRF), in the QSPR-Thesaurus on-line platform.

As part of the CADASTER project, the here proposed QSARs and additional models developed by other Partners were applied to predict aquatic toxicity of over 300 B-TAZs without experimental data, paying particular attention to the applicability domains of the models. Consensus predictions were then proposed for the final assessment of the studied chemicals (**Publication VII**, submitted to *ATLA*). Comparable predictions were obtained when B-TAZs included in the AD of all the models were considered. The fact that different models, based on different descriptors and/or modeling approaches, led to similar predictions adds confidence and reliability to QSAR predictions obtained by Consensus approach.
Considering that some of the screened B-TAZs are included in the ECHA pre-registration list, the here generated QSAR predictions (individual as well as consensus predictions) are very useful for both filling data gaps and prioritize the most dangerous compounds for further experimental testing. In particular, 20 B-TAZs, predicted as toxic ($EC(LC)_{50} \leq 10$ mg/L) or very toxic ($EC(LC)_{50} \leq 1$ mg/L) to algae, daphnids and fish, have been selected for inclusion into a priority list for environmental tests.

Concluding, the predictions generated by the here proposed QSARs for all the studied chemicals can be used by regulators to support the use of Weight of Evidence- and non testing-based approaches for the classification and risk assessment of chemicals. In order to facilitate these procedures, the proposed models and predictions will be made available on-line in the QSPR-Thesaurus Database.

Chapter 5

QSAR Modeling of Ready Biodegradability of Fragrance materials

5.1 Introduction

Environmental half-life of a chemical is a key element for assessing the potential exposure and risk that a chemical presents for humans and wildlife. Chemicals that persist in the environment may be subject to a wider dispersion and remain available to biota for long periods of time, increasing concerns for potential bioaccumulation and long-term adverse effects (Rücker and Kümmerer, 2012).

Environmental half-life is the result of many abiotic processes, including hydrolysis, photolysis, oxidation, reaction of OH radicals, and sorption, as well as biological processes, such as microbial biodegradation, that can lead to the removal or transformation of a chemical substance in the environment.

For the water, soil and sediment compartments, biodegradation is often the most important transformation process. Microbial degradation is also a fundamental process in sewage treatment plants (STP). In biodegradation processes, organic compounds are used by microorganisms as a source of energy and building blocks. In some cases a full mineralization occurs (*ultimate biodegradation*), where the chemical is completely degraded into carbon dioxide, sulphate, nitrate and new biomass. In the *primary biodegradation*, a structural change of the compound is performed by microorganisms resulting in new transformation products, not necessarily less persistent and/or toxic than the parental compound (ECHA, 2008b).

In the context of REACH, information on the degradability of chemicals may be used for hazard assessment (e.g. for classification and labelling), risk assessment (for chemical safety assessment) and persistency assessments in the evaluation of PBT and vPvB chemicals (ECHA, 2008b).

Biodegradation is among the basic information required for hazard and persistency assessments, in particular for the aquatic environment and STPs. Testing biodegradation is therefore a crucial step in the evaluation and registration of chemicals. In the last decades, all industrialized nations developed biodegradation tests which were harmonized by the Organization for Economic Co-operation and Development (OECD) and later in the European Union (EU), in order to guarantee standardized testing procedures and a mutual international acceptance of test results. These tests include simple screening tests of ultimate biodegradability (e.g. OCED 301 ready biodegradation and OECD 302 inherent biodegradation tests), aiming at identifying those chemicals for which more detailed, and costly, studies are needed, and relatively complex higher tiered simulation types of tests (e.g. the OECD 303 aerobic sewage treatment and the OECD 309 aerobic and anaerobic transformation in surface water).

The potential for biodegradation is also basic information that needs to be incorporated in the design and development of new safer chemical products. This is in line with the *Green Chemistry* philosophy,

which encourages the design of chemicals and chemical processes that reduce or eliminate the use or generation of hazardous substances[1] (Horváth and Anastas, 2007).

In particular, one of the 12 Principles of Green Chemistry[2] (Principle 10), originally published by Dr. Paul Anastas (ex EPA Assistant Administrator) and Dr. John Warner in *Green Chemistry: Theory and Practice* (Oxford University Press: New York, 1998), states that "chemical products should be designed so that at the end of their function they break down into innocuous products that do not persist in the environment" (Anastas and Eghbali, 2010).

The "benign by design" concept requires the knowledge of the potential biodegradability to be available in the earliest phases of the chemical production processes, i.e. even before synthesis (Rücker and Kümmerer, 2012). Industrial research and development is increasingly applying this approach to optimize, in terms of costs and time, the long industrial processes involved in the commercialization of new chemicals.

In this context, *in silico* approaches like (Q)SARs, which predict biodegradation potential on the basis of the chemical structure, is a powerful tool for a rational design of new chemicals (e.g. by comparing candidate substances during product development), and can be applied for screening and prioritization purposes, before making a decision on the necessity for testing.

A wide collection of qualitative and quantitative structure-biodegradability relationships is currently available. In literature these models are commonly identified with the acronym (Q)SBRs, but they can be simply considered as (Q)SPRs. In the last years several reviews have been published summarizing the state of the art of (Q)SPRs for biodegradation available both in literature and/or implemented in commercial or freely-available software (e.g. EPI Suite, MultiCase , TOPKAT, Catabol, etc…) (Jaworska *et al.,* 2003, Pavan and Worth, 2008; Rücher and Kümmerer, 2012). In a recent review published in Green Chemistry by Rücher and Kümmerer (2012), the authors critically assess various approaches to predict aquatic aerobic biodegradation, providing practical considerations regarding model's availability, applicability domains, predictive performances and limitations. It is to note that the majority of the available (Q)SPR models for biodegradation estimation have been developed based on training set data consisting of results from ready biodegradability tests, in particular MITI-I data (see section 5.2). Because of the relatively low costs and simplicity of ready biodegradation tests, most of biodegradation data currently available for commercial chemicals are derived using one or more of these methods. A trend in

[1] The concept of *hazardous substance* include the evaluation of toxicity to organisms and ecosystems, persistence and bioaccumulation potential in organisms or environment, and safety with respect to handling and use (http://www.epa.gov/greenchemistry/pubs/about_gc.html).

[2] http://www.epa.gov/sciencematters/june2011/principles.htm

the development of (Q)SPRs for biodegradation over the years is evident, with an evolution from chemical-class specific models, built on small sets of homologous compounds (Rücker and Kümmerer, 2012 and references therein), to broadly applicable models, based on highly heterogeneous datasets (e.g. Cheng *et al.,* 2012). Several modeling approaches have been proposed (linear and nonlinear regression models, partial least square regression, artificial neural networks, classification methods, expert systems, etc…) based on both theoretical molecular descriptors or substructures fragments, automatically selected by specific algorithms or defined by experts.

According to the ECHA guidelines for chemical safety assessment under REACH, the use of (Q)SPR predictions is considered at the screening level for a preliminary identification of substances with a potential for persistency. Concerning the use of (Q)SPRs for the assessment of biodegradation potential, the ECHA guidelines suggest the combined use of results from different estimation models, such as the BioWIN models implemented in EPI suite, paying particular attention to both the validation status of any QSAR model and whether the substance is included in the applicability domain of the model.

In this study the attention was focused on a specific class of substances studied within the CADASTER Project, i.e. fragrance materials, for which specific QSPR models have been developed for the prediction of ready biodegradability. Despite the large quantities used and the continuous exposure to these chemicals, limited information is actually available regarding their health effects and environmental fate. The need to fill data gaps for an ever-increasing number of fragrances and to find commercially and environmentally compatible safer alternatives highlights the importance and utility of *in silico* approaches, like QSAR.

To the best of our knowledge, no QSPR models for ready biodegradability specifically developed for fragrance materials are currently available.

5.2 Methods

5.2.1 Modeled endpoint

Ready biodegradability is an important endpoint that is used, and required in regulation, for the assessment of biodegradability of chemicals on a screening level.

Ready biodegradability tests are stringent screening tests, conducted under aerobic conditions, in which substances are tested in high concentrations when compared to those normally found in the environment (but not inhibiting bacterial growth). Small amounts of a polyvalent inoculum, generally taken from a municipal STP, river water, and/or soil suspension (not artificially pre-adapted to the test substance), are used in order to represent a realistic spectrum of degrading organisms present in the

environment. The test substance is provided as the sole source of carbon for energy and growth and is incubated in the dark for 28 days under conditions favoring biodegradation with respect to pH-value, O_2-content, and temperature (ECHA, 2008b; Beek *et al.,* 2001).

Ultimate biodegradation can be indirectly determined by measuring parameters like Dissolved Organic Carbon (DOC) removal, Biochemical Oxygen Demand (BOD) and CO_2 production. Specific "pass levels" have been chosen for biodegradation test results. Substances are considered "completely" biodegraded or mineralized when > 60% of the ThOD (theoretical oxygen demand) or ThCO2 (theoretical carbon dioxide), or > 70% DOC removal is reached within a certain time span (normally 28 days).

Table 5.1 gives an overview of the internationally standardized OECD 301 tests with relative pass levels for biodegradation.

Table 5.1. Standardized screening tests for ready biodegradability with relative pass levels.

Test Name	**Guideline**		**Endpoint**	**Pass level**
	OECD [ref]	**EU [ref]**		**(threshold)**
DOC Die Away-Test	301 A	C.4-A	DOC removal	70%
CO_2 evolution-Test	301 B	C.4-C	$ThCO_2$[a]	60%
Modified MITI I-Test	301 C	C.4-F	ThOD[b]	60%
Closed Bottle-Test	301 D	C.4-E	ThOD	60%
Modified OECD Screening-Test (MOST)	301 E	C.4-B	DOC removal	70%
Manometric Respirometry-Test	301 F	C.4-D	theoretical oxygen demand (ThOD)	60%

[a] ThCO2, theoretical carbon dioxide: the amount of CO_2 that theoretically can be produced if the test substance is completely oxidized by microorganisms. [b] ThOD, Theoretical oxygen demand: the amount of oxygen that theoretically can be consumed if the test substance is completely oxidized by microorganisms.

Ready biodegradability of tested chemicals is evaluated on the basis of degradation percentage after 28 days and fulfilment of the 10 days-window criterion[3].

In particular, if ≥ 60% of ThOD or ≥ 60% of $ThCO_2$ or ≥ 70% DOC removal occurs within the 10 day-window, than the chemical will be assessed as "readily biodegradable".

If the pass levels are not met within the 10-day window, the substance is classified as "not readily biodegradable".

A positive result in a test for ready biodegradability can be considered as indicative of rapid and ultimate degradation in most environments. In such cases, no further investigation of the biodegradability of the

[3] The pass levels have to be reached in a 10-day window within the 28-day period of the test. The 10-day window begins when the degree of biodegradation has reached 10% DOC removal, ThOD or $ThCO_2$ and must end before or at day 28 of the test.

chemical, or of the possible environmental effects of transformation products, is normally required. If a substance is evaluated as ready biodegradable, than it doesn't meet the P-criterion which is important for PBT assessment.

The stringent test conditions may sometimes lead to conflicting tests results obtained by different methods. In some cases, these differences could for instance be due to the origin of the inoculum used in the tests, which may differ in the adaptation to the test substance.

However, a negative result in a test for ready biodegradability does not necessarily mean that the chemical will not be degraded under realistic environmental conditions.

Rücher and Kümmerer (2012) stated that, even if standardized methods are used, biodegradation data derived from screening tests are intrinsically highly variable and poorly reproducible. According to them, the stringent, and not realistic, conditions of different tests as well as the diversity of microorganisms used as inoculum are the main sources of variability, since they can be the determining factor in assessing a chemical as ready or not ready biodegradable. To overcome this problem, two approaches can be suggested: i) compare data within a set of compounds only if obtained in the same test and using the same inoculum (Rücher and Kümmerer, 2012), or ii) compare data obtained by different testing methods and use only data where consistent positive or negative results are observed, as performed in this study.

5.2.2 Data set

5.2.2.1 Training set

Ready biodegradability data used for the development of QSAR models were mainly collected from the Japanese MITI *(Ministry of International Trade and Industry)* database and publicly available ECHA dossiers. Data were collected by one of the CADASTER Partners involved in WP3 (IDEA group) and provided to the other Partners as an SDF-file. The molecular structures of chemicals were provided in the form of SMILES notations and MOL-files. The provided SDF-file included ready biodegradability data for nearly 1800 heterogeneous organic chemicals. For each compound, information related to chemical identity (i.e. Name, SMILES string, CAS number, InChI code and EC number), ready biodegradation percentages obtained by one or more testing methods (i.e. OECD 301A, B, C, D, E, F), as well as an overall assessment of ready biodegradability (i.e. “RB” if ready biodegradable and “NRB” if not ready biodegradable) were provided. The IDEA Partner performed a preliminary cleaning of the dataset by removing duplicates, salts and chemicals with inconsistent biodegradation results among different testing methods, thus reducing the size of the dataset to 1250 chemicals.

Since the primary objective of the present study was to propose QSAR models for the classification of biodegradability of fragrances, an intensive filtering of the dataset was needed in order to remove any redundant and not useful structural information (section 5.2.4). The final dataset was composed of 136 chemicals (70 RB and 66 NRB), including only fragrance materials and fragrance-like chemicals (i.e. compounds structurally similar to fragrances). Starting from the MOL-files provided by IDEA, mono- and bi-dimensional descriptors were generated by using the commercial software DRAGON (ver 5.5) and the open source software PaDEL-Descriptor (ver 2.12).

5.2.2.2 Validation set

In the context of the CADASTER Project, additional data of ready biodegradation became available specifically for fragrance materials. In particular:

- Experimental data for 64 substituted musks/fragrances, derived from different testing methods (OECD 301 B, C, D, F) were provided by Dr. Dan Salvito from RIFM (*Research Institute for Fragrance Materials)*. A preliminary analysis of the dataset allowed us to clean the dataset from duplicates, chemicals with no OECD 301 data and chemicals with inconsistent biodegradation results among different methods. Additionally, six fragrances having biodegradation percentages in the range of 40-70% (close to the cut-off values of different methods) and/or with chemical structures more similar to chemicals in the opposite class were removed from the dataset. This was done in order to minimize the experimental uncertainty deriving from erroneous structural and/or response information included in the dataset. The final "RIFM dataset" consisted of 35 chemicals belonging to different chemical classes (assigned by RIFM), i.e. Aldehydes/Cyclic, Esters/Salicylates, Heterocycles/Oxygen Containing/Pyrans, Ketals/Cyclic and Phenols/Alkoxy. According to the biodegradation percentage and the testing method used, each chemical was assigned to the class "ready biodegradable" (RB) or "not ready biodegradable" (NRB).

- Ready biodegradability percentages were measured for 11 fragrance materials by the CADASTER Partner PHI (*Public Health Institute Maribor)* by using the OECD 301D test. These chemicals were previously selected by our Research Unit as priority compounds on the basis of the available information on potential toxicity (cyto-toxicity and mammalian toxicity) and structural representativeness (Papa *et al.,* 2009). The tested fragrances were separated into ready and not ready biodegradable according to the biodegradation percentage, i.e. "RB" if biodegradation > 60% (pass level for OECD 301D test) and "NRB" if biodegradation < 60%.

The two datasets of fragrances (RIVM and PHI datasets) were compared, and one overlapping compound (ID 6259-76-3, RB for both the datasets) was removed from the "RIFM dataset". A final dataset of 45 fragrances was defined and used for the external validation of the developed classification models.

The chemical structures of the 45 fragrances were drawn, verified for their correctness and minimized to their lowest energy conformation by the semiempirical AM1 method using the HYPERCHEM software. 3D structures (*.hin files) were then converted into SMILES notations and MOL files by the software Open Babel 2.3.0.

As was done for the training set chemicals, mono- and bi-dimensional molecular descriptors were calculated using the software DRAGON 5.5 (starting from HIN-files) and PaDEL-Descriptors 2.12 (starting from MOL-files).

5.2.3 QSAR modelling and Applicability Domain

For each pool of descriptors (calculated using DRAGON and PaDEL-Descriptor software), separate *k*-NN classification models were developed on the same training set (136 chemicals), and will further be referred to as "DRAGON Models" and "PaDEL Models". The *k*-NN method was applied to auto scaled data, and the *a priori* probability to belong to a class was set as proportional to the number of chemicals in the two classes of ready biodegradability; the predictive power of the model was checked for *k* values between 1 and 10.

As explained in chapter 2 (section 2.4.2), performances of classification models are normally evaluated on the basis of the confusion matrix (Table 5.2), which can be generated both for training and validation set chemicals. In this case, the confusion matrix was used to calculate the percentages of RB and NRB chemicals correctly classified as well as the overall accuracy (OA %).

Table 5.2. Confusion matrix for the classification of ready biodegradability.

	Assigned Class			
Real Class	RB	NRB	Accuracy	
RB	a	b	RB %	(a/a+b)*100
NRB	c	d	NRB %	(d/c+d)*100
			OA %	[(a+d)/a+b+c+d]*100

In order to verify the ability of the proposed models to generate reliable predictions, also for new chemicals, the applicability domain (AD) was assessed taking into account the theoretical structural

space defined by the descriptors used in the models. Two approaches were used, one based on the range of descriptors selected in each model, the other one based on the leverage approach.

In accordance with the first method, chemicals with descriptor values within the range of those of the training set compounds were considered as being inside the AD of the model. Compounds falling outside the descriptors' space were considered as structural outliers and thus outside the AD of the model.

According to the second method, the limit of model domain was quantitatively defined by the leverage cut-off (h^*) set as 3(p+1)/n; leverage values greater than h^* mean that the query compound is outside of the structural model AD.

Predictions of compounds lying outside the ADs of the proposed models were considered as extrapolations, thus less reliable.

5.2.4 Dataset cleaning and balancing

Definition of the training set chemicals is one of the most important phases in QSAR modelling. The selection of a representative dataset, in terms of both structure and response, is a requirement for the development of a robust structure-activity relationship.

In this case study, the composition of the dataset provided by IDEA (1250 chemicals) was structurally highly heterogeneous and just a small portion of the structural domain was covered by fragrances.

A first modelling attempt was made using a training set composed of 146 chemicals measured with the OECD 301D test (Closed Bottle Test). This decision was taken in order to limit the experimental variability derived by different testing methods and assuming that chemicals tested with the same method should share similar structural and/or physico-chemical properties. The selection of the OECD 301D test was based on the fact that this specific method was used by the CADASTER Partner PHI for the measurement of ready biodegradability of fragrances, whose data were included in the validation set.

Despite these assumptions, the selected training set included many chemicals highly different from the most common structural classes characterising fragrance materials. Several internally robust QSARs were developed using this training set (OA=75-80%). However the structural information encoded in the modelling descriptors was not able to correctly classify ready biodegradability of fragrances included in the validation set (RB<25%, NRB=80-90%).

For this reason, further analysis and cleaning of the MITI dataset was performed in order i) to remove structural information not representative for fragrance materials and ii) to extend the training set to all the fragrances included in the big dataset collected by IDEA. In a step-by-step approach, where each step involved QSAR development and validation, many chemicals were excluded from the original dataset and

a final "fragrance dataset" of 187 chemicals, including 100 fragrances and 87 "fragrance-like" chemicals, was defined.

A preliminary SAR (structure-activity relationship) analysis of the dataset was performed in order to characterize the composition of the selected dataset as well as identify any particular structural feature characterizing RB or NRB compounds.

As it is shown in Figure 5.1, it was possible to recognize specific functional groups and sub-fragments most frequently included either in RB chemicals (e.g. alcohols, aldeydes, ketons, esters, amino-groups, etc..) or in NRB chemicals (e.g. aromatic rings, nitro-groups, quaternary carbons, etc..). However, it is known that interactions with other fragments present in the molecule can also play an important role in ready biodegradability (Loonen *et al.,* 1999). In our dataset, for example, chemicals containing ethers or phenols groups were more likely to be RB or NRB according to the type of additional fragments included in the molecules (e.g., concurrent presence of alcohol groups enhanced RB potential, while the presence of quaternary/tertiary carbons or aromatic rings characterized NRB chemical).

All the above observations find a confirmation in literature (Loonen *et al.,* 1999; Boethling *et al.,* 2007; Boethling, 2011; Cheng *et al.,* 2012) and suggest that we have selected a meaningful and representative training set from both a structural and response points of view.

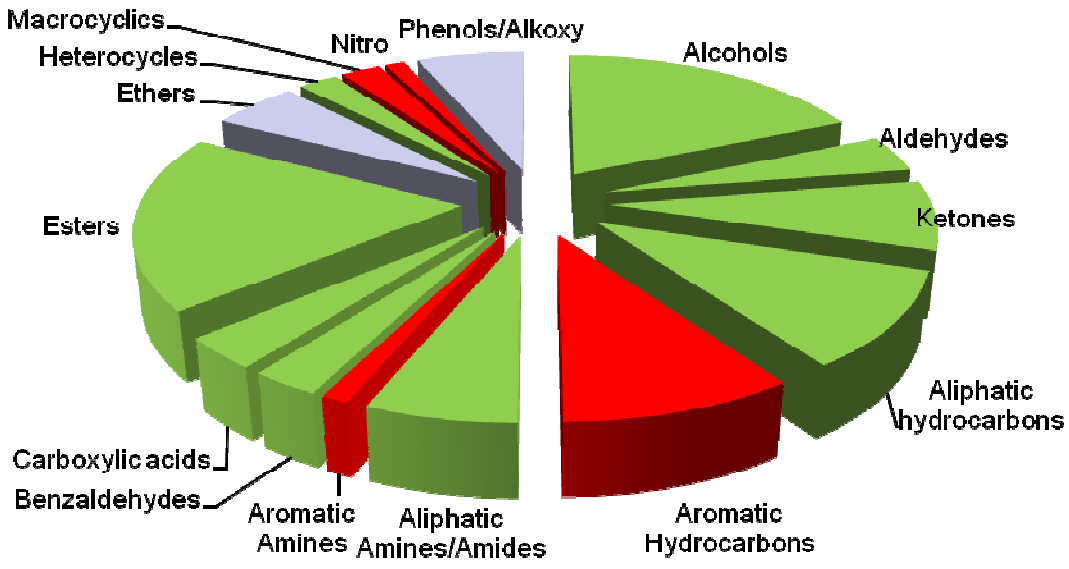


Figure 5.1. Graphical representation of ready biodegradability of 187 chemicals in the dataset. Functional groups/sub-fragments most frequent in RB chemicals are labelled in green, those most frequent in NRB chemicals are labelled in red.

In this dataset, a problem was identified concerning the response distribution. In fact, the number of RB chemicals was much larger than that of NRB compounds (i.e. 121 RB vs 66 NRB).

This problem of unbalanced datasets has been widely discussed in literature. Many studies have shown that unbalanced class distributions result in poor classification performances from standard classification algorithms, which assume a relatively balanced class distribution and equal misclassification costs (Sun *et al.,* 2007; Maloof, 2003; Zhang and Mani, 2003). Classification rules that predict the small class tend to be fewer and weaker than those that predict the prevalent class, often leading to a high misclassification of test samples belonging to the small class.

A method commonly used to deal with the problem of unbalanced data consists of under-sampling the majority class (Kubat & Matwin, 1997). Therefore, before model development, it was decided to pre-process the dataset by reducing the dimension of the most representative class, i.e. RB.

The sampling of the chemicals to remove from the RB class was performed by applying the factorial analysis (FA) (section 2.3.2). The factorial analysis was performed on the first four principal components (derived by a PCA based on 1D and 2D DRAGON descriptors), thus generating 16 groups of structural similarity (Table 5.3). According to the dimension of each group, 1 to 5 chemicals were sampled as representative compounds of the group and included in the final training set. This procedure allowed to reduce the number of RB chemicals (from 121 to 70) by conserving a structural representation of the RB class.

Table 5.3. Data set balancing performed by applying factorial analysis.

FA Gruop	N	$N_{deleted}$	$N_{sampled}$
a	7	2	5
b	4		4
c	6	2	4
d	4		4
e	12	8	4
f	1		1
g	10	5	5
h	8	4	4
i	15	6	9
j	5		5
k	4		4
l	9	4	5
m	4		4
n	3		3
o	16	11	5

FA Gruop	N	$N_{deleted}$	$N_{sampled}$
p	13	9	4
	121	51	70

The dataset used for QSAR development was finally composed of 136 chemicals, 70 RB and 66 NRB. DRAGON (ver. 5.5) and PaDEL (ver 2.12) descriptors were recalculated for the 136 training chemicals (from MOL-files). Constant, near-constant and pair-correlated descriptors (pair-wise correlation > 0.98) were excluded from the pool of descriptors used for QSAR modelling. Additionally, DRAGON descriptors no longer supported or having different values in the last updated version of DRAGON (ver. 6.0) were deleted[4]. At the end of this procedure a final set of 222 Dragon descriptors and 241 PaDEL descriptors were separately used as input variables in the model development.

Before modelling, a preliminary analysis of the dataset was performed by principal component analysis on the bases of the calculated 1D and 2D Dragon descriptors (Figure 5.2). Despite the low information explained by the first two components (C.E.V. 33%), this analysis highlighted that:

i) fragrances included in the validation set shared a similar structural space of the training set, thus indicating that they could be reasonably used for model validation, and

ii) no linear separation surfaces could be defined between RB and NRB chemicals. This highlighted the complexity of the studied dataset (despite the pre-processing procedures performed) and supported the classification algorithm selected, i.e. *k-NN*.

[4] Changes between Dragon 5.5 and Dragon 6.0, available at:
http://www.talete.mi.it/help/dragon_help/index.html?changes_dragon_5_5_6.htm

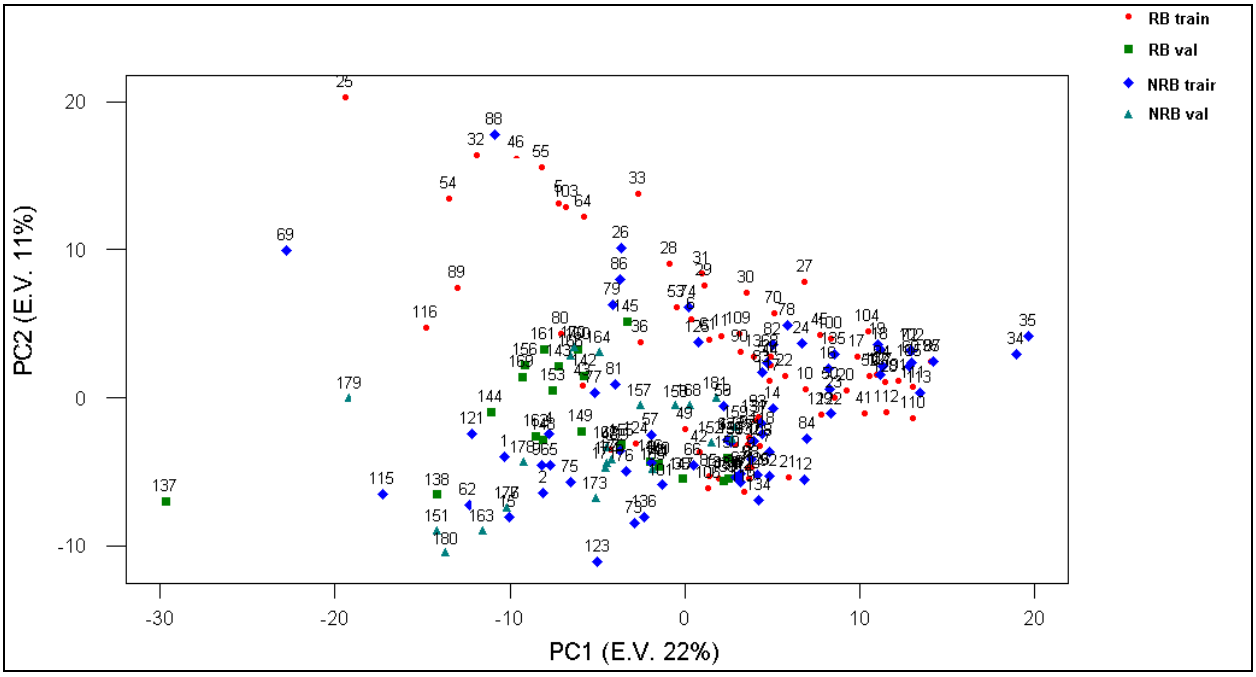


Figure 5.2. Principal Component Analysis (PC1 vs PC2) of training and validation sets based on 1D-2D DRAGON descriptors. RB and NRB chemicals included in training and validation sets are marked with different labels.

5.3 QSAR models for ready biodegradation of fragrances

Using an in-house software, models up to 5 variables were explored by maximizing the overall percentage of correct assignments (OA%). In this procedure, all the models based on one or two variables, obtained calculating all the possible pair-wise combinations of the molecular descriptors (All Subset variable selection method), were first explored. The genetic algorithm was then applied to select the best combination of modelling descriptors (maximum model dimension set as 5 variables). Finally, two independent populations of models, one based on DRAGON descriptors and the other one based on PaDEL descriptors, were generated, each consisting of the best 100 models, ordered by OA%.

The two populations were analysed taking into account the following criteria:

i) Model accuracy in classifying training set chemicals. This implied the analysis of fitting and cross-validated RB%, NRB% and OA% of the developed models, which are indices of the goodness-of-fit and robustness.

ii) Model accuracy in classifying validation set chemicals. RB%, NRB% and OA% were calculated from the confusion matrix obtained for the validation set, and were used to evaluate predictive ability of the models.

iii) Interpretability of modelling descriptors. Particular attention was paid to the possible mechanistic interpretation of the molecular descriptors selected by GA as relevant for classifying RB and NRB chemicals.

5.3.1 Classification models based on DRAGON descriptors

According to the selection criteria explained above, three "best" models based on different combinations of DRAGON descriptors, each combination encoding for different structural information, were identified. Classification performances and modelling descriptors of the three DRAGON Models are reported in Table 5.4.

Table 5.4. Modelling descriptors and classification performances of the three best Dragon Models for ready biodegradability of fragrances.

			Training set (N=136)			**Validation set (N=45)**		
Model ID	**Descriptors**	***k***	**RB%**	**NRB%**	**OA%**	**RB%**	**NRB%**	**OA%**
M1	nCIC, X0A, nR=Ct, F01[C-O]	*5*	86	74	80	73	61	67
M2	TI1, Vindex, nCq, H-052, B06[C-O]	*4*	83	73	78	68	61	64
M3	Sv, Qindex, MAXDP, GGI5, nCq	*6*	87	77	82	73	70	71

As can be observed in Table 5.4, the three proposed models are able to correctly classify both RB and NRB chemicals, with an average OA over 80% for training set chemicals. However, a higher specificity of the proposed models for the classification of RB chemicals was noted. This can be due to a higher structural heterogeneity among the NRB chemicals which is not exhaustively explained by the molecular descriptors included in the models.

Lower classification accuracy was observed for the validation set, where M3 was the best model with a classification accuracy of 73% and 70% for RB and NRB respectively. For further application of the proposed models to new chemicals, predictions obtained in a consensus approach are here suggested. It is known that the consensus approach, by combining different models, often enhances predictive power and reduce misclassification errors (Boethling, 2004). As can be observed in Table 5.5, improvement in classification performances was observed if consensus predictions, i.e. class assignments obtained in agreement with at least two models, were considered. Class assignments of individual models and consensus model are provided in **Appendix III** (Tables A-9, A-10).

Table 5.5. Classification performances of the Consensus Model for ready biodegradability of fragrances.

		Training set (N=136)			Validation set (N=45)		
Model ID	***k***	**RB%**	**NRB%**	**OA%**	**RB%**	**NRB%**	**OA%**
M1	5	86	74	80	73	61	67
M2	4	83	73	78	68	61	64
M3	6	87	77	82	73	70	71
Consensus		*94*	*80*	*88*	*73*	*74*	*73*

As commented in the previous chapter (section 4.4), consensus modelling, which averages predictions obtained from various models based on different descriptors and/or different approaches, often lead to significant improvement of predictive performances. In our case, in a population of 100 models, many possible "best" models, based on different molecular descriptors, are included, each representing different views to describe the structural features that are related to the studied endpoint. In this sense, the Consensus approach can help to combine the structural information encoded by different models and to complement the deficiencies of individual models with the support of the other.

5.3.2 Classification models based on PaDEL descriptors

In the population of 100 models based on PaDEL descriptors, one "best" model was selected and is reported in Table 5.6.

Table 5.6. Modelling descriptors and classification performances of the best PaDEL Model for ready biodegradability of fragrances.

			Training set (N=136)			Validation set (N=45)		
Model ID	**Descriptors**	***k***	**RB%**	**NRB%**	**OA%**	**RB%**	**NRB%**	**OA%**
M4	maxHBa, maxHssNH, maxssssC, WTPT-2	7	81	79	80	73	70	71

The model based on PaDEL descriptors, proposed as an alternative to DRAGON models, shows high classification performances, since it is able to correctly classify around 80% of the training set chemicals and 70% of the fragrances included in the validation set. In this case, comparable accuracy in classifying both RB and NRB chemicals can be observed.

Differently than observed for the DRAGON models, consensus modelling, obtained by combining different good models included in the population of PaDEL models, didn't lead to a significant improvement of the classification performances. This can be explained by the similar structural domains

covered by the molecular descriptors included in different PaDEL models, implying similar predictions and, hence, classification errors for the same molecules. Therefore, in order to provide a simpler applicability of the model, only one model based on the freely calculable PaDEL descriptors was proposed.

Class assignments of the proposed model are provided in **Appendix III** (Tables A-9, A-10).

5.3.3 Interpretation of modeling descriptors

The selection of the "best" models was also based on the interpretability of modelling descriptors.

As can be observed in Tables 5.4 and 5.6, all the proposed models are based on a limited number (4 or 5) of mono- and bi-dimensional theoretical descriptors, encoding for various structural features that have been automatically selected by GA as relevant for classifying RB and NRB chemicals in the studied dataset. To help the understanding of the selected descriptors, a list of the descriptors with the corresponding definitions (as provided by DRAGON and PaDEL-Descriptor supporting information) is reported in Table 5.7.

Table 5.7. Modelling molecular descriptors and corresponding definitions.

Descriptor	Descriptor type	Definition	Model
nCIC	Constitutional descriptors	number of rings	M1 (Dragon)
X0A	Connectivity indeces	average connectivity index chi-0	M1 (Dragon)
nR=Ct	Functional group counts	number of aliphatic tertiary carbons	M1 (Dragon)
F01[C-O]	2D frequency fingerprints	frequency of C-O at topological distance 01	M1 (Dragon)
TI1	Topological descriptor	First Mohar index	M2 (Dragon)
Vindex	Information indices	Balaban V index	M2 (Dragon)
nCq	Functional group counts	number of total quaternary Carbons	M2, M3 (Dragon)
H-052	Atom centred fragments	number of H atteched to C0 (sp3) with 1X attached to the next C	M2 (Dragon)
B06[C-O]	2D binary fingerprints	presence/absence of C-O at topological distance 06	M2 (Dragon)
Sv	Constitutional descriptors	sum of atomic van der Waals volumes, scaled on Carbon atoms	M3 (Dragon)
Qindex	Topological descriptors	Quadratic index	M3 (Dragon)
MAXDP	Topological index	maximal electrotopological positive variation	M3 (Dragon)
GGI5	Topological charge indices	topological charge index of order 5	M3 (Dragon)
maxHBa	E-State descriptors[a]	Maximum E-States for (strong) Hydrogen Bond acceptors	M4 (PaDEL)
maxHssNH	E-State descriptors[a]	Maximum atom-type H E-State: -NH-	M4 (PaDEL)
maxssssC	E-State descriptors[a]	Maximum atom-type E-State: >C<	M4 (PaDEL)
WTPT-2	Weighted Path descriptor	Molecular ID / number of atoms	M4 (PaDEL)

[a] Atom type electrotopological state descriptors.

As a general comment, we observed that the majority of molecular descriptors selected by GA encoded for structural information already highlighted (also in literature) as important for characterising compounds as ready or not ready biodegradable. Among the descriptors that can be easily interpreted *nCIC,* which counts the number of rings (including aromatic rings) in the molecule, is a known fragment increasing resistance to ready biodegradation. The importance of this descriptor for discriminating between RB and NRB chemicals was already stressed by Cheng *et al.* (2012), who found a high significant difference between the mean number of rings of RB and NRB substances. Cheng *et al.* conclude that "higher number of rings is unfavourable for chemical biodegradability". In the present dataset, *nCIC* was found to be highly correlated with descriptors *TI1* (corr.=70%), *Qindex* (corr.=87%) and *WTP-2* (corr.=83%), which were selected as modelling variables in models M2, M3 and M4 respectively.

Other important descriptors highly related to NRB potential are *nR=Ct* (included in M1) and *nCq* (included in both M2 and M3), which count the number of tertiary and quaternary carbons, respectively and which are also known structural alerts for NRB chemicals. Similar information is encoded by the PaDEL descriptor *maxssssC*, which is an E-state descriptor identifying the presence of carbon atoms bonded to four other carbon atoms with single bonds (>C<), i.e. quaternary carbons.

Two important descriptors, i.e. *F01[C-O]* and *B06[C-O],* relevant for RB potential were selected, which identify the presence of functional groups containing oxygen (such as alcohol groups, aldehydes, ketones, carboxylic acids and esters). As has been reported before, chemicals including these groups are more easily degraded by oxidation, hydrolysis or conjugation reactions (Loonen *et al.,* 1999; Boethling, 2007; Chen *et al.,* 2012).

An additional PaDEL descriptor of easy interpretation selected by GA was *maxHBa*, an E-State descriptor for (strong) hydrogen bond acceptors. The fact that hydrogen binding ability is relevant for classifying RB and NRB chemicals is in contrast with what was found in the publication by Chen *et al.* (2012), where the authors state that "hydrogen binding ability is not a key factor for chemical biodegradation". However, it is well known that biodegradation is a complex process that depends on many factors, including physico-chemical and structural features of the chemicals but also environmental conditions. Differences found between our findings and literature could be due to the different experimental information used to develop the models (Jawroska *et al.,* 2003; Rücher and Kümmerer, 2012).

5.3.4 AD analysis: structural and response domain

The here proposed classification models were analyzed both for their structural and response applicability domains, the first for the identification of compounds structurally most diverse from the training set, the latter for the detection of misclassified chemicals.

To define the structural domain of the models, two different approaches were used:

i) the leverage approach, with *h** set at 0.11 for M1/M4 and 0.13 for M2/M3;

ii) the range of modelling descriptors within training set chemicals (Table 5.8).

Table 5.8. Range of molecular descriptors (minimum and maximum values) selected in the four proposed models (M1, M2, M3 Dragon models and M4 PaDEL model).

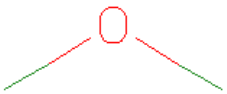
Model	Descriptor	min	max
M1	*nClC*	0	3
	X0A	0.67	0.902
	nR=Ct	0	3
	F01[C-O]	0	12
M2	*TI1*	-87.084	40.11
	Vindex	0.276	2.042
	nCq	0	2
	H-052	0	30
	B06[C-O]	0	1
M3	*Sv*	4.31	35
	Qindex	0	17
	MAXDP	0	5.165
	GGI5	0	1.618
	nCq	0	2
M4	*maxHBa*	0	12.23918
	maxHssNH	0	0.607986
	maxssssC	0	0.606481
	WTPT-2	1.638071	2.10786

The leverage approach was applied for the training set in order to highlight chemicals which are very influential in selecting modelling variables and, for the validation set to identify chemicals outside the model domain. The AD determined by the range of descriptors was based on training set chemicals and applied for the validation set.

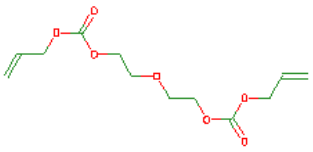
A good agreement was found among the structural outlier chemicals identified by different models and by applying the two methods to define structural AD. However, none of the identified chemicals were particularly far from the structural domains of the models. It is to note that the leverage approach turned out to be more restrictive in comparison to the domain defined by the range of descriptors.

On the basis of the leverage approach, we highlight seven chemicals included in the training set characterized by the highest *h* values in different models. These chemicals were therefore particularly influential in selecting molecular descriptors in the respective models. Chemical structures of the highlighted chemicals, with respective *h* values, are shown in Figure 5.3.

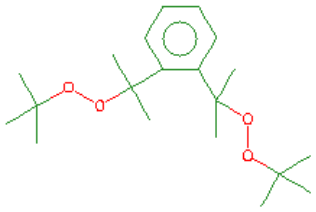
ID 34 – CAS 115-07-1 (NRB)
h=0.22 (M2)



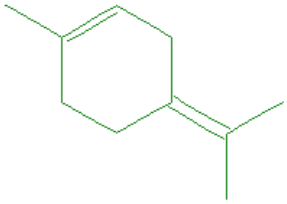
ID 35 – CAS 115-10-6 (NRB)
h=0.22 (M2)



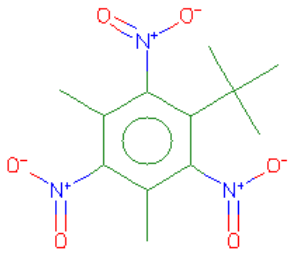
ID 54 - CAS 142-22-3 (RB)
h=0.22 (M1)



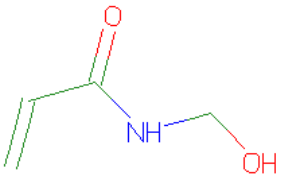
ID 69 – CAS 25155-25-3 (NRB)
h=0.41 (M2)



ID 93 – CAS 586-62-9 (RB)
h=0.24 (M1)



ID 115 – CAS 81-15-2 (NRB)
h=0.46 (M3)



ID 122 – CAS 924-42-5 (NRB)
h=0.56 (M4)

Figure 5.3. High leverage chemicals within the training set.

Combining the two approaches and taking into account the structural AD defined by different models, three fragrances included in the validation set were recognized as outliers for all the models (Figure 5.4). This is mainly related to high number of rings included in the molecules. However, despite these chemicals are not covered by the structural domains of the proposed models, two of them (ID 27 and ID 44) are correctly classified by all the models (except M2 that misclassifies ID27 as RB).

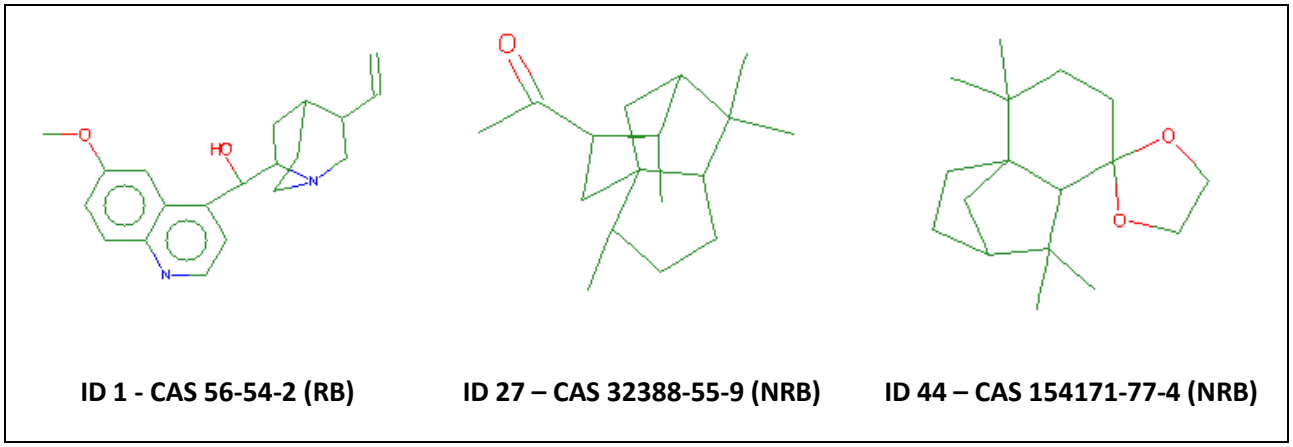
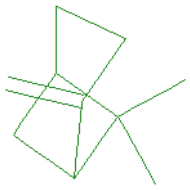


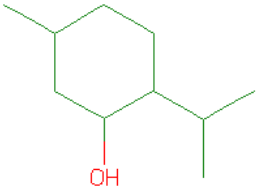
Figure 5.4. Structural outliers fragrances included in the validation set.

The analysis of misclassified chemicals was also performed. Figures 5.5 and 5.6 show the chemical structures of training and validation set compounds, respectively, misclassified by both Dragon Consensus and PaDEL models.

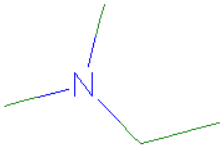
RB chemicals classified as NRB



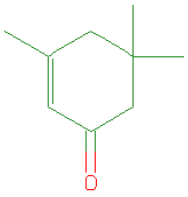
ID 47 - 127-91-3



ID 59 - 15356-60-2



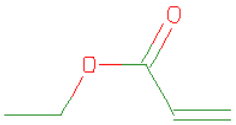
ID 95 - 598-56-1



ID 108 - 78-59-1

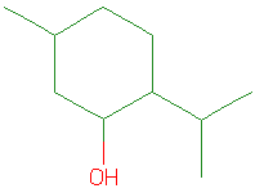
NRB chemicals classified as RB

ID 26 - CAS 111-81-9

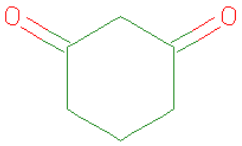


ID 50 - CAS 140-88-5

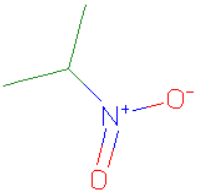
ID 51 - CAS 141-32-2



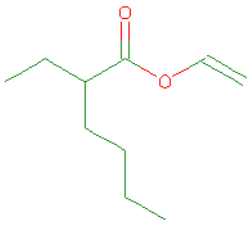
ID 56 – CAS - 1490-04-6



ID 84 - CAS 504-02-9



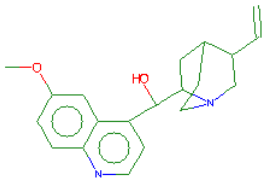
ID 113 - CAS 79-46-9



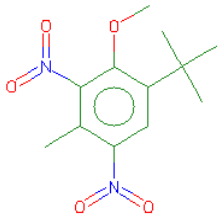
ID 125 - CAS 94-04-2

Figure 5.5. Training set chemicals misclassified by both Dragon Consensus and PaDEL models.

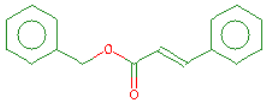
RB chemicals classified as NRB



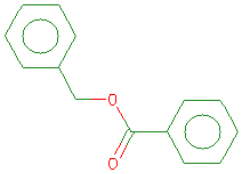
ID 1 - CAS 000056-54-2



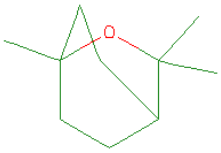
ID 2 - CAS 000083-66-9



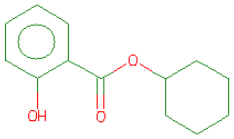
ID 8 - CAS 000103-41-3



ID 12 - CAS 000120-51-4

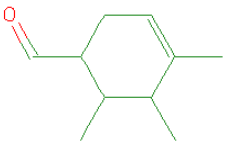


ID 14 - CAS 000470-82-6

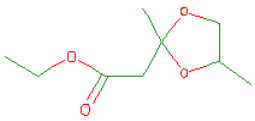


ID 26 - CAS 025485-88-5

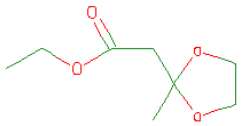
NRB chemicals classified as RB



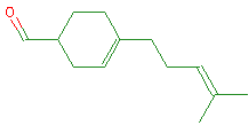
ID 16 - CAS 001335-66-6



ID 21 - CAS 006290-17-1



ID 22 - CAS 006413-10-1



ID 28 - CAS 037677-14-8

Figure 5.6. Validation set chemicals misclassified by both Dragon Consensus and PaDEL models.

A possible explanation of the misclassifications generated by the proposed models can be provided, also taking into account the interpretation of modeling descriptors discussed in the previous paragraph (section 5.3.3).

Looking at the chemical structures reported in figures 5.5 and 5.6, it is evident that all the RB chemicals classified as NRB present structural features and fragments that could be mainly associated to NRB chemicals, such as the presence of one or more tertiary and quaternary carbons, aromatic rings and nitro groups. It is particularly surprising how the nitro-musk included in the validation set (ID 2) was measured to be 100% ready biodegradable. In fact, it is well known from literature that nitro-musks, like the one

here studied, actually incorporate a series of structural features (e.g. *tert*-butyl group, multiple nitro groups, high degree of ring substitution) that make them particularly resistant to aerobic biodegradation (Boethling *et al.,* 2007). Therefore, in this case we are more likely to support the QSAR prediction rather than the experimental data.

Concerning the misclassification of NRB compounds, the main explanation is related to the presence in their structure of aliphatic chains, multiple oxygen atoms and ester groups, which are structural elements that are commonly known to enhance ready biodegradability.

5.4 BIOWIN Tool

A comparison of the here proposed models, developed *ad hoc* for fragrances, and the widely used BIOWIN Tool is here presented.

The Biodegradation Probability Program for Windows (BIOWIN, ver. 4.10) is a freely available tool implemented in the EPI Suite Software that calculates the probability of rapid aerobic and anaerobic biodegradation with mixed cultures of microorganisms. BIOWIN contains seven separate models, whose main features are summarized in Table 5.9. Five of these models, i.e. Biowin 1, 2, 3, 5 and 6, are intended to be used for the prediction of ready biodegradability.

Table 5.9. Summary of biodegradation models available in BIOWIN.

BIOWIN Model	Brief Description	Result interpretation
Biowin1 = linear probability model **Biowin2** = nonlinear probability model	General indication of biodegradation under aerobic condition	Probability ≥0.5: Biodegrades Fast Probability <0.5: Does NOT Biodegrade Fast
Biowin3 = expert survey ultimate biodegradation model **Biowin4** = expert survey primary biodegradation model	Estimates for the time required to achieve complete ultimate (3) and primary (4) biodegradation in a typical or "evaluative" aquatic environment	5.00 = hours; 4.00 = days; 3.00 = weeks; 2.00 =months; 1.00 = longer
Biowin5 = MITI linear model **Biowin6** = MITI nonlinear model	Predictive models for assessing a compound's biodegradability in the Japanese MITI ready biodegradation test; i.e. OECD 301C.	Probability ≥0.5: Readily Degradable Probability <0.5 : NOT Readily Degradable
Biowin7 = anaerobic biodegradation model	Predicts probability of rapid degradation in the "serum bottle" anaerobic biodegradation screening test. This endpoint is assumed to be predictive of degradation in a typical anaerobic digester.	Probability ≥0.5: Biodegrades Fast Probability <0.5: Does NOT Biodegrade Fast

Biodegradability estimates generated by BIOWIN models are based upon fragment constants that were developed using multiple linear or non-linear regression analyses, depending on the model. Biowin1, Biowin2, Biowin5 and Biowin6 estimate the likelihood of rapid biodegradation; in particular, Biowin1 and Biowin2 are based on a dataset of weight-of-evidence biodegradability evaluations for 264 chemicals, while Biowin5 and Biowin6 are based on a set of MITI data for 884 compounds. Biowin3 provides semi-quantitative estimates of rates of ultimate biodegradation on the basis of an evaluation of several US-EPA experts who analysed 200 substances. All the BIOWIN models are based on a set of preselected substructures (the same set of 36 fragments is used in Biowin 1, Biowin2 and Biowin3, and a slightly different set of 42 substructures is used in Biowin5 and Biowin6) and molecular weight as independent variables.

Finally, the BIOWIN software provides an overall YES or NO prediction, derived by a combination of Biowin3 and Biowin5 estimations. More in detail, if Biowin3 result is "weeks" or faster (i.e. ≥2.75) and Biowin5 estimate is "readily degradable" (i.e. ≥0.5), then the prediction is YES (readily biodegradable). If these conditions are not satisfied contemporarily, the prediction is NO (not readily biodegradable).

Classification accuracy of Biowin6 toward a selected group of fragrances, i.e. musks, was already evaluated in literature (Boethling, 2011). In that study, the author applied a threshold of 0.3 for classifying RB and NRB chemicals (instead of the criterion set by Biowin6 at 0.5 - Table 5.9) and was able to correctly classify ready biodegradability of more than 90% of chemicals (36/39 correct classifications).

In the present work, it was decided to evaluate the predictive ability of the BIOWIN tool toward the fragrances considered in our study. In particular, estimations generated by Biowin6 (using the same threshold of 0.3, as applied by Boethling in the before mentioned study) and results obtained using a BIOWIN combined approach (Biowin3 and Biowin5), as recommended in literature (Boethling et al., 2004) were used.

A comparison of classification performances among the models developed in this study, i.e. Consensus model based on Dragon descriptors and the model based on PaDEL descriptors (M4), and BIOWIN, i.e. Biowin6 and BioWIN WoE (combined approach of Biowin3 and Biowin5) are reported in Table 5.10.

Table 5.10. Classification performances of the here proposed models and BIOWIN.

	Training set (N=136)			**Validation set (N=45)**		
Model ID	**RB%**	**NRB%**	**OA%**	**RB%**	**NRB%**	**OA%**
Consensus (Dragon)	94.3	80.3	87.5	72.7	73.9	73.3
M4 (PaDEL)	81.4	78.8	80.1	72.7	69.6	71.1
Biowin6 (0.3)	88.6	15.2	52.9	77.3	34.8	55.6
BioWIN WoE	72.8	62.1	67.6	63.6	78.3	71.1

Despite the very good predictive performances showed in the publication by Boethling (Boethling, 2011), it is evident that Biowin6 is not adequate for the prediction of the chemicals considered in this study. The general overestimation of ready biodegradability of chemicals included in our two datasets leaded to a very low overall classification accuracy (OA%=15-35%). On the contrary, the combined approach of Biowin3 and Biowin5 seems to be more appropriate for the fragrances here studied, correctly classifying ~68-70% of chemicals. The good performances of this approach can be explained by the wider structural and response information obtained by combining two different models and again highlights the importance of consensus modelling.

It is also important to note that the development of new QSARs *ad hoc* for a specific group of chemicals, in this case fragrances, significantly improved the classification performances (OA% of the here proposed models increased 10-20% when compared with the OA% of BIOWIN).

Nevertheless, a higher specificity toward NRB chemicals of the validation set was found for the Biowin WoE approach, which is able to correctly classify 78% of NRB fragrances (74% of correct assignments performed by our proposed models). The lower performances of our models can be due to a lack of specific structural information on NRB fragrances included in the modelling descriptors; this information is probably covered by the large set of sub-fragments included in the BIOWIN models. However, the overall classification accuracy of BIOWIN toward the validation set is lowered by the worse ability to recognize RB chemicals (RB% = 63.4%). The higher specificity of BIOWIN toward NRB chemicals in respect to RB compounds was already recognised in various validation studies of BIOWIN (ECHA, 2008b).

5.5 Conclusions

In this study, robust and externally predictive classification models have been developed for the prediction of ready biodegradability of fragrance materials. Ready biodegradation is among the basic information required for risk assessment for the evaluation of environmental fate and persistence of chemicals.

Two alternative QSPR models, characterized by good and comparable classification performances, have been proposed:

- a Consensus model, derived by integrating the predictions generated by the best three QSPRs based on DRAGON descriptors, and
- a QSPR model based on molecular descriptors that can be calculated with the freely available software PaDEL-Descriptor.

The proposal of the PaDEL model was done in order to allow a wider and feasible application of the model by any user, bearing in mind the potential use of the proposed models as support tools for risk assessment and decision-making.

It is important to stress that the use of the genetic algorithm for the automatic selection of variables allowed for the selection, among hundreds of theoretical descriptors, of descriptors really relevant for the classification of chemicals according to their biodegradation potential. In fact, the selected molecular descriptors encoded for important structural features known to enhance the biodegradability of chemicals, such as the presence of alcohol groups, ketones, carboxylic acids and ester, or, conversely, to increase the resistance to ready biodegradation, such as the presence of rings, tertiary and quaternary carbons.

It was confirmed that the combination of multiple models allows, in the majority of cases, to improve classification performances. The use of the consensus approach is therefore promoted in order to enhance predictive power of QSARs and reduce misclassifications.

Concluding, the here proposed QSPRs represent useful tools to be applied, in combination with already existing and accepted models (e.g. BIOWIN), for the hazard and risk assessment of chemicals; more in detail, they are particularly suitable for fragrance materials. Additionally, in line with the *Green Chemistry* philosophy, they could be used *a priori* to design new alternative compounds, which are potentially less persistent in the environment.

Chapter 6

Overall Conclusions and Future Perspectives

Under the European REACH regulation there is an urgent need to acquire, by the next few years, a large amount of information necessary to assess and manage the potential risk of thousands of industrial chemicals. Meanwhile, in order to reduce costs and experimental time, and to ensure good animal welfare, REACH aims at reducing animal testing by promoting the intelligent and integrated use of alternative methods, such as *in vitro* testing and *in silico* techniques. Among these methods, models based on quantitative structure-activity relationships (QSAR) are useful tools to fill data gaps and to support the hazard and risk assessment of chemicals.

The present thesis was performed in the context of the CADASTER Project, which aims at providing a practical guidance for the integration of *in-silico* models in risk assessment procedures. The main topic of this thesis was the development of QSAR/QSPR models for the characterization of the (eco-)toxicological profile and environmental behaviour of four classes of emerging pollutants studied within the CADASTER Project (i.e. brominated flame retardants (BFRs), fragrances, perfluorinated compounds (PFCs) and (benzo)-triazoles (B-TAZs)), for which limited experimental information is currently available to perform a complete risk assessment.

The major outcomes of this thesis can be summarised as following.

- Several QSAR/QSPR models have been proposed to predict endpoints that are relevant, under the REACH regulation, for the evaluation of the environmental and/or human toxicological hazard (i.e. aquatic toxicity and endocrine disrupting potential), and for the estimation of the environmental persistence (i.e. ready biodegradability). In particular, data for acute toxicity in the aquatic species as well as information on ready biodegradability potential, are key endpoints required in regulation to perform the hazard assessment and to identify substances that fulfil the toxicity (T) and persistence (P) criteria in the PBT assessment. No clear guidance are currently available for the assessment of endocrine disrupting chemicals; however the evaluation of ED properties is important for the assessment of SVHCs.

- The here presented QSARs have been developed on the basis of the OECD principles for QSAR acceptability for regulatory purposes. Particular attention has been paid to the external validation procedure and to the statistical definition of the applicability domain of the models. The relevance of QSAR models to screen and rank chemicals without experimental data has also been highlighted.

- A strong effort was made in order to propose QSAR models that could be easily applied and reproducible by future users (e.g. expert users from industry and regulation). To reach this purpose, in addition to the models based on molecular descriptors generated by the widely used commercial software DRAGON, “alternative” QSARs have been proposed, which were

based on descriptors calculated with the freely available software PaDEL-Descriptors. Moreover, models for the prediction of aquatic toxicity of B-TAZs have been implemented in the QSPR-Thesaurus database, which is the web-platform developed within the CADASTER Project that will allow for a wide distribution and easy free application of the QSARs developed in this thesis.

- It was demonstrated, in all the presented case-studies (Chapters 3-5), that higher prediction accuracy is obtained when models based on different molecular descriptors are combined and, possibly, calculated by different modelling approaches. This confirms the validity of the consensus approach.
- The models developed in this thesis have been used in the CADASTER project to screen large amount of chemicals (in particular, 243 BFRs, 54 PFCs and 386 B-TAZs). These screening have been performed to prioritise few chemicals, which have been predicted as the most hazardous (in relation to the modelled endpoint), and which have been suggested for further experimental tests (Durjava *et al*., submitted to *ATLA*). The application of this screening procedure demonstrated how the QSAR approach can be applied to extract useful information, also when starting from limited amount of data.
- An important output of this thesis within the CADASTER Project consisted in the integration of the developed QSAR models in hazard and risk assessment procedures (Golsteijn *et al.,* submitted to *Environmental Science & Technology*; Iqbal *et al.,* in press). As an example, QSAR predictions generated for aquatic toxicity of B-TAZs were used as input data for the assessment of the potential ecotoxicological impact (i.e., Comparative Toxicity Potentials (CTPs)) of triazoles. This case-study highlighted that the uncertainty in QSAR predictions had low impact in the CTP assessment of triazoles. This issue has not been discussed in this thesis, but demonstrated the potential application of QSAR predictions within risk assessment procedures.

Concluding, the QSAR models developed within this thesis are useful tools to support hazard and risk assessment of specific classes of emerging pollutants, and show how non-testing information can be used for regulatory decisions, thus minimizing costs, time and saving animal lives.

Beyond their use for regulatory purposes, the here proposed QSARs can find application in the rational design of new safer compounds that are potentially less hazardous for human health and environment.

References

Ahlers J., Stock F. and Werschkun B. (2008) Integrated testing and intelligent assessment—new challenges under REACH. *Environmental Science and Pollution Research* 15, 565–572.

Alaee M., Arias P., Sjödin A., and Bergman Å. (2003) An overview of commercially used brominated flame retardants, their applications, their use patterns in different countries/regions and possible modes of release. *Environment International* 29, 683-689.

Anastas P. and Eghbali N. (2010) Green Chemistry: Principles and Practice. *Chemical Society Reviews* 39, 301–312.

Aoshima H., Hamamoto K. (1999) Potentiation of GABAA receptors expressed in Xenopus oocytes by perfume and phytoncid. *Bioscience, Biotechnology and Biochemistry* 63, 743–748.

Atkinson, A. C. (1985) Plots, Transformations and Regression, Clarendon Press, Oxford.

Beek B., Böhling S., Franke C., Jöhncke U., Studinger G., Thumm E. (2001) The assessment of biodegradation and persistence. From "The Handbook of Environmental Chemistry" Vol. 2 Part K. Springer-Verlag Berlin, Heidelberg (Germany).

Benfenati E., Gini G., Piclin N., Roncaglioni A., Vari M.R. (2003) Predicting logP of pesticides using different softwares. *Chemosphere* 53, 1155-1164.

Bergman A.A., Ostman C., Nybom R., Sjodin A., Carlsson H., Nilsson U., and Wachtmeister C.A. (1997) Flame retardants and plasticisers on particulate in the modern computerized indoor environment. *Organohalogen Compounds* 33, 414–419.

Bhhatarai B. and Gramatica P. (2011) Modelling physico-chemical properties of (benzo)triazoles, and screening for environmental partitioning. *Water Research* 45, 1463-1471.

Bhhatarai B., Teetz W., Liu T., Oberg T., Jeliazkova, N., Kochev N., Pukalov O., Tetko I., Kovarich S., Papa E. Gramatica P. (2011) CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Molecular informatics* (proceedings EuroQSAR2010) 30, 189-204.

Bickers D.R., Calow P., Greim H.A., Hanifin J.M., Rogers A.E., Saurat J-H., Sipes G., Smith R.L., Tagami H. (2003) The safety assessment of fragrance materials. *Regulatory Toxicology and Pharmacology* 37, 218–273.

Biffinger J. C., Kim H. W., and DiMagno S. G. (2004). The polar hydrophobicity of fluorinated compounds. *ChemBioChem* 5, 622–627.

Boethling R. S., Lynch D. G., Jaworska J. S., Tunkel J. L., Thom G. C., and Webb S. (2004). Using BIOWIN, Bayes and batteries to Predict Ready Biodegradability. *Environmental Toxicology & Chemistry* 23, 911–920.

Boethling R.S. (2011) Incorporating environmental attributes into musk design. *Green Chemistry* 13, 3386-3396.

Boethling R.S. and Lynch D.G. (2007) Biodegradation of US premanufacture notice chemicals in OECD tests. *Chemosphere* 66, 715–722.

Box G. E. P., Hunter W. G., and Hunter J. S. (1978) Statistics for experimenters: An introduction to design, data analysis, and model building. John Wiley & sons, New York.

Breedveld G.S., Roseth R., Sparrevik M., Hartnik T., Hem L.J. (2003) Persistence of the de-icing additive benzotriazole at an abandoned airport. *Water, Air, & Soil Pollution: Focus* 3, 91-101.

Bridges B. (2002) Fragrance: emerging health and environmental concerns. *Flavour and Fragrance Journal* 17, 361–371.

Brouwer A., Morse D.C., Lans M.C., Schuur A.G., Murk A.J., Klasson-Wehler E., Bergman Å., Visser T.J. (1998) Interactions of persistent environmental organohalogens with the thyroid hormone system: mechanisms and possible consequences for animal and human health. *Toxicology and Industrial Health.* 14, 59-84.

Burreau S., Zebuhr Y., Broman D., and Ishaq R. (2006) Biomagnification of PBDEs and PCBs in food webs from the Baltic Sea and the northern Atlantic ocean. *Science of the Total Environment* 366, 659–672.

Cancilla D., Martinez J., and Van Aggelen G. (1998) Detection of aircraft deicing/antiicing fluid additives in a perched water monitoring well at an international airport*. Environmental Science & Technology* 32, 3834-3835.

Carballa M., Omil F., Lema J.M., Llompart M., Garcia-Jares C., Rodriguez I., Gomez M., and Ternes T. (2004) Behavior of pharmaceuticals, cosmetics and hormones in a sewage treatment plant. *Water Research* 38, 2918-2926.

Chang S.-C., Thibodeaux J. R.,. Eastvold M. L, Ehresman D. J., Bjork J. A., Froehlich J. W., Lau C., Singh R. J., Wallace K. B., and Butenhoff J. L. (2008) Thyroid hormone status and pituitary function in adult rats given oral doses of perfluorooctane sulfonate (PFOS). *Toxicology* 243, 330–339.

Chen G., Konstantinov A. D., Chittim B. G., Joyce E. M., Bols N. C., and Bunce N. J. (2001) Synthesis of polybrominated diphenyl ethers and their capacity to induce CYP1A by the Ah Receptor mediated pathway. *Environmental Science & Technology* 35, 3749–3756.

Cheng F., Ikenaga Y., Zhou Y.,Yu Y., Li W., Shen J., Du Z., Chen L., Xu C., Liu G., Lee P.W., and Tang Y. 2012. In Silico Assessment of Chemical Biodegradability. *Journal of chemical Information and Modelling* 52, 655-69.

Chirico N. and Gramatica P. (2011) Real External Predictivity of QSAR Models: How To Evaluate It? Comparison of Different Validation Criteria and Proposal of Using the Concordance Correlation Coefficient. *Journal of Chemical Information and Modeling* 51, 2320–2335.

Chirico N. and Gramatica P. (2012) Real External Predictivity of QSAR Models. Part 2. New inter-comparable thresholds for different validation criteria and the need for scatter plot inspection. *Journal of Chemical Information and Modeling* 52, 2044– 2058.

Colborn T. (1995) Environmental estrogens: health implications for humans and wildlife. *Environmental Health Perspectives* 103, 135–136.

Colborn T., vom Saal F.S., Soto A.M. (1993) Developmental Effects of Endocrine-Disrupting Chemicals in Wildlife and Humans. *Environmental Health Perspective* 101, 378-384.

Consonni V.; Ballabio D.; Todeschini R.(2010) Evaluation of model predictive ability by external validation techniques. *Journal of chemometrics* 24, 194–201.

Consonni V., Ballabio D., and Todeschini R. (2009) Comments on the Definition of the Q^2 Parameter for QSAR Validation. *Journal of Chemical Information and Modeling* 49, 1669-1678.

Consonni V., Todeschini R., and Pavan M. (2002) Structure/response correlations and similarity/diversity analysis by GETAWAY descriptors 1. Theory of the novel 3D molecular descriptors. *Journal of Chemical Information and Computer Science* 42, 682–692.

Cooper J.A., Saracci R., Cole P. (1979) Describing the validity of carcinogen screening tests. *British Journal of Cancer* 39, 87–89.

Cruciani G., Baroni M., Clementi S., Costantino G., Riganelli D., Skagerberg B. (1992) Predictive ability of regression models. Part I: Standard deviation of prediction errors (SDEP). *Journal of Chemometrics* 6, 335-346.

Daughton C.G. (2005) "Emerging" Chemicals as Pollutants in the Environment: a 21st Century Perspective. *Renewable Resources Journal*. 23 (4), 6-23.

de Escobar G.M., Obregòn M.J., and del Rey F.E. (2004) Maternal thyroid hormones early in pregnancy and fetal brain development. *Best Practice & Research Clinical Endocrinology & Metabolism* 8, 225–248.

de Wit C.A. (2002) An overview of brominated flame retardants in the environment. *Chemosphere* 46, 583-624.

Dom N., Knapen D., Benoot D., Nobels I., and Blust R. (2010) Aquatic multi-species acute toxicity of (chlorinated) anilines: Experimental versus predicted data. *Chemosphere* 81, 177-186.

Dreyer A., Weinber I., Temme C., and Ebinghaus R. (2009) Polyfluorinated compounds in the atmosphere of the Atlantic and Southern Oceans: Evidence for a global distribution. *Environmental Science & Technology* 43, 6507–6514.

Durjava M.K., Kolar B., Arnus L., Papa E., Kovarich S., Sahlin U., Peijnenburg W. Experimental assessment of the environmental fate and effects of triazoles and benzotriazoles. *Submitted to ATLA* (as proceedings of the 2nd CADASTER Workshop in Munich, October 2012).

ECHA - European Chemical Agency (2008a) Guidance on information requirements and chemical safety assessment. Chapter R.6: QSARs and grouping of chemicals.

ECHA - European Chemical Agency (2008b) Guidance on information requirements and chemical safety assessment. Chapter R.7b: Endpoint specific guidance.

ECHA - European Chemical Agency (2008c) Guidance on information requirements and chemical safety assessment. Chapter R.11: PBT Assessment.

Eriksson L., Jaworska J., Worth A.P., Cronin M.T.D., McDowell R.M., and Gramatica P. (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environmental Health Perspectives 111, 1361–1375.

Eriksson P., Jakobsson E., and Fredriksson A. (2001) Brominated flame retardants: a novel class of developmental neurotoxicants in our environment? *Environmental Health Perspectives* 109, 903-908.

European Commission (1991). Directive 67/548/EEC, 1991 (Annex VI). General Classification and Labelling Requirements for Dangerous Substances and Preparations. Official Journal of the European Communities No. L180/45, 9/07/1991.

Fang H., Tong W., Perkins R., Soto A.M., Prechtl N.V., and Sheehan D.M. (2000). Quantitative Comparisons of in Vitro Assays for Estrogenic Activities. *Environmental Health Perspectives* 108, 723-729.

Fang H., Tong W., Branham W. S., Moland C. L., Dial S. L., Hong H., Xie Q., Perkins R., Owens W., and Sheehan D. M. (2003) Study of 202 natural, synthetic, and environmental chemicals for binding to the androgen receptor. *Chemical Research in Toxicology* 16, 1338 – 1358.

Fourches D., Muratov E., and Tropsha A. (2010) Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling* 50, 1189-1204.

Frank I.E. and Friedman J.H. (1989) Classification: oldtimers and newcomers. *Journal of Chemometrics* 3, 463-475.

Free S.M. and Wilson J.W. (1964) A mathematic contribution to Structure-Activity studies. *Journal of Medicinal Chemistry* 7, 395-399.

Fromme H., Tittlemier S.A., Völkel W., Wilhelm M., and Twardella D. (2009) Perfluorinated compounds – Exposure assessment for the general population in western countries. *International Journal of Hygiene and Environmental Health* 212, 239–270.

Gasteiger J. and Zupan J. (1993) Angewandte *Chemie International Edition(* English) 32, 503-527.

Giesy J. and Kannan K. (2001) Global distribution of perfluorooctane sulfonate in wildlife. *Environmental Science & Technology* 35, 1339–1342.

Giger W., Schaffner C., and Kohler H.P. (2006) Benzotriazole and tolyltriazole as aquatic contaminants. 1. Input and occurrence in rivers and lakes. *Environmental Science & Technology* 40, 7186-7192.

Gramatica P. (2001) QSAR Approach to the evaluation of Chemicals. *Chemistry Today* 18-24.

Gramatica, P. (2007) Principles of QSAR models validation: Internal and external. *QSAR and Combinatorial Science* 26, 694–701.

Gramatica P., Pilutti P. and Papa E. (2004) Validated QSAR Prediction of OH Tropospheric degradability: splitting into training-test set and consensus modeling. *Journal of Chemical Information and Computer Science* 44, 1794-1802.

Gramatica P., Cassani S., Roy P.P., Kovarich S., Yap C.W., Papa E. (2012). QSAR Modeling is not "Push a Button and Find a Correlation": A Case Study of Toxicity of (Benzo-)triazoles on Algae. *Molecular Informatics* 31, 817 – 835.

Gutshall D. M., Pilcher G. D., and Langley A. E. (1989) Mechanism of the serum thyroid hormone lowering effect of perfluoro-n-decanoic acid (PFDA) in rats. *Journal of Toxicology and Environmental Health* 28, 53–65.

Hale R.C., La Guardia M.J., Harvey E., Gaylor M.O., and Mainor, T.M. (2006) Brominated flame retardant concentrations and trends in abiotic media. *Chemosphere.* 64, 181-186.

Hamers T., Kamstra J.H., Sonneveld E., Murk A.J., Kester M.H.A., Andersson P.L., Legler J., and Brouwer A. (2006) *In vitro* profiling of the endocrine-disrupting potency of brominated flame retardants. *Toxicological Science* 92, 157-173.

Hamers T., Kamstra J. H., Sonneveld E., Murk A. J., Visser T. J., Van Velzen M. J. M., Brouwer A., and Bergman A. (2008) Biotransformation of brominated flame retardants into potentially endocrine-disrupting metabolites, with special attention to 2,2′,4,4′-tetrabromodiphenyl ether (BDE-47). *Molecular Nutrition & Food* Research 52, 284–298.

Hansch C., and Fujita T. (1964) ρ–σ–π analysis. A method for the correlation of biological activity and chemical structure. *Journal of American Chemical Society* 86, 1616-1626.

Hansch C., Maloney P.P., Fujita T., and Muir R.M. (1962) Correlation of biological activity of phenoxyacetic acids with Hammett Substituents Constants and partition coefficients. *Nature* 194, 178-180.

Harada K.H. and Koizumi A. (2009) Environmental and biological monitoring of persistent fluorinated compounds in Japan and their toxicities. *Environmental Health and Preventive Medicine* 14, 7–19.

Harju M., Hamers T., Kamstra J. H., Sonneveld E., Boon J. P., Tysklind M., and Andersson P. L. (2007) Quantitative structureactivity relationship modeling on in vitro endocrine effects and metabolic stability involving 26 selected brominated flame retardants. *Environmental Toxicology and Chemistry* 26, 816–826.

Harrad S., Wijesekera R., Hunter S., Halliwell C., Baker R. (2004) Preliminary Assessment of U.K. Human Dietary and Inhalation Exposure to Polybrominated Diphenyl Ethers. *Environmental Science & Technology* 38, 2345-2350.

Harrison P. T. C., Humfrey C. D. N., Litchfield M., Peakall D., and Shuker L. K. (1995) IEH Assessment on Environmental Oestrogens: Consequences to Human Health and Wildlife. Medical Research Council, Institute for Environment and Health, Page Bros., Norwich.

Helleday T., Touminem K.L., Bergman Å., Jenssen D. (1999) Brominated flame retardants induce intragenic recombination in mammalian cells. *Mutation Research* 439, 137-147.

Hem L., Hartnik T., Roseth R., and Breedveld, G. (2003) Photochemical Degradation of Benzotriazole. *Journal of Environmental Science and Health*, Part A, 38, 471-481.

Hirsch R., Ternes T., Haberer K., and Kratz K-L. (1999) Occurrence of antibiotics in the aquatic environment. The *Science of the Total Environment* 225, 109-118.

Holzapfel W. (1966) Uses of fluorinated surfactants. *Fette Seifen Anstrichm* 68, 837–842.

HorváthI.T. and Anastas P.T. (2007) Introduction: Green Chemistry. *Chemical Reviews* 107, 2167-2168.

Houde M., Martin J.W., Letcher R.J., Solomon K.R., and Muir D.C.G. (2006) Biological monitoring of polyfluoroalkyl substances: A review. *Environmental Science & Technology* 40, 3463–3473.

Hu J., Eriksson L., Bergman Å., Jakobsson E., Kolehmainen E., Knuutinen J., Suontamo R., Wei X. (2005) Molecular orbital studies on brominated diphenyl ethers. Part II – reactivity and quantitative structure–activity (property) relationships. *Chemosphere* 59, 1043–1057.

Hu K. and Bunce N. J. J. (1999) Metabolism of polychlorinated dibenzo-p-dioxins and related dioxin-like compounds. *Journal of Toxicology and Environmental Health, Part B.* 2, 183-210.

Ikonomou M.G., Rayne S., Addison R. (2002). Exponential increases of brominated flame retardants, polybrominated diphenylethers, in the Canadian Arctic from 1981–2000. *Environmental Science & Technology* 36, 1886–1892.

Jaworska J. S., Boethling R. S., and Howard P. H. (2003) Recent developments in broadly applicable structure-biodegradability relationships. *Environmental Toxicology and Chemistry* 2, 1710–1723.

Jaworska J., Comber M., Auer C., and van Leeuwen C. J. (2003) Summary of a Workshop on Regulatory Acceptance of (Q)SARs for Human Health and Environmental Endpoints. *Environmental Health Perspectives* 111, 1358 – 1360.

Jensen A.A. and Leffers H. (2008) Emerging endocrine disrupters: Perfluoroalkylated substances. *International Journal of Andrology* 31, 161–169.

Jensen T.K., Toppari J., Keiding N., and Skakkebaek N.E. (1995) Do environmental estrogens contribute to the decline in male reproductive health? *Clinical Chemistry* 41, 1896–1901.

Jirovetz L., Jager W., Buchbauer G., Nikiforov A., Raverdino V. (1991) Investigations of animal blood samples after fragrance drug inhalation by gas chromatography/mass spectrometry with chemical ionization and selected ion monitoring. *Biological Mass Spectrometry* 20, 801–803.

Jolliffe I. T. (1986) Principal Component Analysis. Springer-Verlag. pp. 487. doi:10.1007/b98835

Kadar E., Dashfield S., and Hutchinson T.H. (2010) Developmental toxicity of benzotriazole in the protochordate *Ciona intestinalis* (Chordata, Ascidiae). *Analytical and Bioanalytical Chemistry* 396, 641-647.

Kavlock, R.J. (1996) Research needs for risk assessment of health and environmental effects of endocrine disruptors: A review of the U.S. EPA-sponsored workshop. *Environmental Health Perspectives* 104, 715-740.

Kester M. H. A., Bulduk S., Van Toor H., Tibboel D., Meinl W., Glatt H., Falany C. N., Coughtrie M. W. H., Schuur A. G., Brouwer A., and Visser T. J. (2002) Potent inhibition of estrogen sulfotransferase by hydroxylated metabolites of polyhalogenated aromatic hydrocarbons reveals alternative mechanism for estrogenic activity of endocrine disrupters. *The Journal of Clinical Endocrinology and Metabolism. 87*, 1142–1150.

Kojima H., Katsura E., Takeuchi S., Niiyama K., and Kobayashi K. (2004) Screening for estrogen and androgen receptor activities in 200 pesticides by in vitro reporter gene assays using Chinese hamster ovary cells. *Environmental Health Perspectives* 112, 524–531.

Kolpin, D.W., Furlong,E.T., Meyer,M.T., Thurman,E.M., Zaugg,S.D., Barber,L.B., and Buxton,H.T. (2002) Pharmaceuticals, hormones, and other organic wastewater contaminants in US streams, 1999-2000: A national reconnaissance. *Environmental Science & Technology* 36, 1202-1211.

Kubat M. and Matwin S. (1997) Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the Fourteenth International Conference on Machine Learning (pp. 179-186). San Francisco, CA: Morgan Kaufmann.

Lau C., Anitole K., Hodes C., Lai D., Pfahles-Hutchens A., and Seed J. (2007) Perfluoroalkyl acids: A review of monitoring and toxicological findings. *Toxicological Science* 99, 366–394.

Law R. J., Alaee M., Allchin C. R., Boon J. P., Lebeuf M., Lepom P., and Stern G. A. (2003) Levels and trends of polybrominated diphenylethers and other brominated flame retardants in wildlife. *Environment International* 29, 757–770.

Law R.J., Allchin C.R., de Boer J., Covaci A., Herzke D., Lepom P., Morris S., a, Tronczynski J., and de Wit C.A. (2006) Levels and trends of brominated flame retardants in the European environment. *Chemosphere* 64, 187–208.

Le T.X. and Munekage,Y. (2004) Residues of selected antibiotics in water and mud from shrimp ponds in mangrove areas in Viet Nam. *Marine Pollution Bulletin* 49, 922-929.

Leardi R., Boggia R., and Terrile M. (1992) Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics* 6, 267–281.

Legler J. and Brouwer A. (2003) Are brominated flame retardants endocrine disruptors? *Environmental International* 29, 879–885.

Li J. and Gramatica P. (2010) Classification and virtual screening of androgen receptor antagonists *Journal of Molecular Graphics and Modelling* 50, 861–874.

Li J. and Gramatica P. (2010) The importance of molecular structures, endpoints' values, and predictivity parameters in QSAR research: QSAR analysis of a series of estrogen receptor binders. *Molecular Diversity* 14, 687–696.

Lilienthal H., Hack A., Roth-Harer A., Grande S.W., Talsness C.E. (2006) Effects of developmental exposure to 2,2',4,4',5-pentabromodiphenyl ether (PBDE-99) on sex steroids, sexual development, and sexually dimorphic behavior in rats. *Environmental Health Perspectives* 114, 194–201.

Lindgren F., Hansen B., Karcher W., Sjöström M., Eriksson L. (1996) Model validation by permutation tests: Applications to variable selection. *Journal of Chemomometrics* 10, 521-532.

Liu C., Deng J., Yu L.,. Ramesh M, and Zhou B. (2010) Endocrine disruption and reproductive impairment in zebrafish by exposure to 8:2 fluorotelomer alcohol. *Aquatic Toxicology* 96, 70–76.

Liu H., Papa E., Walker J.D., and Gramatica P. (2007) In silico screening of estrogen-like chemicals based on different nonlinear classification models. *Journal of Molecular Graphics and Modelling* 26, 135–144.

Loonen H., Lindgren F., Hansen B., Karcher W., Niemelä J., Hiromatsu K., Takatsuki L., Peijnenburg W., Rorije E., Struijs J. (1999) Prediction of biodegradability from chemical structure: modelling of ready biodegradation test data. *Environmental Toxicology and Chemistry* 18, 1763–1768.

Maloof M.A. (2003) Learning When Data Sets are Imbalanced and When Costs are Unequal and Unknown. Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC.

McNeill K.S. and Cancilla D.A. (2008) Detection of triazole deicing additives in soil samples from airports with low, mid, and large volume aircraft deicing activities. *Bulletin of Environmental Contamination and Toxicology* 70, 868-875.

Meerts I.A.T.M., Letcher R.J., Hoving S., Marsh G., Bergman A., Lemmen J.G., van der Burg B., and Brouwer A. (2001) *In vitro* estrogenicity of polybrominated diphenyl ethers, hydroxylated PBDEs, and polybrominated bisphenol A Compounds. *Environmental Health Perspectives* 109, 399-407.

Meerts I.A.T.M., van Zanden J.J., Luijks E.A.C., van Leeuwen-Bol I., Marsh G., Jakobsson E., Bergman A., Brouwer A. (2000) Potent competitive interactions of some brominated flame retardants and related compounds with human transthyretin in vitro. *Toxicological Science* 56, 95–104.

Meironyté Guvenius D., Bergman Å., and Norén K. (2001) Polybrominated Diphenyl Ethers in Swedish Human Liver and Adipose Tissue. *Archives of Environmental Contamination and Toxicology* 40, 564–570.

Moore D., Breton R., and MacDonald D. (2003) A Comparison of Model Performance for Six QSAR Packages that Predict Acute Toxicity to Fish. *Environmental Toxicology and Chemistry* 22, 1799-1809.

Moore M., Mustain M., Daniel K., Chen I., Safe S., Zacharewski T., Gillesby B., Joyeux A., and Balaguer P. (1997) Antiestrogenic Activity of Hydroxylated Polychlorinated Biphenyl Congeners Identified in Human Serum. *Toxicology and Applied Pharmacology* 142, 160–168.

Muehlbacher M., Kerdawy A.E., Kramer C., Hudson B., and Clark T. (2011) Conformation-dependent QSPR models: logPOW. . *Journal of Chemical Information and Modeling*, 51, 2408-2416.

Muir D. C. G., Backus S., Derocher A. E., Dietz R., Evans T. J., Gabrielsen G. W., Nagy J., Norstrom R. J., Sonne C., Stirling I., Taylor M. K., and Letcher R. J. (2006) Brominated flame retardants in polar bears (*Ursus maritimus)* from Alaska, the Canadian Arctic, East Greenland, and Svalbard. *Environmental Science & Technology* 40, 449–455.

Murk A.J., Legler J., Denison M.S., Giesy J.P., van de Guchte C., and Brouwer A. (1996) Chemical-activated luciferase gene expression (CALUX): a novel *in vitro* bioassay for Ah Receptor active compounds in sediments and pore water. *Fundamental and Applied Toxicology* 33, 149-160.

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., and Yang C. (2005) Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations of ECVAM Workshop 52. *Alternatives to Laboratory Animals (ATLA)* 33, 155–173.

Norén K. and Meironyté D. (1998) Contaminants in Swedish human milk. Decreasing levels of organochlorine and increasing levels of organobromine compounds. *Organohalogen Compounds* 38, 1–4.

OECD - Organisation for Economic Co-Operation and Development (2004) OECD Principles for the validation, for regulatory purposes, of (Quantitative) Structure-Activity Relationship Models. http://www.oecd.org/env/chemicalsafetyandbiosafety/assessmentofchemicals/37849783.pdf (Accessed 11/12/2012).

OECD - Organisation for Economic Co-Operation and Development (2007) Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] models. ENV/JM/MONO(2007)2. http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono%282007%292&doclanguage=en (Accessed 11/12/2012).

Okey A.B., Riddick D.S., Harper P.A. (1994) The Ah receptor: mediator of the toxicity of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and related compounds. *Toxicological Letters*, 70, 1-22.

Orellana M.A. (2006) Europe's Reach: A New Chapter in International Chemicals Law. *Sustainable Development Law & Policy* 6 (3), 21-28, 65.

Palm A., Cousins I.T., Mackay D., Tysklinnd M., and Alaee M. (2002) Assessing the environmental fate of chemicals of emerging concern: a case study of the polybrominated diphenyl ethers. *Environmental Pollution.* 117, 195-213.

Papa E., Villa F., and Gramatica P. (2005) Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemical in Pimephales promelas (Fathead Minnow), *Journal of Chemical Information and Modeling* 45, 1256-1266.

Papa E., Luini M., and Gramatica P. (2009) QSAR modelling of oral acute toxicity and cytotoxic activity of fragrance materials in rodents. *SAR & QSAR in Environmental Research* 20, 767–779.

Pavan M. and Worth A. P. (2008) Review of Estimation Models for Biodegradation. *QSAR and Combinatorial Science* 27, 32–40.

Pearson K. (1901) On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine* 2, 559–572.

Peters A.K., Nijmeijer S., Gradin K., Backlund M., Bergman Å., Poellinger L., Denison M.S., Van den Berg M. (2006) Interactions of polybrominated diphenyl ethers with the aryl hydrocarbon receptor pathway. *Toxicological Science* 92, 133–142.

Peterson S.A., Klabunde T., Lashuel H.A., Purkey H., Sacchettini J.C., and Kelly J.W. (1998) Inhibiting transthyretin conformational changes that lead to amyloid fibril formation. *Proceedings of the National Academy of Sciences USA* 95, 12956–12960.

Porcelli C., Boriani E., Roncaglioni A., Chana A., and Benfenati E. (2008) Regulatory perspectives in the use and validation of QSAR. A case study: DEMETRA model for Daphnia toxicity. *Environmental Science & Technology* 42, 491-496.

PPDB - Pesticide Properties DataBase (2009), http://sitem.herts.ac.uk/aeru/footprint/en/ (accessed 03.10.2012)

Prevedouros K. Cousins I. T., Buck R. C., and Korzeniowski S.H. (2006) Sources, fate and transport of perfluorocarboxylates. *Environmental Science & Technology* 40, 32–44.

QSPR-THESAURUS online platform, http://qspr-thesaurus.eu/login/show.do?render-mode=full (accessed 15.12.2012)

Renner R. (2002) The Kow Controversy. *Environmental Science & Technology* 36, 410A-413A.

Reuschenbach P., Silvani M., Dammann M., Warnecke D., and Knacker T. (2008) ECOSAR model performance with a large test set of industrial chemicals. *Chemosphere* 71, 1986-1995.

Richardson S.D. and Ternes TA. (2011) Water analysis: emerging contaminants and current issues. *Analytical Chemistry* 83 (12), 4614-48.

Roncaglioni A., Piclin N., Pintore M., Benfenati E. (2008) Binary classification models for endocrine disrupter effects mediated through the estrogen receptor. *SAR & QSAR in Environmental Research* 19, 697–733.

Roy P.P., Kovarich S., and Gramatica P. (2011) QSAR Model reproducibility and applicability: a case study of rate-constants of Hydroxyl radical reaction models applied to Polybrominated Diphenyl Ethers and (Benzo-) Triazoles. *Journal of Computational Chemistry* 32, 2386-2396.

Rücker C. and Kümmerer K. (2012) Modeling and predicting aquatic aerobic biodegradation – a review from a user's perspective. *Green Chemistry* 14, 875-887.

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

Salvito D.T., Vey M.G.H., and Senna R.J. (2004) Fragrance materials and their environmental impact. *Flavour and Fragrance Journal* 19, 105-108.

Schaafsma G., Kroese E.D., Tielemans E.L.J.P., Van de Sandt J.J.M., Van Leeuwen C.J. (2009) REACH, non-testing approaches and the urgent need for a change in mind set. *Regulatory Toxicology and Pharmacology* 53, 70–80.

Schüürmann G., Ebert R., Chen J., Wang B., and Kühne R. (2008) External Validation and Prediction Employing the Predictive Squared Correlation Coefficients Test Set Activity Mean vs Training Set Activity Mean. *Journal of Chemical Information and Modeling* 48, 2140–2145.

Sharaf M. A., Illman D. L., and Kowalski B.R. (1986) Chemometrics, Wiley Interscience, New York.

Shi L. M., Fang H., Tong, W., Wu J., Perkins R., Blair R. M., Branham W. S., Dial S. L., Moland C. L., and Sheehan D. M. (2001) QSAR Models Using a Large Diverse Set of Estrogens. *Journal of Chemical Information and Computer Science* 41, 186-195.

Shi Z., Zhang H., Ding L., Feng Y., Xu M., and Dai J. (2009) The effect of perfluorododecanonic acid on endocrine status, sex hormones and expression of steroidogenic genes in pubertal female rats. *Reproductive Toxicology* 27, 352–359.

Singh R., Artaza J.N., Taylor W.E., Braga M., Yuan X., Gonzalez-Cadavid N.F., and Bhasin S. (2006) Testosterone inhibits adipogenic differentiation in 3T3-L1 cells: nuclear translocation of androgen receptor complex with beta-catenin and T-cell factor 4 may bypass canonical Wnt signaling to down-regulate adipogenic transcription factors. *Endocrinology* 147, 141–54.

Sinha-Hikim I., Taylor W.E., Gonzalez-Cadavid N.F., Zheng W., Bhasin S. (2004) Androgen receptor in human skeletal muscle and cultured muscle satellite cells: up-regulation by androgen treatment. *The Journal of Clinical Endocrinology and Metabolism* 89, 5245–55.

Sjodin A., Carlsson H., Thuresson K., Sjolin S., Bergman A A., and Ostman C. (2001) Flame retardants in indoor air at an electronics recycling plant and at other work environments. *Environmental Science & Technology* 35, 448–454.

Sjödin A., Patterson D. G., and Bergman Å. (2003) A review on human exposure to brominated flame retardants - particularly polybrominated diphenyl ethers. *Environment International* 29, 829– 839.

Spencer P.S., Bischoff-Fenton M.C., Moreno O.M., Opdyke D.L., and Ford R.A. (1984) Neurotoxic properties of musk ambrette. *Toxicology and Applied Pharmacology* 75, 571–557.

Spencer P.S., Sterman A.B., Bischoff M., Horoupian D., and Foster G.V. (1978) Experimental myelin disease and ceroid accumulation produced by the fragrance compound acetyl ethyl tetramethyl tetralin. *Transactions of the American Neurological Association* 103, 185–187.

Stern G.A. and Ikonomou M.G. (2000). Temporal trends of polybrominated diphenyl ethers in SE Baffin beluga: increasing evidence of long range atmospheric transport. *Organohalogen Compounds* 47, 81–84.

Streets S.S., Henderson S.A., Stoner A.D., Carlson D.L., Simcik M.F., and Swackhamer D.L. (2006) Partitioning and bioaccumulation of PBDEs and PCBs in Lake Michigan. *Environmental Science & Technology.* 40, 7263-7269

Sun Y., Kamel M. S., Wong A.K.C., Wang Y. (2007) Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition* 40, 3358 – 3378.

Sushko I., Novotarskyi S., , Robert Körner R., Pandey A.K., Kovalishyn V.V., Prokopenko V.V., Tetko I.V. (2010) Applicability domain for in silico models to achieve accuracy of experimental measurements. *Journal of Chemometrics* 24, 202-208.

Taxvig C., Hass U., Axelstad M., Dalgaard M., Boberg J., Handeasen H.R., and Vinggaard A.M. (2007) Endocrine-disrupting activities *in vivo* of the fungicides tebuconazole and epoxiconazole. *Toxicological Science* 100, 464-473.

Thiele-Bruhn S. (2003) Pharmaceutical antibiotic compounds in soils - a review. *Journal of Plant Nutrition and Soil Science-Zeitschrift fur Pflanzenernahrung und Bodenkunde* 166, 145-167.

Todeschini R. and Consonni V. (2000) Handbook of Molecular Descriptors, Wiley-VCH, Weinheim (Germany).

Todeschini R., Maiocchi A., and Consonni V. (1999) The *K* correlation index: Theory development and its application in chemometrics. *Chemometrics and Intelligent Laboratory Systems* 46, 13 –29.

Todeschini, R. (1998) Introduzione alla chemiometria. EdiSES, Napoli.

Tomy G.T., Palace,V.P., Halldorson T., Braekevelt E., Danell R., Wautier K., Evans B., Brinkworth L., Fisk A.T. (2004) Bioaccumulation, biotrasformation and biochemical effects of brominated diphenyl ethers in Jouvenile Lake Trout (*Salvelinus namaycush*). *Environmental Science & Technology.* 38, 1496-1504.

Tropsha A. (2010) Best Practices for QSAR Model Development, Validation, and Exploitation *Molecular Informatics* 29, 476 – 488.

Tropsha A., Gramatica P., and Gombar V. K. (2003) The importance of being Earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR and Combinatorial Science 22, 69–77.

Tunkel J., Mayo K., Austin C., Austin C, Hickerson A., and Howard P. (2005) Practical Considerations of the Use of Predictive Methods for Regulatory Purposes. Environmental Science & Technology 39, 2188-2199.

Ucán-Marin F., Arukwe A., Mortensen A. S., Gabrielsen G. W., and Letcher R. J. (2010) Recombinant albumin and transthyretin transport proteins from two gull species and human: chlorinated and brominated contaminant binding and thyroid hormones. *Environmental Science & Technology* 44, 497-504.

van Leewuen C.J., Vermeire T.G. (2007). Risk Assessment of Chemicals: An introduction. 2nd Edition. Published by Springer, P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Vos J.G., Becher G., van den Berg M., de Boer J., Leonards P.E.G. (2003) Brominated flame retardants and endocrine disruption. *Pure and Applied Chemistry* 75, 2039–2046.

Wania F. and Dugani C.B. (2003) Assessing the long-range transport potential of polybrominated diphenyl ethers: a comparison of four multimedia models. *Environmental Toxicology and Chemistry* 22, 1252–1261.

Watanabe I., Kawano M., Wang T., Chen Y., Tatsukawa R. (1992) Polybrominated dibenzo-*p*-dioxins (PBDDs) and –dibenzofurans (PBDFs) in atmospheric air in Taiwan and Japan. *Organohalogen Compounds* 9, 309-312.

Weigel,S., Kuhlmann,J., and Huhnerfuss,H. (2002) Drugs and personal care products as ubiquitous pollutants: occurrence and distribution of clofibric acid, caffeine and DEET in the North Sea. *Science of the Total Environment* 295, 131-141.

Weiss J. M., Andersson P. L., Lamoree M. H., Leonards P. E. G., van Leeuwen S. P. J., and Hamers T. (2009) Competitive binding of poly- and perfluorinated compounds to the thyroid hormone transport protein transthyretin. *Toxicological Science* 109, 206–216.

Whitlock J. P. (1993) Mechanistic aspects of dioxin action. *Journal of Chemical Research in Toxicology* 6, 754-763.

Wollenberger L., Halling-Sorensen B., and Kusk K.O. (2000) Acute and chronic toxicity of veterinary antibiotics to Daphnia magna. *Chemosphere* 40, 723-730.

Wolschke H., Xie Z., Möller A., Sturm R., and Ebinghaus R. (2011) Occurrence, distribution and fluxes of benzotriazoles along the German large river basins into the North Sea. *Water Research* 45, 6259-6266.

Wu X., Chou N., Lupher D., and Davis L. C. (1998) Benzotriazoles: Toxicity and Degradation. In Conference on Hazardous Waste Research (pp. 374-382). Retrieved from https://www.engg.ksu.edu/HSRC/98Proceed/32Wu/32wu.pdf

Yang W., Mu Y., Giesy J.P., Zhang A., Yu H. (2009) Anti-androgen activity of polybrominated diphenyl ethers determined by comparative molecular similarity indices and molecular docking. *Chemosphere* 75, 1159-1164.

Young D., Martin T., Venkatapathy R., and Harten P. (2008) Are the chemical structures in your QSAR correct? *QSAR & Combinatorial Science* 2008, 27, 1337-1345.

Zhang J. and Mani I. (2003) kNN Approach to unbalanced data distribution: A case study involving information extraction. Workshop on Learning from Imbalanced Data Sets II, ICML, Washington DC.

Zhao Y-Y., Tao F-M., Zeng E.Y. (2008) Theoretical study on the chemical properties of polybrominated diphenyl ethers. *Chemosphere* 70, 901–907.

Zheng W. F. and Tropsha A. (2000) Novel Variable Selection Quantitative Structure-Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *Journal of Chemical Information and Computer Sciences* 40, 185–194.

Zhu H., Tropsha A., Fourches D.; Varnek A.; Papa E.; Gramatica P.; Oberg T., Phuong D., Cherkasov A., Tetko I.V. (2008) Combinatorial QSAR modeling of chemical toxicants tested against tetrahymena pyriformis. *Journal of Chemical Information and Modeling* 48, 766-784.

Software

- HYPERCHEM ver. 7.03 for Windows, 2002, Autodesk Inc., Sausalito, CA (USA).

- MOBY DIGS – MOdels BY Descriptors In Genetic Selection – ver. 1 beta for Windows (2004). Todeschini R., Ballabio D., Consonni V., Mauri A., and Pavan M. Talete S.r.l., Milan (Italy).

- DRAGON - Software for the Calculation of Molecular Descriptors - ver.5.5 (2007) Todeschini R., Consonni V., Mauri A., Pavan M., Talete srl, Milan (Italy). www.talete.mi.it

- PaDEL-Descriptor ver. 2.12. (2012). Ref: Yap C.W. (2011) PaDEL-Descriptor: an open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* 32, 1466-1474. Available online at http://padel.nus.edu.sg/software/padeldescriptor/index.html (accessed 15.12.2012)

- Open Babel ver. 2.3.0 (2010). Ref: O'Boyle N.M., , Banck M., James C.A., Morley C., Vandermeersch T., and Hutchison G.R. (2011) Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 3, 33.

- QSARINS, Software for QSAR MLR Model Development and Validation (2012) Chirico N., Papa E., Kovarich S., Cassani S., Gramatica P. University of Insubria, Varese, Italy (http://www.qsar.it).

- KOALA – KOhonen Artificial LAyers – ver. 1.1 (2001) Todeschini R., Milano Chemometrics and QSAR Research Group.

- SCAN – Software for Chemometric Analysis, ver. 1.1 for Windows (1995) Minitab (USA).

- ToxMatch ver. 1.06 (2008) Developed by Ideaconsult Ltd. for European Chemicals Bureau.

- ECOSAR Class Program ver. 1.11 (2012) Ecological Structure-Activity Relationship Model Class Program for the estimation of toxicity of industrial chemicals to aquatic organisms. United States Environmental Protection Agency (US EPA), Washington, DC, USA.
http://www.epa.gov/oppt/newchems/tools/21ecosar.htm (accessed 15.12.2012)

- EPI Suite - Estimation Programs Interface Suite - ver. 4.10 (2012) United States Environmental Protection Agency (US EPA), Washington, DC, USA.
http://www.epa.gov/oppt/exposure/pubs/episuite.htm (accessed 15.12.2012)

Acknowledgments

When I started my PhD "adventure" I couldn't imagine it would have impress my life so much, and now that I'm close to conclude this journey there is a long list of people I wish to say thanks…

My first and sincere gratitude goes to my supervisors, Prof. Paola Gramatica and Dr. Ester Papa, for their endless support and help, for everything they had taught me and, most important, for having taught me how to learn and to be a good scientist. Thanks for having shown me that, if we truly believe in something, we have to be ready to "fight" for it. Thanks for the many beautiful experiences and for all the opportunities you gave me over the past six years. Thanks for being so excellent guides, friends (when needed) and source of inspiration: I'm so proud to have worked, and grown, in your group!

I wish to thank all the fantastic people working, or that have worked in the these years, in the QSAR Unit at Insubria University: Ester (the boss), Prof. Gramatica (the big boss!!), Stefano (my twin), Nicola (the statistician), Lidia and Alessandra ("le molecoline"), the visiting scientists Partha, Barun and Giagio, and the colleagues of the past (but friends of the present) Mara, Elisa and Ale Moccia… I would never have come this far without the help and contribution of each of you! Thanks you all for showing me the importance to work in a team! Special thanks go to Leon, the expert ecotoxicologist of the group… thanks for your constructive criticism and advices, for giving us always a different point of view, and for all the time you have spent correcting my English..!

I would like to thank Prof. Palmisano, the coordinator of this doctoral school, for his kind availability during these three years, for all the interesting and useful lessons organized for the PhD students, and for making me feel always the welcome in his school of chemistry…!

I'm deeply grateful to the CADASTER Project, for the financial support, and to all the CADASTER Partners, and friends: Willie, Mojca, Igor, Stefan, Nina, Ullrika, Tomas, Magnus, Sara, Mark, Laura, Safraz… Thanks for giving me the opportunity to keep my mind open to the different facets of the fascinating world of ecotoxicology.

I wish to express my gratitude to the two groups that gave me "hospitality" during the months I spent abroad: Dr. Igor Tetko, and his research group (Igor, Stefan, Wolfram, Jury, Robert, Sergii, Ahmed and Eva), at the Helmholtz Zentrum München, and all the staff of WCA Environment in Oxford (Claire, Graham, Becky, Helen, Chris, Paul, Jhon, Rhiannon, Dawn, Ed, Dean, Adam, Peter, Iain,…). Thanks for everything I could learn from these two beautiful experiences and for making me feel at home.

I would like to say thanks to the friends and colleagues of the "red floor" and lunch times (Enrico, Viviana, Fabrizio, Andrea, Prof. Banfi and Prof. Barbieri, Zazza, Pupi, Matteo, Eleonora, Francesca, Bif, Luca, Marina, Annagiulia, and all the students passing from our labs during these years), as well as all my friends outside the university, both the old friends (Vale, Mario, Francy, Ceci, Nicky,…) and the new ones I met on the way during this long journey (Daniela, Viola, Elena, Jamie, the "ECO guys", and so on...): it was a great help to me to be able to forget this thesis during the breaks and my free time.

A deep thank goes also to the group of "founders ecotoxicologists" at Insubria University (Prof. Calamari, Prof. Di Guardo, Prof. Gramatica, Luca, Fofo, Sara, Francesco and Ester) and to the ARG group of "young" scientists of the Italian Chemical Society (Fabrizio, Andrea, Grazia, Lucia, Gianluigi, Pierluigi and, again, Ester): you have always been my muses and examples to follow.

Finally, I'm truly thankful to my love, Davide, and to my dear "big" family: my wonderful mum, Piera, my grandmother (nonna Nina), who left us just few months ago, my brother Gianluca and my little niece Francesca (if you are asking me again…yes, I'm at 100% now!), my uncles and my cousin Valentina (and of course the newcomer Federico!). A special thought also to whom has always guided me from "above". There are no words to express my earnest gratitude for all your support, for your patience, for all the time I couldn't spend with you because of this thesis, for your presence and closeness also when I was far away from home, for your faith in me and in my capabilities, for encouraging me in every situation, for letting me free to take my own decisions (also the wrong ones..!), for laying the bases of my education, and for making me build myself to become the person I am today. For all this and much more… ***Thank you!***

APPENDIX I

Table A-1. Datasets used for the development of regression and classification models (Note: pE(I)C50 = Log1/E(I)C50).

ID	MolID	LogRBA	pEC_{50} $EROD_{ind}$	pEC_{50} DR_{ag}	pEC_{50} ER_{ag}	pIC_{50} PR_{ant}	Log $T4_{REP}$	Log E_2SULT_{REP}	T4-TTR	E_2SULT	DR_{ag}	DR_{ant}	AR_{ant}	PR_{ant}	ER_{ag}	ER_{ant}
3	BDE-003	-3.89	--	--	--	--	--	--	--	--	--	--	--	--	--	--
15	BDE-015	-3.42	--	--	--	--	--	--	--	--	--	--	--	--	--	--
17	BDE-017	-3.64	--	--	--	--	--	--	--	--	--	--	--	--	--	--
19	BDE-019	--	--	--	-0.38	0.10	--	-1.26	1	2	1	1	2	2	2	1
28	BDE-028	-2.92	--	--	-1.18	-1.18	--	-1.82	1	2	1	2	2	2	2	1
38	BDE-038	--	--	0.30	-0.79	-1.18	-2.66	--	2	1	2	1	2	2	2	1
39	BDE-039	--	--	--	--	-1.18	--	--	1	1	1	2	2	2	1	1
47	BDE-047	-3.25	--	--	-1.08	-1.18	-2.66	-0.73	2	3	1	2	2	2	2	1
49	BDE-049	-4.17	--	-1.18	-0.96	-0.79	-2.66	-0.94	2	2	2	2	2	2	2	1
66	BDE-066	-2.70	0.36	--	--	--	--	--	--	--	--	--	--	--	--	--
71	BDE-071	-3.87	--	--	--	--	--	--	--	--	--	--	--	--	--	--
75	BDE-075	-3.40	--	--	--	--	--	--	--	--	--	--	--	--	--	--
77	BDE-077	-2.66	1.21	--	--	--	--	--	--	--	--	--	--	--	--	--
79	BDE-079	--	--	-0.15	-1.18	-1.18	--	--	1	1	2	1	2	2	2	2
85	BDE-085	-1.72	-0.15	--	--	--	--	--	--	--	--	--	--	--	--	--
99	BDE-099	-3.85	--	-1.18	--	-1.18	--	--	1	1	2	2	2	2	1	1
100	BDE-100	-4.11	-0.02	--	-0.85	-0.53	--	-2.00	1	2	1	2	2	2	2	1
119	BDE-119	-2.96	0.86	--	--	--	--	--	--	--	--	--	--	--	--	--
126	BDE-126	-2.57	1.47	--	--	--	--	--	--	--	--	--	--	--	--	--
127	BDE-127	--	--	--	--	-1.18	-2.60	-1.82	2	2	1	2	2	2	1	1
153	BDE-153	-4.60	-0.62	--	--	-0.76	--	--	1	1	2	1	2	2	1	1
154	BDE-154	-4.64	--	--	--	--	--	--	--	--	--	--	--	--	--	--
155	BDE-155	--	--	--	-0.91	-0.58	--	-2.00	1	2	1	1	2	2	2	1
169	BDE-169	--	--	--	--	--	-2.66	-1.82	2	2	1	2	1	1	1	1
181	BDE-181	--	--	-0.65	--	-0.56	-2.10	--	2	1	2	1	2	2	1	2
183	BDE-183	-3.60	-0.55	-0.30	--	-0.42	--	-1.82	1	2	2	1	2	2	1	2
185	BDE-185	--	--	--	--	-0.56	-2.13	--	2	1	1	1	2	2	1	2
190	BDE-190	--	--	-0.08	--	-0.68	-2.21	-1.40	2	2	2	1	2	2	1	2

ID	MolID	LogRBA	pEC_{50} $EROD_{ind}$	pEC_{50} DR_{ag}	pEC_{50} ER_{ag}	pIC_{50} PR_{ant}	Log $T4_{REP}$	Log E_2SULT_{REP}	T4-TTR	E_2SULT	DR_{ag}	DR_{ant}	AR_{ant}	PR_{ant}	ER_{ag}	ER_{ant}
206	BDE-206	--	--	--	--	--	--	-1.82	1	2	1	1	1	1	1	1
209	BDE-209	--	--	--	--	--	--	-1.82	1	2	1	1	1	1	1	1
214	TBBPA	--	--	--	--	--	0.25	0.97	3	3	1	1	1	1	1	1
215	BDE-215	--	--	--	--	-1.18	1.06	-0.26	3	3	1	1	2	2	1	2
216	246-TBP	--	--	-0.11	--	-0.70	-0.51	-0.97	3	3	2	1	2	2	1	2
221	6OH-BDE-47	--	--	--	--	-0.15	--	--	1	1	1	2	2	2	1	2
222	TBBPA-DBPE	--	--	--	--	--	-1.98	-0.26	2	3	1	1	1	1	1	1
224	4OH-BDE-42	--	--	--	--	--	0.54	0.81	3	3	--	--	--	--	--	--
225	3OH-BDE-47	--	--	--	--	--	0.60	0.81	3	3	--	--	--	--	--	--
226	5OH-BDE-47	--	--	--	--	--	0.48	0.15	3	3	--	--	--	--	--	--
227	4'OH-BDE-49	--	--	--	--	--	0.54	0.92	3	3	--	--	--	--	--	--
228	2'OH-BDE-66	--	--	--	--	--	-0.19	-1.08	3	2	--	--	--	--	--	--
	N_{tot}	**18**	**8**	**8**	**8**	**19**	**17**	**21**	**29**	**29**	**24**	**24**	**24**	**24**	**24**	**24**

Table A-2. List of 243 Brominated Flame Retardants (BFRs) studied within this thesis.

ID	Abbreviation	Name	CAS
1	BDE-001	2-monoBDE	036563-47-0
2	BDE-002	3-monoBDE	006876-00-2
3	BDE-003	4-monoBDE	000101-55-3
4	BDE-004	2,2'-diBDE	051452-87-0
5	BDE-005	2,3-diBDE	053563-56-7
6	BDE-006	2,3'-diBDE	147217-72-9
7	BDE-007	2,4-diBDE	171977-44-9
8	BDE-008	2,4'-diBDE	147217-71-8
9	BDE-009	2,5-diBDE	
10	BDE-010	2,6-diBDE	
11	BDE-011	3,3'-biBDE	006903-63-5
12	BDE-012	3,4-diBDE	189084-59-1
13	BDE-013	3,4'-diBDE	083694-71-7
14	BDE-014	3,5-diBDE	
15	BDE-015	4,4'-diBDE	002050-47-7
16	BDE-016	2,2',3-triBDE	
17	BDE-017	2,2',4-triBDE	147217-75-2
18	BDE-018	2,2',5-triBDE	
19	BDE-019	2,2',6-triBDE	
20	BDE-020	2,3,3'-triBDE	
21	BDE-021	2,3,4-triBDE	
22	BDE-022	2,3,4'-triBDE	
23	BDE-023	2,3,5-triBDE	
24	BDE-024	2,3,6-triBDE	
25	BDE-025	2,3',4-triBDE	147217-77-4
26	BDE-026	2,3',5-triBDE	
27	BDE-027	2,3',6-triBDE	
28	BDE-028	2,4,4'-triBDE	041318-75-6
29	BDE-029	2,4,5-triBDE	
30	BDE-030	2,4,6-triBDE	155999-95-4
31	BDE-031	2,4',5-triBDE	
32	BDE-032	2,4',6-triBDE	189084-60-4
33	BDE-033	2,3',4'-triBDE	147217-78-5

ID	Abbreviation	Name	CAS
123	BDE-123	2,3',4,4',5'-pentaBDE	
124	BDE-124	2,3',4',5,5'-pentaBDE	
125	BDE-125	2,3',4',5',6-pentaBDE	
126	BDE-126	3,3',4,4',5-pentaBDE	366791-32-4
127	BDE-127	3,3',4,5,5'-pentaBDE	
128	BDE-128	2,2',3,3',4,4'-hexaBDE	
129	BDE-129	2,2',3,3',4,5-hexaBDE	
130	BDE-130	2,2',3,3',4,5'-hexaBDE	
131	BDE-131	2,2',3,3',4,6-hexaBDE	
132	BDE-132	2,2',3,3',4,6'-hexaBDE	
133	BDE-133	2,2',3,3',5,5'-hexaBDE	
134	BDE-134	2,2',3,3',5,6-hexaBDE	
135	BDE-135	2,2',3,3',5,6'-hexaBDE	
136	BDE-136	2,2',3,3',6,6'-hexaBDE	
137	BDE-137	2,2',3,4,4',5-hexaBDE	446254-95-1
138	BDE-138	2,2',3,4,4',5'-hexaBDE	182677-30-1
139	BDE-139	2,2',3,4,4',6-hexaBDE	
140	BDE-140	2,2',3,4,4',6'-hexaBDE	243982-83-4
141	BDE-141	2,2',3,4,5,5'-hexaBDE	
142	BDE-142	2,2',3,4,5,6-hexaBDE	
143	BDE-143	2,2',3,4,5,6'-hexaBDE	
144	BDE-144	2,2',3,4,5',6-hexaBDE	
145	BDE-145	2,2',3,4,6,6'-hexaBDE	
146	BDE-146	2,2',3,4',5,5'-hexaBDE	
147	BDE-147	2,2',3,4',5,6-hexaBDE	
148	BDE-148	2,2',3,4',5,6'-hexaBDE	
149	BDE-149	2,2',3,4',5',6-hexaBDE	
150	BDE-150	2,2',3,4',6,6'-hexaBDE	
151	BDE-151	2,2',3,5,5',6-hexaBDE	
152	BDE-152	2,2',3,5,6,6'-hexaBDE	
153	BDE-153	2,2',4,4',5,5'-hexaBDE	068631-49-2
154	BDE-154	2,2',4,4',5,6'-hexaBDE	207122-15-4
155	BDE-155	2,2',4,4',6,6'-hexaBDE	035854-94-5

ID	Abbreviation	Name	CAS
34	BDE-034	2,3',5'-triBDE	
35	BDE-035	3,3',4-triBDE	147217-80-9
36	BDE-036	2,3',5-triBDE	
37	BDE-037	3,4,4'-triBDE	147217-81-0
38	BDE-038	3,4,5-triBDE	
39	BDE-039	3,4',5-triBDE	
40	BDE-040	2,2',3,3'-tetraBDE	
41	BDE-041	2,2',3,4-tetraBDE	
42	BDE-042	2,2',3,4'-tetraBDE	
43	BDE-043	2,2',3,5-tetraBDE	
44	BDE-044	2,2',3,5'-tetraBDE	
45	BDE-045	2,2',3,6-tetraBDE	
46	BDE-046	2,2',3,6'-tetraBDE	
47	BDE-047	2,2',4,4'-tetraBDE	005436-43-1
48	BDE-048	2,2',4,5-tetraBDE	
49	BDE-049	2,2',4,5'-tetraBDE	243982-82-3
50	BDE-050	2,2',4,6-tetraBDE	
51	BDE-051	2,2',4,6'-tetraBDE	
52	BDE-052	2,2',5,5'-tetraBDE	
53	BDE-053	2,2',5,6'-tetraBDE	
54	BDE-054	2,2',6,6'-tetraBDE	
55	BDE-055	2,3,3',4-tetraBDE	
56	BDE-056	2,3,3',4'-tetraBDE	
57	BDE-057	2,3,3',5-tetraBDE	
58	BDE-058	2,3,3',5'-tetraBDE	
59	BDE-059	2,3,3',6-tetraBDE	
60	BDE-060	2,3,4,4'-tetraBDE	
61	BDE-061	2,3,4,5-tetraBDE	
62	BDE-062	2,3,4,6-tetraBDE	
63	BDE-063	2,3,4',5-tetraBDE	
64	BDE-064	2,3,4',6-tetraBDE	
65	BDE-065	2,3,5,6-tetraBDE	
66	BDE-066	2,3',4,4'-tetraBDE	189084-61-5
67	BDE-067	2,3',4,5-tetraBDE	
68	BDE-068	2,3',4,5'-tetraBDE	

ID	Abbreviation	Name	CAS
156	BDE-156	2,3,3',4,4',5-hexaBDE	
157	BDE-157	2,3,3',4,4',5'-hexaBDE	
158	BDE-158	2,3,3',4,4',6-hexaBDE	
159	BDE-159	2,3,3',4,5,5'-hexaBDE	
160	BDE-160	2,3,3',4,5,6-hexaBDE	
161	BDE-161	2,3,3',4,5',6-hexaBDE	
162	BDE-162	2,3,3',4',5,5'-hexaBDE	
163	BDE-163	2,3,3',4',5,6-hexaBDE	
164	BDE-164	2,3,3',4',5',6-hexaBDE	
165	BDE-165	2,3,3',5,5',6-hexaBDE	
166	BDE-166	2,3,4,4',5,6-hexaBDE	189084-58-0
167	BDE-167	2,3',4,4',5,5'-hexaBDE	
168	BDE-168	2,3',4,4',5',6-hexaBDE	
169	BDE-169	3,3',4,4',5,5'-hexaBDE	
170	BDE-170	2,2',3,3',4,4',5-heptaBDE	327185-13-7
171	BDE-171	2,2',3,3',4,4',6-heptaBDE	
172	BDE-172	2,2',3,3',4,5,5'-heptaBDE	
173	BDE-173	2,2',3,3',4,5,6-heptaBDE	
174	BDE-174	2,2',3,3',4,5,6'-heptaBDE	
175	BDE-175	2,2',3,3',4,5',6-heptaBDE	
176	BDE-176	2,2',3,3',4,6,6'-heptaBDE	
177	BDE-177	2,2',3,3',4,5',6'-heptaBDE	
178	BDE-178	2,2',3,3',5,5',6-heptaBDE	
179	BDE-179	2,2',3,3',5,6,6'-heptaBDE	
180	BDE-180	2,2',3,4,4',5,5'-heptaBDE	
181	BDE-181	2,2',3,4,4',5,6-heptaBDE	189084-67-1
182	BDE-182	2,2',3,4,4',5,6'-heptaBDE	
183	BDE-183	2,2',3,4,4',5',6-heptaBDE	207122-16-5
184	BDE-184	2,2',3,4,4',6,6'-heptaBDE	
185	BDE-185	2,2',3,4,5,5',6-heptaBDE	
186	BDE-186	2,2',3,4,5,6,6'-heptaBDE	
187	BDE-187	2,2',3,4',5,5',6-heptaBDE	
188	BDE-188	2,2',3,4',5,6,6'-heptaBDE	
189	BDE-189	2,3,3',4,4',5,5'-heptaBDE	
190	BDE-190	2,3,3',4,4',5,6-heptaBDE	189084-68-2

ID	Abbreviation	Name	CAS
69	BDE-069	2,3',4,6-tetraBDE	327185-09-1
70	BDE-070	2,3',4',5-tetraBDE	
71	BDE-071	2,3',4',6-tetraBDE	189084-62-6
72	BDE-072	2,3',5,5'-tetraBDE	
73	BDE-073	2,3',5',6-tetraBDE	
74	BDE-074	2,4,4',5-tetraBDE	
75	BDE-075	2,4,4',6-tetraBDE	189084-63-7
76	BDE-076	2,3',4',5'-tetraBDE	
77	BDE-077	3,3',4,4'-tetraBDE	093703-48-1
78	BDE-078	3,3',4,5-tetraBDE	
79	BDE-079	3,3',4,5'-tetraBDE	
80	BDE-080	3,3',5,5'-tetraBDE	
81	BDE-081	3,4,4',5-tetraBDE	
82	BDE-082	2,2',3,3',4-pentaBDE	
83	BDE-083	2,2',3,3',5-pentaBDE	
84	BDE-084	2,2',3,3',6-pentaBDE	
85	BDE-085	2,2',3,4,4'-pentaBDE	182346-21-0
86	BDE-086	2,2',3,4,5-pentaBDE	
87	BDE-087	2,2',3,4,5'-pentaBDE	
88	BDE-088	2,2',3,4,6-pentaBDE	
89	BDE-089	2,2',3,4,6'-pentaBDE	
90	BDE-090	2,2',3,4',5-pentaBDE	
91	BDE-091	2,2',3,4',6-pentaBDE	
92	BDE-092	2,2',3,5,5'-pentaBDE	
93	BDE-093	2,2',3,5,6-pentaBDE	
94	BDE-094	2,2',3,5,6'-pentaBDE	
95	BDE-095	2,2',3,5',6-pentaBDE	
96	BDE-096	2,2',3,6,6'-pentaBDE	
97	BDE-097	2,2',3,4',5'-pentaBDE	
98	BDE-098	2,2',3,4',6'-pentaBDE	
99	BDE-099	2,2',4,4',5-pentaBDE	060348-60-9
100	BDE-100	2,2',4,4',6-pentaBDE	189084-64-8
101	BDE-101	2,2',4,5,5'-pentaBDE	
102	BDE-102	2,2',4,5,6'-pentaBDE	
103	BDE-103	2,2',4,5',6-pentaBDE	

ID	Abbreviation	Name	CAS
191	BDE-191	2,3,3',4,4',5',6-heptaBDE	
192	BDE-192	2,3,3',4,5,5',6-heptaBDE	
193	BDE-193	2,3,3',4',5,5',6-heptaBDE	
194	BDE-194	2,2',3,3',4,4',5,5'-octaBDE	085446-17-9
195	BDE-195	2,2',3,3',4,4',5,6-octaBDE	
196	BDE-196	2,2',3,3',4,4',5,6'-octaBDE	
197	BDE-197	2,2',3,3',4,4',6,6'-octaBDE	
198	BDE-198	2,2',3,3',4,5,5',6-octaBDE	
199	BDE-199	2,2',3,3',4,5,5',6'-octaBDE	
200	BDE-200	2,2',3,3',4,5,6,6'-octaBDE	
201	BDE-201	2,2',3,3',4,5',6,6'-octaBDE	
202	BDE-202	2,2',3,3',5,5',6,6'-octaBDE	
203	BDE-203	2,2',3,4,4',5,5',6-octaBDE	
204	BDE-204	2,2',3,4,4',5,6,6'-octaBDE	446255-54-5
205	BDE-205	2,3,3',4,4',5,5',6-octaBDE	
206	BDE-206	2,2',3,3',4,4',5,5',6-nonaBDE	
207	BDE-207	2,2',3,3',4,4',5,6,6'-nonaBDE	
208	BDE-208	2,2',3,3',4,5,5',6,6'-nonaBDE	876310-29-1
209	BDE-209	2,2',3,3',4,4',5,5',6,6'-decaBDE	001163-19-5
210	BPA	bisphenol-A	000080-05-7
211	MBBPA	3-monobromobisphenol-A	006073-11-6
212	DiBBPA	3,3-dibromobisphenol-A	029426-78-6
213	TriBBPA	3,3,5-tribromobisphenol-A	006386-73-8
214	TBBPA	3,5,3,5-tetrabromobisphenol-A	000079-94-7
215	246-TBP	2,4,6-tribromophenol	000118-79-6
216	6OH-BDE-47	6-OH-2,2',4,4'-tetrabromodiphenyl ether	079755-43-4
217	4-phenoxyphenol	4-phenoxyphenol	000831-82-3
218	4'-HO-BDE-30	4'-HO-BDE-30	
219	4'-HO-BDE-69	4'-HO-BDE-69	
220	4'-HO-BDE-121	4'-HO-BDE-121	
221	HBCD γ	hexabromocyclododecane γ	003194-55-6
222	TBBPA-DBPE	tetrabromobisphenol-A-bis(2,3)dibromopropyl ether	021850-44-2
223	HBB	hexabromobenzene	000087-82-1
224	4-OH-BDE-42	4-OH-2,2',3,4'-tetrabromodiphenyl ether	
225	3-OH-BDE-47	3-OH-2,2',4,4'-tetrabromodiphenyl ether	

ID	Abbreviation	Name	CAS
104	BDE-104	2,2',4,6,6'-pentaBDE	
105	BDE-105	2,3,3',4,4'-pentaBDE	373594-78-6
106	BDE-106	2,3,3',4,5-pentaBDE	
107	BDE-107	2,3,3',4',5-pentaBDE	
108	BDE-108	2,3,3',4,5'-pentaBDE	
109	BDE-109	2,3,3',4,6-pentaBDE	
110	BDE-110	2,3,3',4',6-pentaBDE	
111	BDE-111	2,3,3',5,5'-pentaBDE	
112	BDE-112	2,3,3',5,6-pentaBDE	
113	BDE-113	2,3,3',5',6-pentaBDE	
114	BDE-114	2,3,4,4',5-pentaBDE	
115	BDE-115	2,3,4,4',6-pentaBDE	
116	BDE-116	2,3,4,5,6-pentaBDE	189084-65-9
117	BDE-117	2,3,4',5,6-pentaBDE	
118	BDE-118	2,3',4,4',5-pentaBDE	
119	BDE-119	2,3',4,4',6-pentaBDE	189084-66-0
120	BDE-120	2,3',4,5,5'-pentaBDE	417727-71-0
121	BDE-121	2,3',4,5',6-pentaBDE	
122	BDE-122	2,3,3',4',5'-pentaBDE	

ID	Abbreviation	Name	CAS
226	5-OH-BDE-47	5-OH-2,2',4,4'-tetrabromodiphenyl ether	
227	4'-OH-BDE-49	4'-OH-2,2',4,5'-tetrabromodiphenyl ether	
228	2'-OH-BDE-66	2'-OH-2,3',4,4'-tetrabromodiphenyl ether	
229	2-OH-BDE-28	2'-OH-2,4,4'-triBDE	
230	6-OH-BDE-99	6-OH-2,2',4,4',5-pentaBDE	
231	6-CH_3O-BDE-47	6-CH_3O-2,2',4,4'-tetrabromodiphenyl ether	
232	4-BP	4-bromophenol	106-41-2
233	2,4,6-TBA	2,4,6-tribromoanisole	607-99-8
234	4'-CH_3O-BDE-49	4'-CH_3O-2,2',4,5'-tetrabromodiphenyl ether	
235	4'-OH-BDE-17	4'-OH-2,2',4-tribromodiphenyl ether	
236	2'-OH-BDE-68	2'-OH-2,3',4,5'-tetrabromodiphenyl ether	
237	6'-OH-BDE-49	6'-OH-2,2',4,5'-tetrabromodiphenyl ether	
238	6-OH-BDE-90	6-OH-2,2',3,4',5-pentabromodiphenyl ether	
239	PBP	pentabromophenol	608-71-9
240	TBBPA-DE	tetrabromobisphenol-A-diallyl ether	25327-89-3
241	DBDE	Decabromo Dipheyl Ethane	84852-53-9
242	EBTPI	ethylene bistetrabromo phthalimide	32588-76-4
243	TBPE	bis(tribromophenoxy) ethane	37853-59-1

Table A-3. Equations and performances of the MLR-OLS models developed for the prediction of ED potency of BFRs.

Endpoint	N_{TR}	Equations	R^2	Q^2_{LOO}	Q^2_{EXT}	$RMSE_{TR}$	$RMSE_P$
LogRBA	18	Y = -10.31(±0.92) + 0.79(±0.10) L1v + 8.89(±1.54) Mor22u	0.82	0.73	0.76*	0.31	0.42*
Log1/IC_{50}PRant	19	Y = -3.67(±0.51) + 0.01(±0.001) RDF045m + 2.69(±0.54) GATS4m	0.87	0.82	0.86*	0.14	0.15*
Log$T4_{REP}$	17	Y = -8.60 (±0.62) + 38.23(±3.03) qpmax + 2.89 (±0.80) MATS6v	0.94	0.91	0.90*	0.35	0.47*
Log$E2SULT_{REP}$	21	Y = -0.61 (±0.23) + 2.11(±0.19) B08[C-O] - 2.53 (±0.63) GGI7	0.88	0.84	0.88*	0.36	0.36*
Log1/EC_{50}ERODind	8	Y = 11.17 (±1.84) - 0.12 (±0.02) piID	0.85	0.75	--	0.29	--
Log1/EC_{50}DRag	8	Y = -0.27(±0.06) - 2.74(±0.35) Mor08e	0.91	0.85	--	0.15	--
Log1/EC_{50}ERag	8	Y = 0.99(±0.19) - 0.50(±0.05) RGyr	0.95	0.88	--	0.06	--

* Parameters calculated for the Split Models

Table A-4. Modelling descriptors and classification accuracy of *k*-NN models developed for the prediction of ED potency of BFRs.

Endpoint	Descriptors	*k*	Real class	N_{TR}	Assigned class			NER_{class}%	NER%	Sn	Sp
					1	2	3				
DR agonism	F04[O-Br] RDF055v	4	1	15	14	1		93.3	95.8	1	0.93
			2	9	0	9		100			
DR antagonism	Jhetm BEHm7	1	1	15	13	2		86.7	91.7	1	0.87
			2	9	0	9		100			
ER agonism	Ms BEHv7	1	1	16	15	1		93.7	95.8	1	0.94
			2	8	0	8		100			
ER antagonism	QW nArOH	1	1	16	15	1		93.7	95.8	1	0.94
			2	8	0	8		100			
AR/PR	GGI8	1	1	5	5	0		100	100	1	1
			2	19	0	19		100			
T4-TTR	DISPe nArOH	3	1	12	10	2	0	83.3	89.6	0.94	0.83
			2	9	1	8	0	88.9			
			3	8	0	0	8	100			
E2SULT inhibition	Mor21v qnmax	1	1	8	8	0	0	100	89.6	0.95	1
			2	12	1	10	1	83.3			
			3	9	0	1	8	88.9			

Table A-5. List of 57 Perfluorinated compounds (PFCs) studied within this thesis.

ID	MolID	Abbreviation	Name	SET
1	000307-24-4	PFHxA	Perfluorohexanoic acid	Training
2	000307-55-1	PFDoA	Perfluorododecanoic acid	Training
3	000335-67-1	PFOA	Perfluorooctanoic acid	Training
4	000335-76-2	PFDcA	Perfluorodecanoic acid	Training
5	000355-46-4	PFHxS(A)	Perfluorohexane sulfonic acid	Training/Validation
6	000375-22-4	PFBA	Perfluorobutyric acid	Training
7	000375-73-5	PFBS(A)	Nonafluorobutane sulfonic acid	Training/Validation
8	000375-85-9	PFHpA	Perfluoroheptanoic acid	Training
9	000375-95-1	PFNA	Perfluorononanoic acid	Training
10	000376-06-7	PFTdA	Perfluorotetradecanoic acid	Training
11	000647-42-7	FTOH (6:2)	2-Perfluorohexyl ethanol	Training
12	000678-39-7	FTOH (8:2)	2-Perfluorooctyl ethanol	Training
13	000754-91-6	FOSA	Perfluorooctane sulfonamide	Training
14	001546-95-8	7H-PFHpA	7H-Perfluoroheptanoic acid	Training
15	001691-99-2	N-EtFOSE	2-(N-ethylperfluoro-1-octane sulfonamido) ethanol	Training
16	001763-23-1	PFOS(A)	Perfluorooctane sulfonic acid	Training/Validation
17	002058-94-8	PFUnA	Perfluoroundecanoic acid	Training
18	004151-50-2	N-EtFOSA	N-ethyl perfluorooctane sulfonamide	Training
19	024448-09-7	N-MeFOSE	2-(N-methylperfluoro-1-octane sulfonamido) ethanol	Training
20	031506-32-8	N-MeFOSA	N-methyl perfluorooctane sulfonamide	Training
21	FTUA	FTUA (6:2)	2H-Perfluoro-2-octenoic acid (6:2)	Training
22	Me2FOSA	N,N-Me2FOSA	N,N-dimethyl perfluorooctane sulfonamide	Training
23	PFDS-A	L-PFDS(A)	Perfluorodecane sulfonic acid	Training/Validation
24	PFOSi-A	L-PFOSi(A)	Perfluorooctane sulfinic acid	Training/Validation
25	000076-21-1		Hexadecafluoro-nonanoic acid	Unknown
26	000336-08-3		Hexanedioic acid, 2,2,3,3,4,4,5,5-octafluoro-	Unknown
27	000355-80-6	FTOH(4:1)	1-Pentanol, 2,2,3,3,4,4,5,5-octafluoro-	Unknown
28	000356-27-4		Butanoic acid, heptafluoro-, ethyl ester	Unknown
29	000375-81-5		Perfluoropentane-1-sulphonyl fluoride	Unknown
30	000375-92-8	PFHpS	Pentadecafluoro-1-heptanesulfonic acid	Unknown
31	000376-53-4		Adiponitrile, perfluoro	Unknown
32	000376-72-7		Octafluoropentanoic acid	Unknown
33	000376-89-6		Hexafluoroglutaronitrile	Unknown

ID	MolID	Abbreviation	Name	SET
34	000377-38-8		Perfluorosuccinic acid	Unknown
35	000423-50-7		Perfluorohexanesulphonyl fluoride	Unknown
36	000423-54-1		Perfluorooctanamide	Unknown
37	000559-11-5		1,1-Dihydroperfluoroheptyl acrylate	Unknown
38	000756-91-2		3-Penten-1,5-diol, 3-methyl-1,1,5,5-tetrakis(trifluoromethyl)-	Unknown
39	000865-86-1	10:2 FTOH	1,1,2,2-Tetrahydroperfluoro dodecanol	Unknown
40	000918-21-8		Perfluoropinacol	Unknown
41	001478-61-1		Hexafluoroacetone bisphenol A	Unknown
42	001765-48-6		11-Eicosafluoroundecanoic acid	Unknown
43	002043-55-2		Hexane, 1,1,1,2,2,3,3,4,4-nonafluoro-6-iodo-	Unknown
44	002706-90-3	PFPeA	Perfluoropentanoic acid	Unknown
45	002706-91-4	PFPeS	Perfluoropentanesulfonic acid	Unknown
46	013695-31-3		Heptafluorobutyl methacrylate	Unknown
47	017527-29-6		1,1,2,2-Tetrahydroperfluorooctyl acrylate	Unknown
48	034449-89-3		1-Butanesulfonamide, N-ethyl-1,1,2,2,3,3,4,4,4-nonafluoro-N-(2-hydroxyethyl)-	Unknown
49	039239-77-5	12:2 FTOH	1,1,2,2-Tetrahydroperfluoro-1-tetradecanol	Unknown
50	057729-98-3		Benzenamine, N-(2,4-difluorophenyl)-2,4-dinitro-6-(trifluoromethyl)-	Unknown
51	065592-51-0		Benzenamine, N-methyl-2,4-dinitro-6-(trifluoromethyl)-N-(3-(trifluoromethyl)phenyl)-	Unknown
52	067939-33-7		2-Propenoic acid, 2-methyl-, 2- ethyl (nonafluorobutyl)sulfonyl amino ethyl ester)	Unknown
53	068259-12-1	PFNS	Nonadecafluoro-1-nonanesulfonic acid	Unknown
54	107350-42-5		Cyclohexanecarboxamide, 1,2,2,3,3,4,4,5,5,6,6-undecafluoro-N-(2,3,4,5-tetrachlorophenyl)-	Unknown
55	XXX001	PFTriA	Perfluorotridecanoic acid	Unknown
56	XXX002	PFPA	Perfluoropentadecanoic acid	Unknown
57	002043-47-2	FTOH(4:2)	2-Perfluorobutylyl ethanol	Unknown

APPENDIX II

APPENDIX II

Table A-6. List of 386 triazoles and benzotriazoles (B-TAZs) studied within this thesis.

ID	CAS	Name
1	000061-82-5	Amitrole
2	000094-97-3	1H-Benzotriazole, 6-chloro-
3	000095-14-7	Benzotriazole
4	000130-34-7	2-(p-nitrophenyl)-2H-naphtho[1,2-d]triazole-6,8-disulphonic acid
5	000131-43-1	EINECS 205-022-6
6	000134-58-7	8-Azaguanine
7	000136-85-6	6-methylbenzotriazole
8	000273-40-5	8-Azapurine
9	000288-36-8	1H-1,2,3-Triazole
10	000288-88-0	1,2,4-Triazole
11	000584-13-4	4H-1,2,4-Triazole, 4-amino-
12	000932-64-9	5-Nitro-1,2,4-triazol-3-one
13	000938-56-7	1H-Benzotriazole-1-ethanol
14	000939-07-1	4H-1,2,4-Triazol-5-ol, 2-phenyl-
15	000939-08-2	Pyridine, 4-(3-hydroxy-4H-1,2,4-triazol-5-yl)-
16	000944-91-2	3H-1,2,4-Triazol-3-one, 2,4-dihydro-2,4-dimethyl-5-phenyl-
17	000947-85-3	3H-1,2,4-Triazol-3-one, 2,4-dihydro-4-ethyl-2-methyl-5-phenyl-
18	000974-29-8	Stannane, (1H-1,2,4-triazol-1-yl)triphenyl-
19	001028-08-6	Phosphonic diamide, (5-amino-3-pentyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
194	075020-35-8	(1)Benzopyrano(2,3-d)-1,2,3-triazol-9(1H)-one, 6,7-dimethyl-
195	075736-33-3	Diclobutrazol
196	076608-88-3	triapenthenol (Ref: NTN 811)
197	076674-21-0	Flutriafol
198	076738-62-0	Paclobutrazol
199	077314-77-3	3-Amino-5-(2-(ethylamino)-4-pyridyl)-1,2,4-triazole
200	078149-96-9	Phosphonic amide, (5-amino-3-pentyl-1H-1,2,4-triazol-1-yl)-N,N-dimethyl-, propyl ester
201	078150-00-2	Phosphonic diamide, (5-amino-3-benzyl-1H-1,2,4-triazol-1-yl)-N,N'-dimethyl
202	078150-02-4	WP 254
203	078218-51-6	Phosphonic diamide, (5-amino-3-p-chlorophenyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
204	078218-52-7	Phosphonic diamide, (5-amino-3-ethyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
205	078218-53-8	Phosphonic diamide, (5-amino-3-heptyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
206	078218-54-9	Phosphonic diamide, (5-amino-3-isopropyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
207	078218-55-0	Phosphonic diamide, (5-amino-3-p-methoxyphenyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
208	078218-56-1	Phosphonic diamide, (5-amino-3-methylbenzylidene-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
209	078218-57-2	Phosphonic diamide, (5-amino-3-methyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
210	078218-58-3	Phosphonic diamide, (5-amino-3-phenyl-1H-1,2,4-triazol-1-yl)N,N'-dimethyl
211	078218-59-4	Phosphonic diamide, (5-amino-3-propyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
212	078218-60-7	Phosphonic diamide, (5-amino-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetraethyl

ID	CAS	Name
20	001031-47-6	Triamiphos
21	001123-54-2	8-Azaadenine
22	001325-58-2	(1,2-ethenediyl)bis[5-nitrobenzenesulfonic acid], reduced
23	001326-66-5	C.I. Sulphur Yellow 2
24	001455-77-2	Guanazole
25	001468-26-4	8-Azaxanthine
26	001600-61-9	delta(sup 2)-1,2,4-Triazolin-5-one, 1-methyl-3-(5-nitro-2-furyl)-
27	001680-44-0	1H-1,2,3-Triazole, 4-phenyl-
28	001704-66-1	Acetamide, N-(3-(5-nitro-2-furyl)-s-triazol-5-yl)-
29	002338-12-7	1H-Benzotriazole, 6-nitro-
30	002440-22-4	Drometrizole
31	002592-95-2	1-Hydroxybenzotriazole
32	002683-90-1	8-Azahypoxanthine
33	003142-42-5	2-(2H-benzotriazol-2-yl)-4-dodecylphenol
34	003147-75-9	Octrizole [USAN:INN]
35	003147-76-0	Phenol, 2-(2H-benzotriazol-2-yl)-4-(1,1-dimethylethyl)-
36	003232-84-6	Urazole
37	003310-68-7	N-phenyl-1H-1,2,4-triazole-3,5-diamine
38	003333-62-8	7-(2H-naphtho[1,2-d]triazol-2-yl)-3-phenyl-2-benzopyrone
39	003357-42-4	1H-1,2,4-Triazole, 5-phenyl-
40	003641-10-9	3-Cyano-1,2,4-triazole
41	003652-22-0	Pyridine, 4-(3-ethylthio-5-(4H-1,2,4-triazolyl))-
42	003652-23-1	Pyridine, 4-(3-propylthio-5-(4H-1,2,4-triazolyl))-
43	003652-25-3	Pyridine, 4-(3-allylthio-5-(4H-1,2,4-triazolyl))-

ID	CAS	Name
213	078218-61-8	Phosphonic diamide, (5-amino-3-undecyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
214	078218-65-2	Phosphonothioic diamide, (5-amino-3-methyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
215	078218-66-3	Phosphonothioic diamide, (5-amino-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
216	078324-76-2	1H-anthra[2,3-d]triazole-5,10-dione
217	078371-72-9	1H-1,2,4-Triazole-3-carboxylic acid, 5-amino-1-(N,N,N',N'-tetramethyldiaminophosphinyl), ethyl ester
218	078371-73-0	1H-1,2,4-Triazole-3-carboxylic acid, 5-amino-1-(N,N,N',N'-tetramethyldiaminophosphinyl), pentyl ester
219	078371-74-1	1H-1,2,4-Triazole-3-carboxylic acid, 5-amino-1-(N,N,N',N'-tetramethyldiaminophosphinyl), propyl ester
220	078592-90-2	Phosphonic diamide, (5-amino-3-isobutyl-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
221	079983-71-4	Hexaconazole [BSI:ISO]
222	080301-64-0	1H-Benzotriazole-1-methanamine, N,N-bis(2-ethylhexyl)-
223	080584-88-9	-[[(5-methyl-1H-benzotriazol-1-yl)methyl]imino]bisethanol
224	080584-89-0	-[[(4-methyl-1H-benzotriazol-1-yl)methyl]imino]bisethanol
225	080584-90-3	N,N-bis(2-ethylhexyl)-4-methyl-1H-benzotriazole-1-methylamine
226	080595-74-0	N,N-bis(2-ethylhexyl)-5-methyl-1H-benzotriazole-1-methylamine
227	081518-26-5	
228	081518-27-6	
229	081518-28-7	
230	081518-29-8	
231	081518-31-2	
232	081518-32-3	
233	081518-37-8	Phenol, 2-(5-(butylthio)-4-phenyl-4H-1,2,4-triazol-3-yl)-
234	081518-41-4	Phenol, 2-(4-(4-bromophenyl)-5-(propylthio)-4H-1,2,4-triazol-3-yl)-
235	082200-72-4	1H-1,2,4-Triazole-1-ethanol, beta-(4-chlorophenoxy)-alpha-(1,1-dimethyl)-,(alphaR,betaR)-rel-
236	083044-89-7	octyl 3-[3-tert-butyl-4-hydroxy-5-(5-chloro-2H-benzotriazol-2-yl)phenyl]propionate

ID	CAS	Name
44	003652-27-5	Pyridine, 4-(3-(2-pyridylmercapto)-5-(4H-1,2,4-triazolyl))-
45	003652-31-1	Pyridine, 3-(3-mercapto-4-methyl-5-(4H-1,2,4-triazolyl))-
46	003652-32-2	Pyridine, 4-(3-mercapto-4-methyl-5-(4H-1,2,4-triazolyl))-
47	003663-24-9	Butylbenzotriazole
48	003683-95-2	1,2,4(H)-Triazole, 3-mercapto-5-(4-bromo-3-methylisothiazol-5-yl)-
49	003770-47-6	Pyridine, 4-(3-(methylthio)-5(4H)-1,2,4-triazol-5-yl)-
50	003846-71-7	2-benzotriazol-2-yl-4,6-di-tert-butylphenol
51	003864-99-1	2,4-di-tert-butyl-6-(5-chlorobenzotriazol-2-yl)phenol
52	003896-11-5	Bumetrizole [USAN:INN]
53	004184-79-6	5,6-dimethyl-1H-benzotriazole
54	004314-22-1	1H-1,2,4-triazol-1-ylacetic acid (Ref: CGA 142856)
55	004343-73-1	ethyl 5-methyl-1H-1,2,3-triazole-4-carboxylate
56	004368-68-7	1-Benzyl-1,2,3-triazole
57	004928-87-4	1,2,4-triazole-5-carboxylic acid
58	004928-88-5	1,2,4-Triazole-3-Carboxylic Acid Methylester
59	005302-27-2	A 10749
60	005369-84-6	4H-1,2,4-Triazole, 3-amino-4-(2-(2,6-xylyloxy)ethyl)-
61	005472-71-9	1-(4-morpholinomethyl)-1H-benzotriazole
62	005516-20-1	2-[2-(4-chlorophenyl)vinyl]-5-(2H-naphtho[1,2-d]triazol-2-yl)benzonitrile
63	005873-30-3	C.I. Direct Violet 72
64	006054-53-1	6-[(2-chloro-4-nitrophenyl)azo]-1,2,3,4-tetrahydrobenzo[h]quinoline-3,7-diol
65	006085-94-5	1H-1,2,4-Triazole, 1-benzyl-
66	006299-39-4	4-Nitro-1H-benzotriazole
67	006789-99-7	4,5,6,7-tetrahydro-1H-benzotriazole
68	006818-99-1	3-chloro-1H-1,2,4-triazole
69	006994-51-0	phenyl 4-(2H-naphtho[1,2-d]triazol-2-yl)stilbene-2-sulphonate

ID	CAS	Name
237	083044-90-0	2-ethylhexyl 3-[3-tert-butyl-4-hydroxy-5-(5-chloro-2Hbenzotriazol-2-yl)phenyl]propionate
238	083044-91-1	methyl 3-[3-tert-butyl-4-hydroxy-5-(5-chloro-2Hbenzotriazol-2-yl)phenyl]propionate
239	083366-66-9	Nefazodone [INN:BAN]
240	083657-17-4	Uniconazole-P [ISO]
241	083657-24-3	Diniconazole [ISO]
242	085509-19-9	Flusilazole [ISO]
243	085634-51-1	1H-Benzotriazole, 1-(cyclohexylcarbonyl)-5-(1,4,5,6-tetrahydro-4-methyl-6-oxo-3-pyridazinyl)-
244	086386-73-4	Fluconazole
245	086598-92-7	Imibenconazole
246	088671-89-0	Myclobutanil [ANSI:BSI:ISO]
247	089482-17-7	Triadimenol A
248	089786-04-9	Tazobactam [USAN:INN:BAN]
249	094270-86-7	1H-Benzotriazole-1-methanamine, N,N-bis(2-ethylhexyl)-ar-methyl-
250	094361-06-5	Cyproconazole
251	094667-47-7	1H-benzotriazolesulphonic acid
252	097232-75-2	(R)-glycolic acid, compound with (S)-1,2,3,4,5,6,7,8-octahydro-1-[(4-methoxyphenyl)methyl]isoquinoline (1:1)
253	098518-95-7	
254	098518-96-8	
255	098518-99-1	
256	098519-00-7	
257	098519-01-8	
258	098519-02-9	
259	098519-04-1	
260	098519-05-2	
261	098519-06-3	
262	098519-07-4	

ID	CAS	Name
70	007170-01-6	1H-1,2,4-Triazole, 5-methyl
71	007411-23-6	s-Triazole, 3,5-dibromo-
72	007532-52-7	s-Triazole, 5-amino-3-(5-nitro-2-furyl)-
73	010109-05-4	1,2,4-triazolyl-3-alanine
74	010187-79-8	Acetamide, N-(1-methyl-3-(5-nitro-2-furyl)-s-triazol-5-yl)-
75	010187-84-5	s-Triazole, 5-methylamino-3-(5-nitro-2-furyl)-
76	010187-86-7	s-Triazole, 3-amino-4-methyl-5-(5-nitro-2-furyl)-
77	010187-89-0	s-Triazole, 5-ethylamino-3-(5-nitro-2-furyl)-
78	013091-80-0	6-chloro-4-nitro-1H-benzotriazole
79	013257-88-0	1-(trimethylsilyl)-1H-1,2,3-triazole
80	013351-73-0	1H-Benzotriazole, 1-methyl-
81	014803-99-7	1H-1,2,4-Triazole, 5-(4-pyridyl)-
82	015421-84-8	Trapidil
83	015497-45-7	N,N-Dibutyl-1H-benzotriazole-1-methylamine
84	015805-10-4	1H-Benzotriazole, 6,6'-methylenebis-
85	016515-58-5	7-(5-butoxy-6-methyl-2H-benzotriazol-2-yl)-3-phenyl-2-benzopyrone
86	016584-05-7	1H-Benzotriazole, 1-ethyl-
87	018076-61-4	1H-benzotriazol-4-amine
88	018811-70-6	1-(1H-benzotriazol-5-yl)-3-phenylurea
89	019683-09-1	2H-1-Benzopyran-2-one, 7-(4-methyl-5-phenyl-2H-1,2,3-triazol-2-yl)-3-phenyl-
90	019794-93-5	Trazodone
91	021050-95-3	1-Chloro-benzotriazole
92	021532-04-7	4H-1,2,4-Triazole, 4-amino-3,5-dimethylthio-
93	023633-05-8	s-Triazole, 3-(hydroxymethylamino)-4-methyl-5-(5-nitro-2-furyl)-
94	023711-34-4	
95	024017-47-8	Triazophos [BSI:ISO]
96	024054-57-7	s-Triazole, 5-(hydroxymethylamino)-1-methyl-3-(5-nitro-2-furyl)-
97	025973-55-1	Phenol, 2-(2H-benzotriazol-2-yl)-4,6-bis(1,1-dimethylpropyl)-
98	026621-45-4	1-Methyl-3-nitro-1,2,4-triazole
99	027022-50-0	Abbott 40060

ID	CAS	Name
263	098519-24-5	
264	098519-25-6	
265	098519-26-7	
266	098519-28-9	
267	098519-29-0	
268	098519-30-3	
269	098519-31-4	
270	098519-32-5	
271	098519-33-6	
272	098519-34-7	
273	098519-35-8	
274	098519-37-0	
275	098519-39-2	
276	098519-41-6	
277	098519-43-8	
278	098519-49-4	
279	098532-64-0	
280	098532-65-1	
281	098532-66-2	
282	098532-67-3	
283	098532-68-4	
284	098532-69-5	
285	098532-70-8	
286	098532-71-9	
287	098532-72-0	
288	098532-73-1	
289	098532-74-2	
290	098532-75-3	
291	098532-77-5	
292	098532-80-0	

ID	CAS	Name
100	027210-18-0	1,2,4-Triazolo(3,4-a)isoquinoline, 8,9-dimethoxy-
101	027799-91-3	5-methoxy-1H-benzotriazole
102	028401-89-0	1,2,4-Triazolo(3,4-a)isoquinoline, 8,9-dimethoxy-3-methyl-
103	028911-01-5	Triazolam
104	028981-97-7	Alprazolam
105	029440-31-1	Phosphonic diamide, (5-amino-1H-1,2,4-triazol-1-yl)-N,N,N',N'-tetramethyl
106	029878-31-7	1H-Benzotriazole, 4-methyl-
107	029975-16-4	Estazolam
108	031251-03-3	fluotrimazole
109	031409-18-4	4H-1,2,4-Triazole, 3-(4-chlorophenyl)-4-ethyl-5-(methylthio)-
110	031701-42-5	1,4-Benzenediol, 2-(2H-benzotriazol-2-yl)-
111	032362-89-3	1H-1,2,4-Triazole-3-thiol, 5-(2-pyridyl)-
112	032723-50-5	1,2,4-Triazolo(4,3-b)pyridazine, 3,8-dimethyl-6-phenyl-
113	034771-66-9	2H-1-Benzopyran-2-one, 3-(4-chloro-1H-pyrazol-1-yl)-7-(4-methyl-5-phenyl-2H-1,2,3-triazol-2-yl)
114	035515-45-8	1,2,4-Triazolo(3,4-a)isoquinoline, 5,6-dihydro-3-(trifluoromethyl)-
115	036325-69-6	N-(2-(2-Hydroxyphenyl)-2H-benzotriazol-5-yl)methacrylamide
116	036411-52-6	3-(N-Salicyloyl)amino-1,2,4-triazole
117	036437-37-3	Phenol, 2-(2H-benzotriazol-2-yl)-4-(1,1-dimethylethyl)-6-(1-methylpropyl)
118	036791-04-5	Ribavirin [USAN:INN]
119	037160-06-8	1,2,4-Triazolo(4,3-b)(1,2,4)triazine, 6,7-diphenyl-
120	038942-51-7	4H-1,2,4-Triazole-3-thiol, 4-methyl-5-phenyl-
121	039968-33-7	3H-1,2,3-Triazolo[4,5-b]pyridine, 3-hydroxy-
122	040054-69-1	Etizolam
123	041083-11-8	Azocyclotin
124	041735-28-8	s-Triazole, 5-(N-methyl-N-nitroso)amino-3-(5-nitro-2-furyl)-
125	041735-29-9	s-Triazole, 5-(N-ethyl-N-nitroso)amino-3-(5-nitro-2-furyl)-

ID	CAS	Name
293	098532-81-1	
294	098532-82-2	
295	098532-83-3	
296	098532-85-5	
297	098967-40-9	Flumetsulam [ANSI]
298	099793-38-1	1H-1,2,4-Triazole, 1,5-bis(4-chlorophenyl)-3-(methylthio)-
299	099793-75-6	1H-1,2,4-Triazole, 1-(4-chlorophenyl)-5-(4-fluorophenyl)-3-(methylsulfonyl)-
300	103112-35-2	Fenchlorazole-ethyl
301	103112-36-3	Fenchlorazole
302	103597-45-1	Bisoctrizole
303	103922-48-1	1H-1,2,4-Triazole-3,5-diamine, 1-methyl-N(sup 5)-(4-((4-(1-piperidinylmethyl)-2-pyridinyl)oxy)-2-butenyl)-, (Z)-
304	104958-85-2	RB 6110
305	106325-08-0	Epoxiconazole
306	107534-96-3	Tebuconazole
307	112143-82-5	Triazamate [ISO:BSI]
308	112281-77-3	Tetraconazole [ISO]
309	113518-46-0	
310	114369-43-6	Fenbuconazole [ISO]
311	116255-48-2	Bromuconazole
312	119126-15-7	Flupoxam [ISO:BSI]
313	119446-68-3	Difenoconazole [ISO]
314	122836-35-5	Sulfentrazone [ISO]
315	125116-23-6	Metconazole [ISO]
316	125225-28-7	Ipconazole [ISO]
317	125304-04-3	Benzotriazolyl dodecyl p-cresol
318	125306-83-4	Cafenstrole [ISO]

ID	CAS	Name
126	041735-30-2	s-Triazole, 5-(N-ethyl-N-nitro)amino-3-(5-nitro-2-furyl)-
127	041735-38-0	Acetamide, N-(1-methyl-3-(5-nitro-2-furyl)-s-triazol-5-yl)di-
128	041735-41-5	s-Triazole, 1-methyl-5-methylamino-3-(5-nitro-2-furyl)-
129	041735-42-6	s-Triazole, 3-amino-1-methyl-5-(5-nitro-2-furyl)-
130	041735-44-8	Acetamide, N-(1-methyl-5-(5-nitro-2-furyl)-s-triazol-3-yl)di-
131	041735-45-9	s-Triazole, 1-methyl-3-methylamino-5-(5-nitro-2-furyl)-
132	041735-50-6	Acetamide, N-(4-methyl-5-(5-nitro-2-furyl)-s-triazol-3-yl)di-
133	041735-51-7	s-Triazole, 4-methyl-3-nitramino-5-(5-nitro-2-furyl)-
134	041735-54-0	s-Triazole, 3-chloro-5-(5-nitro-2-furyl)-
135	041735-55-1	s-Triazole, 4-methyl-3-methylthio-5-(5-nitro-2-furyl)-
136	041735-56-2	s-Triazole, 4-methyl-3-methylsulfonyl-5-(5-nitro-2-furyl)-
137	041735-57-3	s-Triazole, 4-methyl-3-methoxy-5-(5-nitro-2-furyl)-
138	041814-78-2	Tricyclazole
139	041834-21-3	s-Triazole, 4-methyl-3-methylsulfinyl-5-(5-nitro-2-furyl)-
140	042509-80-8	Isazofos
141	043029-44-3	1,2,4-Triazolo(3,4-b)(1,3,4)thiadiazole, 3,6-diphenyl-
142	043121-43-3	Triadimefon
143	051627-14-6	Cefatrizine [USAN:INN:BAN]
144	053817-16-6	1H-1,2,3-Triazole-4,5-dicarbonitrile
145	054028-81-8	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-9-chloro-4-methyl-1-phenyl-
146	054028-83-0	4H-(1,2,4)Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-9-chloro-1-(2-methoxyphenyl)-4-methyl-
147	054028-84-1	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-9-chloro-4-methyl-1-(2-naphthalenyl)-
148	054028-85-2	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-

ID	CAS	Name
319	127519-17-9	Benzenepropanoic acid, 3-(2H-benzotriazol-2yl)-5- (1,1-dimethylethyl)-4-hydroxy-, C7-9-branched and linear alkyl esters
320	128625-52-5	PYBOP hexafluorophosphate
321	128639-02-1	Carfentrazone-ethyl [ISO:BSI]
322	129586-32-9	Ssf 109
323	129909-90-6	amicarbazone (Ref: BAY MKH 3586)
324	131983-72-7	Triticonazole [ISO]
325	136426-54-5	Fluquinconazole [ISO]
326	139158-24-0	3H-1,2,4-Triazole-3-thione, 2,4-dihydro-4-methyl-5-tricyclo(3.3.1.1(sup 3,7))dec-1-yl-
327	139158-25-1	3H-1,2,4-Triazole-3-thione, 2,4-dihydro-4-ethyl-5-tricyclo(3.3.1.1(sup 3,7))dec-1-yl-
328	139158-26-2	3H-1,2,4-Triazole-3-thione, 2,4-dihydro-4-(phenylmethyl)-5-tricyclo(3.3.1.1(sup 3,7))dec-yl-
329	139528-85-1	Metosulam [ISO]
330	141078-91-3	
331	141078-92-4	
332	141078-93-5	
333	141078-94-6	
334	141078-95-7	
335	141078-99-1	
336	141079-00-7	
337	141079-01-8	
338	141079-02-9	
339	141079-03-0	
340	141079-06-3	
341	141079-07-4	

ID	CAS	Name
		dihydro-9-chloro-4-methyl-1-(2-thienyl)-
149	054028-86-3	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-9-chloro-1-(2-furyl)-4-methyl-
150	054028-89-6	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-8-chloro-4-methyl-1-phenyl-
151	054028-90-9	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-8-chloro-1-(p-methoxyphenyl)-4-methyl-
152	054028-91-0	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-4-methyl-1-phenyl-
153	054028-92-1	4H-(1,2,4)Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-1-(2-methoxyphenyl)-4-methyl-
154	054028-93-2	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-1-(p-methoxyphenyl)-4-methyl-
155	054028-94-3	4H-s-Triazolo(4,3-a)(1,5)benzodiazepine, 5,6-dihydro-1-phenyl-
156	054028-95-4	
157	054123-06-7	6H-Thieno(3,2-f)(1,2,4)triazolo(4,3-a)(1,4)diazepine, 2-chloro-4-(2-chlorophenyl)-9-methyl-
158	055179-31-2	Bitertanol
159	055219-65-3	Triadimenol
160	055375-40-1	1H-1,2,4-Triazol-3-ol, 5-(2-butenylthio)-1-methyl-, methanesulfonate (ester)
161	055425-38-2	2-phenyl-3-(1H-1,2,4-triazol-5-ylazo)-1H-indole
162	056383-06-3	1,2,4-Triazolo(4,3-b)pyridazine, 8-methyl-6-(4-morpholinyl)-
163	056383-11-0	1,2,4-Triazolo(4,3-b)pyridazine, 3-methyl-6-(4-morpholinyl)-
164	056396-43-1	4H-1,2,4-Triazole-3,4-diamine, 5-(4-chlorophenyl)-N(sup 4)-((4-chlorophenyl)methylene)-
165	056881-36-8	s-Triazolo(4,3-c)pyrimidin-5-ol, 7-methyl-
166	057801-81-7	Brotizolam
167	057801-94-2	6H-Thieno(3,2-f)(1,2,4)triazolo(4,3-a)(1,4)diazepine, 2-bromo-4-(2-bromophenyl)-9-methyl-
168	059026-08-3	Epronaz
169	059338-86-2	Methyl 6-methoxy-1H-benzotriazole-5-carboxylate
170	059338-92-0	6-Methoxy-1H-benzotriazole-5-carboxylic acid
171	059338-93-1	Alizapride [INN]
172	060207-31-0	Azaconazole

ID	CAS	Name
342	141079-08-5	
343	141079-12-1	
344	141079-13-2	
345	141079-14-3	
346	141079-15-4	
347	141079-16-5	
348	141079-17-6	
349	141079-18-7	
350	141079-19-8	
351	141079-20-1	
352	145026-81-9	propoxycarbazone
353	145701-21-9	diclosulam (Ref: XDE 564)
354	145701-23-1	florasulam (Ref: DE 570)
355	147150-35-4	Cloransulam-methyl [ISO]
356	147993-59-7	Guerbetalkoholethoxylatbutylether
357	149508-90-7	simeconazole
358	173980-17-1	bencarbazone (Ref: TM-435)
359	178928-70-6	Prothioconazole [ISO:BSI]
360	212201-70-2	ipfencarbazone
361	219714-96-2	penoxsulam (Ref: XDE 638)
362	317815-83-1	thiencarbazone-methyl (Ref: BYH 18636)
363	348635-87-0	amisulbrom
364	422556-08-9	pyroxsulam (Ref: XDE 742)
365	865318-97-4	ametoctradin (Ref: BAS 650F)

ID	CAS	Name
173	060207-90-1	Propiconazole [BSI:ISO]
174	060207-93-4	Etaconazole [BSI:ISO]
175	060932-58-3	1H-Benzotriazolecarboxylic acid
176	061691-97-2	Ethanol, 2,2'-((1H-benzotriazol-1-ylmethyl)imino)bis-
177	063216-86-4	2-[4-[2-(4-amino-2-sulphophenyl)vinyl]-3-sulphophenyl]-2H-naphtho[1,2-d]triazole-5-sulphonic acid
178	063251-40-1	2H-Naphtho(1,2-d)triazole-6,8-disulfonic acid, 2-(4-aminophenyl)-
179	063870-37-1	1H-naphtho[1,2-d]triazole-6-sulphonic acid
180	064057-50-7	Acetamide, 2-chloro-N-(4-methyl-5-(5-nitro-2-furyl)-5-triazol-3-yl)-
181	064082-38-8	Acetamide, 2,2-dichloro-N-(3-(5-nitro-2-furyl)-5-triazol-5-yl)-
182	066104-34-5	2H-1,2,3-Triazolo(4,5-b)pyridin-5-amine, 2-(4-methoxyphenyl)-
183	066104-44-7	2H-1,2,3-Triazolo(4,5-b)pyridin-5-amine, 2-(3,4-dimethylphenyl)-
184	066246-88-6	Penconazole
185	066492-64-6	1H-1,2,4-Triazole, 1-acetyl-3-(p-chlorophenyl)-5-methyl-
186	066535-86-2	Lotrifen [INN]
187	066975-54-0	diethyl [[[3-(4,7-dihydro-7-oxo-1H-1,2,3-triazolo[4,5-d]pyrimidin-5-yl)-4-propoxyphenyl]amino]methylene]malonate
188	067465-03-6	Pyridine, 4-(3-acetonylthio)-5-(4H-(1,2,4-triazolyl))-
189	067465-05-8	Pyridine, 4-(3-(2-dimethylaminoethyl)-5-(4H-1,2,4-triazolyl))-

ID	CAS	Name
366	XXX002	2-(2,2-difluoroethoxy)-N-(5-hydroxy-8-methoxy[1,2,4]triazolo[1,5-c]pyrimidin-2-yl)-6-(trifluoromethyl)benzenesulfonamide
367	XXX003	(2-hydroxy-3,3-dimethyl-2-[1,2,4]triazol-1-ylmethyl-cyclopentyl)-(4-chlorophenyl)-methanone (Ref: CL 382389)
368	XXX004	3-({[2-(2,2-difluoroethoxy)-6-(trifluoromethyl)phenyl]sulfonyl}amino)-1H-1,2,4-triazole-5-carboxylic acid
369	XXX006	2-(2,4-dichlorophenyl)-3-(1H-1,2,4-triazol-1-yl)propan-1-ol (Ref: M14360-alcohol)
370	XXX007	4-[2-hydroxy-3,3-dimethyl-1-(1H-1,2,4-triazol-1-yl)butoxy]-benzoic acid (Ref: BUE 2684)
371	XXX008	2-(1-chlorocyclopropyl)-1-(2-chlorophenyl)-3-[5-(methylthio)-1H-1,2,4-triazol-1-yl]propan-2-ol
372	XXX009	2-(1-chlorocyclopropyl)-1-(2-chlorophenyl)-3-(3H-1,2,4-triazol-3-yl)propan-2-ol
373	XXX010	2-chloro-3-{2-chloro-5-[4-(difluoromethyl)-3-methyl-5-oxo-4,5-dihydro-1H-1,2,4-triazol-1-yl]-4-fluorophenyl}propanoic acid
374	XXX011	(1R,2E,3S,4R,5R)-2-(4-chlorobenzylidene)-4,5-dimethyl-1-(1H-1,2,4-triazol-1-ylmethyl)cyclopentane-1,3-diol
375	XXX012	(E)-2-(4-chlorobenzylidene)-5,5-dimethyl-1-((1H)-1,2,4-triazol-1-ylmethyl)-cyclopentan-1,3-diol (Ref: RPA 404766)
376	XXX013	1-[2-[2-chloro-4-(4-chloro-phenoxy)-phenyl]-2-1H-[1,2,4]triazol-yl]-ethanol (Ref: CGA 205375)
377	XXX014	N-(2,6-difluorophenyl)-8-fluoro-5-hydroxy[1,2,4]triazolo[1,5-c]pyrimidine-2-sulfonamide
378	XXX015	N-(2,6-difluorophenyl)-5-aminosulfonyl-1H-1,2,4-triazole-3-carboxylic acid
379	XXX016	5-(aminosulfonyl)-1H-1,2,4-triazole-3-carboxylic acid
380	XXX017	trans-5-(4-chlorophenyl)-dihydro-3-phenyl-3- (1H-1,2,4-triazole-1-ylmethyl)-2-3H-furanone (Ref: RH-9130)
381	XXX018	4-(4-chlorophenyl)-2-(methyl-1H-1,2,4-triazole)-4-oxo-2-phenylbutanenitrile (Ref: RH-6467)
382	XXX019	5-amino-N-(2,6-dichloro-3-methylphenyl)-1H-1,2,4-triazole-3-sulfonamide

ID	CAS	Name
190	068049-83-2	Azafenidin [ISO]
191	069141-50-0	3-Octanone, 6-hydroxy-2,2,7,7-tetramethyl-5-(1H-1,2,4-triazol-1-yl)-,(5R,6R)-rel-
192	070292-10-3	1H-1,2,3-Triazole-4-carboxylic acid, 5-methyl-1-(1-naphthalenyl)-
193	070321-86-7	Phenol, 2-(2H-benzotriazol-2-yl)-4,6-bis(1-methyl-1-phenylethyl)-

ID	CAS	Name
383	XXX020	3-(2-((1H-1,2,4-triazol-1-yl)methyl)-2-(2,4-dichlorophenyl)-1,3-dioxolan-4-yl)propan-1-ol (Ref: CGA 118245)
384	XXX021	cis-5-(4-chlorophenyl)-dihydro-3-phenyl-3-(1H-1,2,4-triazole-1-ylmethyl)-2-3H-furanone (Ref: RH-9129)
385	XXX022	N-(2,6-dichloro-3-methylphenyl)-5-methoxy-7-hydroxy-1,2,4-triazolo[1,5-alpha]pyrimidine-2-sulfonamide
386	XXX023	(2RS)-1-(4-chlorophenyl)-4,4-dimethyl-2-(1H-1,2,4-triazol-1-yl)pentan-3-one (Ref: CGA 149907)

Table A-7. QSAR models for the prediction of EC_{50} in *Pseudokirchneriella subcapitata*, EC_{50} in *Daphnia magna*, and LC_{50} in *Onchorynchus mykiss* of B-TAZs.

Endpoint	Model	Full Model Equation	N_{Tr}	R^2	Q^2_{LOO}	Q^2_{ext} [a]	CCC_{ext}	$RMSE_{Tr}$	$RMSE_p$	R^2ys
pEC_{50}-72h ***P. subcapitata***	**DRAGON**	**pEC_{50}**=3.448 + 0.014 *AeigZ* + 0.090 *T(N..S)* + 0.150 *Seigv*	35	0.82	0.77	0.72-0.84[b]	0.86-0.87[b]	0.41	0.43-0.47[b]	0.09
	PaDEL-Descriptor	**pEC_{50}**=1.505 + 0.027 *AMR* + 0.432 *MDEN-22* + 0.472 *maxwHBa*	35	0.82	0.76	0.67-0.89[b]	0.82-0.91[b]	0.42	0.35-0.51[b]	0.08
	QSPR-Thesaurus	**pEC_{50}**= 3.096 + 0.006 *p1p4_5N* + 0.227 *C-C* - 0.103 *p5BE*	35	0.80	0.73	0.71-0.83[b]	0.86-0.87[b]	0.44	0.44[b]	0.09
pEC_{50}-48h ***D.magna***	**DRAGON**	**pEC_{50} =** 3.725 - 0.019 *TPSA(NO)* + 0.009 *Aeigm* + 0.048 *nCar* + 0.192 *nHDon* + 0.027 *H-052*	97	0.77	0.74	0.69-0.83[b]	0.85-0.89[b]	0.39	0.34-0.44[b]	0.05
pLC_{50}-96h ***O.mykiss***	**DRAGON**	**pLC_{50} =** -6.705 + 1.579 *CIC1* + 13.251 *Mp* + 0.135 *H-052* - 0.005 *TPSA (Tot)*	75	0.79	0.76	0.85[c]	0.92[c]	0.48	0.41[c]	0.05
	PaDEL-Descriptor	**pLC_{50} =** 2.325 + 0.392 *VP-1* – 0.049 *SHBint2* – 0.335 *maxHaaCH*	75	0.76	0.73	0.71-0.72[c]	0.82[c]	0.51	0.57[c]	0.04

[a] Range of Q^2_{ext-F1}, Q^2_{ext-F2} and Q^2_{ext-F3}. [b] Range of the external parameters of the *Split models* (by response and by structure). [c] Range of the external parameters calculated for the EV set.

Table A-8. List of prioritized B-TAZs derived from the analysis of Consensus predictions (into AD) for acute toxicity in algae, daphnids and fish. Experimental and predicted E(L)C_{50} values are reported in mg/L.

ID PCA	CAS	*P.subcapitata*			*D.magna*			*O.mykiss*		
		Exp.EC_{50}	Pred.EC_{50}	Class	Exp.EC_{50}	Pred.EC_{50}	Class	Exp.LC_{50}	Pred.LC_{50}	Class
40	055179-31-2	1.38	0.77	Very Toxic	4.46	3.22	Toxic	2.14	1.99	Toxic
44	056396-43-1	-	1.83	Toxic	-	3.25	Toxic	-	1.48	Toxic
61	081518-29-8	-	2.40	Toxic	-	2.95	Toxic	-	0.43	Very Toxic
62	081518-41-4	-	1.59	Toxic	-	2.30	Toxic	-	0.73	Very Toxic
74	098519-02-9	-	0.73	Very Toxic	-	10.60	Harmful	-	1.53	Toxic
86	098519-33-6	-	1.24	Toxic	-	10.08	Harmful	-	2.15	Toxic
98	098532-73-1	-	1.02	Toxic	-	16.99	Harmful	-	2.04	Toxic
101	098532-77-5	-	1.45	Toxic	-	10.77	Harmful	-	2.25	Toxic
102	098532-81-1	-	1.33	Toxic	-	10.55	Harmful	-	2.01	Toxic
105	099793-38-1	-	2.12	Toxic	-	1.76	Toxic	-	0.36	Very Toxic
107	106325-08-0	1.19	2.13	Toxic	8.69	5.35	Toxic	3.14	1.85	Toxic
110	114369-43-6	0.33	0.89	Very Toxic	2.3	2.28	Toxic	1.5	1.28	Toxic
112	119446-68-3	0.032	0.55	Very Toxic	0.77	3.08	Toxic	1.1	0.71	Very Toxic
113	125116-23-6	-	1.76	Toxic	4.2	5.07	Toxic	2.1	3.67	Toxic
114	125225-28-7	-	1.30	Toxic	1.7	4.50	Toxic	>0.76	2.16	Toxic
116	128639-02-1	-	1.88	Toxic	>9.8	7.68	Toxic	1.6	4.22	Toxic
121	xxx008	3.77	2.26	Toxic	2.8	8.40	Toxic	1.8	1.75	Toxic
125	xxx017	-	0.81	Very Toxic	-	3.88	Toxic	-	1.73	Toxic
126	xxx018	-	0.80	Very Toxic	-	4.84	Toxic	-	2.99	Toxic
127	xxx021	-	0.81	Very Toxic	-	3.97	Toxic	-	1.73	Toxic

APPENDIX III

APPENDIX III

Table A-9. List of training set chemicals, experimental and predicted classes of ready biodegradability (1=RB, 2=NRB).

ID	CAS / InChIKey	Exp CLA	DRAGON Models				PaDEL model
			M1	M2	M3	Consensus	M4
1	CAS_101-72-4	**2**	2	2	2	2	2
2	CAS_101-80-4	**2**	2	2	2	2	2
3	CAS_101-84-8	**2**	2	2	2	2	2
4	CAS_102-09-0	**2**	2	2	2	2	2
5	CAS_10233-13-3	**1**	2	1	1	1	1
6	CAS_103-09-3	**1**	2	1	1	1	1
7	CAS_103-64-0	**2**	2	1	2	2	2
8	CAS_104-87-0	**1**	1	1	1	1	1
9	CAS_104-93-8	**2**	1	1	2	2	1
10	CAS_105-45-3	**1**	1	2	1	1	2
11	CAS_106-24-1	**1**	1	1	1	1	1
12	CAS_106-50-3	**2**	2	2	2	2	2
13	CAS_106-65-0	**1**	1	1	1	1	1
14	CAS_106-92-3	**2**	1	1	2	1	2
15	CAS_106-99-0	**2**	2	1	2	2	2
16	CAS_107-45-9	**2**	1	2	2	2	2
17	CAS_108-11-2	**1**	2	1	1	1	1
18	CAS_108-18-9	**2**	2	2	2	2	2
19	CAS_108-20-3	**2**	2	2	2	2	2
20	CAS_108-22-5	**1**	1	1	1	1	1
21	CAS_108-39-4	**1**	1	2	1	1	1
22	CAS_108-59-8	**1**	1	2	1	1	1
23	CAS_108-93-0	**1**	1	1	1	1	1
24	CAS_110-05-4	**2**	2	2	2	2	2
25	CAS_111-03-5	**1**	1	1	1	1	1
26	CAS_111-81-9	**2**	1	1	1	1	1
27	CAS_111-84-2	**1**	1	1	1	1	1
28	CAS_1120-24-7	**1**	1	1	2	1	1
29	CAS_112-05-0	**1**	2	1	1	1	1
30	CAS_112-27-6	**1**	1	1	1	1	1
31	CAS_112-34-5	**1**	1	1	1	1	1
32	CAS_112-62-9	**1**	1	1	1	1	1
33	CAS_112-72-1	**1**	2	1	1	1	2
34	CAS_115-07-1	**2**	2	2	2	2	2
35	CAS_115-10-6	**2**	2	2	2	2	2
36	CAS_115-95-7	**1**	1	1	1	1	1
37	CAS_116-02-9	**2**	2	2	2	2	2
38	CAS_119-36-8	**1**	1	1	1	1	1
39	CAS_119-64-2	**2**	2	2	2	2	2

ID	CAS / InChIKey	Exp CLA	DRAGON Models				PaDEL model
			M1	M2	M3	Consensus	M4
40	CAS_120-14-9	**1**	1	1	1	1	1
41	CAS_120-92-3	**1**	1	2	1	1	1
42	CAS_122-03-2	**1**	1	1	1	1	1
43	CAS_122-40-7	**1**	1	1	1	1	1
44	CAS_123-35-3	**1**	1	1	1	1	1
45	CAS_124-09-4	**1**	1	1	1	1	2
46	CAS_124-28-7	**1**	1	1	1	1	1
47	CAS_127-91-3	**1**	2	2	2	2	2
48	CAS_131-11-3	**1**	1	1	1	1	1
49	CAS_140-11-4	**1**	1	2	1	1	1
50	CAS_140-88-5	**2**	1	1	1	1	1
51	CAS_141-32-2	**2**	2	1	1	1	1
52	CAS_141-78-6	**1**	1	1	2	1	2
53	CAS_142-19-8	**1**	1	1	1	1	1
54	CAS_142-22-3	**1**	1	1	1	1	1
55	CAS_143-28-2	**1**	1	1	1	1	1
56	CAS_1490-04-6	**2**	1	1	1	1	1
57	CAS_1502-22-3	**2**	2	2	2	2	2
58	CAS_150-78-7	**1**	1	1	2	1	2
59	CAS_15356-60-2	**1**	2	2	2	2	2
60	CAS_1634-04-4	**2**	1	2	2	2	2
61	CAS_1724-39-6	**1**	1	1	1	1	1
62	CAS_17796-82-6	**2**	2	2	1	2	2
63	CAS_2049-95-8	**2**	1	2	2	2	2
64	CAS_2146-71-6	**1**	1	1	1	1	1
65	CAS_2173-57-1	**2**	2	2	2	2	2
66	CAS_2409-55-4	**2**	2	2	2	2	2
67	CAS_2416-94-6	**2**	2	2	2	2	2
68	CAS_2436-90-0	**1**	1	1	1	1	1
69	CAS_25155-25-3	**2**	1	2	2	2	2
70	CAS_25265-71-8	**1**	1	1	1	1	1
71	CAS_25340-17-4	**2**	2	2	1	2	2
72	CAS_25377-72-4	**2**	2	2	2	2	2
73	CAS_26780-96-1	**2**	2	2	2	2	2
74	CAS_26896-20-8	**2**	2	2	2	2	2
75	CAS_26896-48-0	**2**	2	1	2	2	2
76	CAS_27776-01-8	**2**	2	2	2	2	2
77	CAS_28219-61-6	**2**	1	2	2	2	2
78	CAS_294-62-2	**2**	2	2	2	2	2
79	CAS_3006-82-4	**2**	2	1	1	1	1
80	CAS_31906-04-4	**1**	1	1	1	1	1
81	CAS_32539-83-6	**2**	2	1	2	2	2
82	CAS_3452-97-9	**2**	1	2	2	2	2
83	CAS_4904-61-4	**2**	2	2	2	2	2

ID	CAS / InChIKey	Exp CLA	DRAGON Models				PaDEL model
			M1	M2	M3	Consensus	M4
84	CAS_504-02-9	**2**	1	1	1	1	1
85	CAS_507-70-0	**1**	1	1	1	1	2
86	CAS_51000-52-3	**2**	2	2	2	2	2
87	CAS_513-35-9	**2**	1	1	1	1	2
88	CAS_5333-42-6	**2**	1	1	1	1	2
89	CAS_54549-24-5	**1**	1	1	1	1	1
90	CAS_54839-24-6	**1**	1	1	1	1	1
91	CAS_563-80-4	**1**	1	1	1	1	1
92	CAS_576-26-1	**2**	2	2	2	2	2
93	CAS_586-62-9	**1**	1	1	1	1	1
94	CAS_590-86-3	**2**	2	2	2	2	1
95	CAS_598-56-1	**1**	2	1	2	2	2
96	CAS_599-64-4	**2**	2	2	2	2	1
97	CAS_60-12-8	**1**	1	1	1	1	1
98	CAS_620-17-7	**1**	1	2	1	1	1
99	CAS_66063-15-8	**2**	2	2	2	2	2
100	CAS_66-25-1	**1**	1	1	2	1	1
101	CAS_68479-98-1	**2**	2	2	2	2	2
102	CAS_68956-55-8	**2**	2	2	2	2	2
103	CAS_693-23-2	**1**	1	1	1	1	1
104	CAS_71-41-0	**1**	1	1	1	1	1
105	CAS_75-91-2	**2**	1	2	2	2	1
106	CAS_76-22-2	**1**	1	1	1	1	2
107	CAS_763-32-6	**1**	1	1	1	1	1
108	CAS_78-59-1	**1**	1	2	2	2	2
109	CAS_78-69-3	**1**	1	1	1	1	1
110	CAS_79-06-1	**1**	1	2	1	1	1
111	CAS_79-20-9	**1**	1	2	1	1	1
112	CAS_79-41-4	**1**	1	1	2	1	1
113	CAS_79-46-9	**2**	2	1	1	1	1
114	CAS_79-92-5	**2**	1	2	1	1	2
115	CAS_81-15-2	**2**	2	2	2	2	2
116	CAS_84-74-2	**1**	1	1	1	1	1
117	CAS_868-77-9	**1**	1	1	1	1	1
118	CAS_873-94-9	**2**	2	1	1	1	2
119	CAS_90-02-8	**1**	2	1	1	1	1
120	CAS_90-05-1	**1**	1	1	1	1	1
121	CAS_90-72-2	**2**	2	2	1	2	1
122	CAS_924-42-5	**2**	2	2	2	2	2
123	CAS_92-84-2	**2**	2	2	2	2	2
124	CAS_93-15-2	**1**	1	1	1	1	2
125	CAS_94-04-2	**2**	1	1	1	1	1
126	CAS_95-87-4	**2**	2	2	2	2	2
127	CAS_96-17-3	**2**	2	1	2	2	1

ID	CAS / InChIKey	Exp CLA	DRAGON Models				PaDEL model
			M1	M2	M3	Consensus	M4
128	CAS_96-26-4	**1**	1	2	1	1	2
129	CAS_97-99-4	**1**	1	1	1	1	1
130	CAS_98-51-1	**2**	2	2	2	2	2
131	CAS_99-85-4	**1**	1	1	1	1	1
132	CAS_99-87-6	**1**	2	1	1	1	1
133	CAS_99-96-7	**1**	1	1	1	1	1
134	EC_415-450-7	**2**	2	2	2	2	2
135	KFRVYYGHSPLXSZUHFFFAOYSAN	**2**	2	2	2	2	2
136	MFCLOAFNUWAGGBUHFFFAOYSAN	**2**	2	2	2	2	2

Table A-10. List of validation set chemicals, experimental and predicted classes of ready biodegradability (1=RB, 2=NRB).

ID	CAS	Exp CLA	DRAGON				PaDEL
			M1	M2	M3	Consensus	M4
1	000056-54-2	**1**	2	2	2	2	2
2	000083-66-9	**1**	2	2	2	2	2
3	000093-51-6	**1**	1	1	1	1	1
4	000097-53-0	**1**	1	1	1	1	1
5	000097-54-1	**1**	1	1	1	1	1
6	000101-85-9	**1**	1	1	1	1	1
7	000101-86-0	**1**	1	1	1	1	1
8	000103-41-3	**1**	2	2	2	2	2
9	000106-02-5	**1**	1	1	1	1	1
10	000118-61-6	**1**	1	1	1	1	1
11	000119-36-8	**1**	1	1	1	1	1
12	000120-51-4	**1**	2	2	2	2	2
13	000122-48-5	**1**	1	1	1	1	1
14	000470-82-6	**1**	2	2	2	2	2
15	001222-05-5	**2**	2	2	2	2	2
16	001335-66-6	**2**	1	1	1	1	1
17	002050-08-0	**1**	1	1	1	1	1
18	003209-13-0	**1**	1	2	1	1	1
19	005595-79-9	**1**	1	1	1	1	1
20	006259-76-3	**1**	1	1	1	1	1
21	006290-17-1	**2**	1	1	2	1	1
22	006413-10-1	**2**	1	1	1	1	1
23	007392-19-0	**2**	2	2	2	2	2
24	018871-14-2	**1**	1	1	1	1	1
25	024851-98-7	**1**	1	1	1	1	1
26	025485-88-5	**1**	2	2	2	2	2
27	032388-55-9	**2**	2	2	2	2	2
28	037677-14-8	**2**	1	1	1	1	1
29	039067-39-5	**2**	2	2	2	2	2

ID	CAS	Exp CLA	DRAGON				PaDEL
			M1	M2	M3	Consensus	M4
30	052475-86-2	**2**	2	2	2	2	1
31	062406-73-9	**2**	2	2	2	2	2
32	063500-71-0	**2**	1	1	1	1	2
33	065405-77-8	**1**	1	1	1	1	1
34	066327-54-6	**2**	1	2	2	2	1
35	067845-30-1	**2**	2	2	1	2	2
36	068039-49-6	**2**	1	1	1	1	2
37	068738-94-3	**2**	2	2	2	2	1
38	068738-96-5	**2**	2	2	2	2	2
39	068991-97-9	**2**	2	2	2	2	2
40	121251-67-0	**2**	2	1	2	2	2
41	121251-68-1	**2**	2	1	2	2	2
42	128489-04-3	**2**	2	2	1	2	2
43	131812-52-7	**2**	1	2	2	2	2
44	154171-77-4	**2**	2	1	2	2	2
45	166301-22-0	**2**	1	2	2	2	2