In silico assessment of aquatic bioaccumulation: advances from chemometrics and QSAR modelling

Francesca Grisoni

February 3, 2016

Version: 1.0

University of Milano-Bicocca

Dept. of Earth and Environmental Sciences

PhD Dissertation

In silico assessment of aquatic bioaccumulation: advances from chemometrics and QSAR modelling

Francesca Grisoni

Supervisors	Prof. R Todeschini
	Dr. V Consonni
	Dr. S Villa

February 3, 2016

Francesca Grisoni
In silico assessment of aquatic bioaccumulation: advances from chemometrics and QSAR modelling
PhD Dissertation, February 3, 2016
Supervisors: Prof. R Todeschini, Dr. V Consonni and Dr. S Villa
Ext. Advisor: Dr. M Pavan

University of Milano-Bicocca
Milano Chemometrics and QSAR Research Group
Dept. of Earth and Environmental Sciences
P.za della Scienza, 1
20126 Milano - Italy

To Diego.

Abstract

This doctoral dissertation addresses some of the current open problems in the prediction of aquatic bioaccumulation of organic chemicals, by exploiting Quantitative Structure-Activity Relationship (QSAR) and chemometric techniques. It aims to advance the mechanistic knowledge about the bioaccumulation processes and to overcome some of the existing modelling gaps. Bioconcentration and dietary bioaccumulation are addressed separately, using fish as the target organism.

The bioconcentration is considered the main bioaccumulation route and, therefore, the Bioconcentration Factor (BCF) has been widely modelled. However, in this work, the comparison of nine well-established models for BCF showed that, in most of the cases, only lipid-driven bioconcentration is well predicted and other influential processes, such as storage within non-lipid tissues or metabolism, are neglected. This set the basis for the development of a classification scheme to identify compounds that (1) are well predicted by their lipophilicity, (2) have additional storage sites (e.g. proteins) and an increased BCF, or (3) are metabolized/eliminated, with a reduced bioconcentration. The classification scheme allowed to gather knowledge about the mechanisms of bioconcentration, and to develop an expert system to choose the most appropriate modelling tool according to the predicted bioconcentration mechanism. The final expert system led to an increased accuracy than its sub-models.

The second target was the dietary bioaccumulation, expressed through the laboratory-based Biomagnification Factor (BMF), which is usually not considered for the bioaccumulation assessment. The comparison between BMF and BCF revealed that, for some chemical classes, the

dietary bioaccumulation could be more relevant than the bioconcentration. On this basis, a QSAR model was developed to predict the BMF of organic chemicals. The mechanistic interpretation of the results unveiled the structural features that may be responsible for a preferential bioaccumulation through diet and those that are shared with the bioconcentration process. The model, which complied with the OECD principles for QSAR validity and regulatory acceptance, can be used as a side tool for the assessment of the bioaccumulation of chemicals.

This dissertation put special attention to the mechanistic interpretation of the selected molecular descriptors and provided a set of efficient tools to estimate the chemical's propensity to bioaccumulate within the food chain, also adding knowledge to the field. Salient features of the developed QSAR approaches are simplicity and interpretability, which can allow for a widespread and transparent application, especially for regulatory purposes. Moreover, this work offers a theoretical basis for the hazard assessment of new emerging contaminants, such as Perfluorinated compounds.

Acknowledgements

I owe my deepest gratitude to my supervisor, Prof Roberto Todeschini, for exposing me to many research challenges and for giving me so many growth opportunities.
I also express my gratitude to Dr Viviana Consonni, for her care, her encouragement and all the fruitful suggestions.

I am grateful to Prof Marco Vighi for introducing me to this project; his guidance was essential in shaping this work. Thanks to Dr Sara Villa for her enthusiasm and support.

I would like to thank Dr Manuela Pavan for reviewing the dissertation manuscript and for the helpful feedbacks.

I acknowledge the present and former members of Milano Chemometrics Group for their help whenever I asked for it: Andrea Mauri, Davide Ballabio, Faizan Sahigara, Kamel Mansouri and Matteo Cassotti.

I would like to thank Prof Gisbert Schneider (ETH-Zurich) and Dr Richard Judson (U.S. EPA) for hosting me in their research teams and involving me in many stimulating projects.

I am deeply grateful to my family, which always supported my choices and believed in me. Thanks to my best friends, for the joy they bring in my life.

Finally, I am thankful to all the people that contributed to my personal and scientific growth, in particular to Biagio Di Mauro, Daniel Reker, Daniela Mingotti, Marta Mapelli and Tomoyuki Miyao.

Preface

In silico toxicology melds advanced computational technology with molecular biology and chemistry, to improve agency prioritization of data requirements and risk assessment of chemicals [1]. It is an area of very active development and great potential, whose importance is expected to grow in the next future [2].

The potential of computational toxicology lies in its broad scale of application, in terms of number of studied chemicals, breadth of endpoints and pathways covered, levels of biological organization and ranges of exposure conditions considered, and in the coverage of life stages, genders, and species. Thus, success in this area would translate into a more efficient determination of the hazards related to many environmental stressors [3]. Moreover, *in silico* methods are gaining increasing momentum to replace, reduce and refine, as much as possible, the use of animal testing for ethical reasons, goals that are pursued by many national and international legislations [4].

This work targets the bioaccumulation in aquatic environment, because of its potential to expose organisms, especially top-predators such as humans, to severe long-term effects difficult to predict. One of the most widely applied *in silico* methods, QSAR (Quantitative Structure-Activity Relationship), was combined with recent chemometric advances, aiming to address some gaps and problems in the state-of-the-art of bioaccumulation modelling.

Contents

I Introduction 1

1 Background 3

1.1 Bioaccumulation 4

1.1.1 Bioaccumulation processes 4

1.1.2 Regulation of bioaccumulative chemicals . . . 9

1.2 Quantitative Structure-Activity Relationship (QSAR) . 10

1.2.1 Theoretical Background 10

1.2.2 REACH Regulation and OECD principles . . . 17

2 Research Framework 19

2.1 Problem Statement 19

2.2 Contributions 20

2.3 Research Boundaries 21

2.4 Thesis Structure 22

II Bioconcentration 23

3 QSAR models for bioconcentration: complexity *versus* accuracy 25

3.1 BCF models 27

3.1.1 K_{OW}-based equations 27

3.1.2 Descriptor-based models 29

3.2 Data and curation 33

3.3 Results and discussion 36

3.3.1 Global Performance 36

3.3.2 Performance on critical compounds 38

3.4 Conclusions 43

4 Mechanisms of bioconcentration: insights from QSAR classification trees 45
4.1 Materials and methods 46
4.1.1 Dataset 46
4.1.2 Molecular Descriptors 47
4.1.3 Modelling strategy 47
4.2 Results and discussion 48
4.2.1 Class definition 49
4.2.2 Classification results 52
4.2.3 Model analysis and interpretation 55
4.2.4 *Consensus* model 66
4.3 Conclusions 67

5 Expert system to predict BCF for regulatory purposes 69
5.1 Materials and Methods 70
5.1.1 Datasets 70
5.1.2 Class prediction 70
5.2 Results and Discussion 71
5.2.1 Approach refinement 71
5.2.2 Analysis of the method 72
5.2.3 Comparison with benchmark BCF models . . 74
5.2.4 Expert-based system 76
5.3 Conclusions 79

III Dietary Bioaccumulation 81

6 Biomagnification factor: critical comparison with K_{OW} and BCF 83
6.1 BMF Dataset 84
6.2 Results and Discussion 86
6.2.1 Classification of bioaccumulative compounds . 86
6.2.2 Continuous values comparison 89
6.2.3 Cluster Analysis 90
6.3 Conclusions 93

7 Modelling the dietary bioaccumulation 95
7.1 Materials and methods 96
7.1.1 Dataset 96

7.1.2 Molecular Descriptors 97
7.1.3 Variable selection 97
7.2 Results and Discussion 98
7.2.1 Modelling strategy 98
7.2.2 Model Development 103
7.2.3 Model Analysis 104
7.3 Conclusions 121

IV Conclusions 123

Bibliography 127

List of Figures

1.1 Complex interactions between organisms of the food web 5
1.2 Simplified scheme of the bioaccumulation processes in food webs. 8

3.1 Graphical representation of K_{OW}-based equations. . . . 28
3.2 Graphical scheme of the chosen descriptor-based models. 32
3.3 Scheme of experimental data sources for merged and external datasets. 33
3.4 Experimental K_{OW} *vs* BCF in comparison with TGD model. 39
3.5 RMSE on compounds that are overestimated, underestimated and correctly estimated by TGD 41
3.6 Experimental *vs* predicted logBCF for compounds overestimated and underestimated by TGD. 42

4.1 Schematic workflow of the modelling approach. 48
4.2 Dataset compounds coloured according to the assigned class. 50
4.3 Classification tree for specifically bioconcentrating compounds 55
4.4 Classification tree for compounds with a reduced bioconcentration 61
4.5 Variation of ON1V according to the number of carbon and hydrogen atoms. 62
4.6 Simplified scheme of the *consensus* model 66

5.1 wR-Hasse Diagrams on the selected BCF models 77
5.2 Simplified scheme of the expert system 79

6.1 Distribution of K_{OW}, BCF and BMF values. 87
6.2 Comparison between BMF, BCF and K_{OW} 89

6.3 Self Organizing Map using K_{OW}, BCF and BMF as descriptors. 91

7.1 Experimental BMF values distribution 96
7.2 Multi-Dimensional scaling on model M1 descriptors . . 105
7.3 Factors that influence nBT descriptor 106
7.4 Factors that influence F06[C-C] descriptor 107
7.5 PCA on M1 descriptors 108
7.6 Applicability Domain assessment for model M1 111
7.7 Factors that influence X0Av and X1Per descriptors . . . 114
7.8 Factors that influence VE1_B(m) descriptor 116
7.9 Standardized regression coefficients of M2 117
7.10 Williams plot for model M2 119
7.11 *Consensus* model 120

List of Tables

1.1 Summary of the criteria for regulatory bioaccumulation assessment . . . 9

3.1 Global model performance. . . . 37
3.2 Model performance on critical compounds. . . . 40

4.1 Distribution within mechanistic classes of some environmentally relevant chemicals. . . . 52
4.2 Statistics of selected classification trees . . . 54
4.3 Examples of PCD, X2Av and piPC09 descriptor values . 58
4.4 Examples of ON1V and N-072 descriptor values 63
4.5 Statistics of the *consensus* classification . . . 67

5.1 Performance of the tested models for K_{OW} prediction. . . 72
5.2 Statistics of TGD model on each predicted class. 73
5.3 Statistics of the benchmark models on each predicted class. . . . 75
5.4 Chosen weighting scheme for the wR-Hasse technique . 77
5.5 Statistics of the expert system . . . 78

6.1 Classification matrix of the bioaccumulation assessment according to different criteria . . . 88
6.2 Contingency table of the bioaccumulation assessment . 88

7.1 Regression methods and variable selection techniques used in the exploratory phase. . . . 100
7.2 Performance of the selected models in fitting, cross-validation and on the test set . . . 104
7.3 Applicability Domain assessment for model M1 111
7.4 AD assessment for the *consensus* model. . . . 121

Part I

Introduction

1 Background

> *The ultimate test of man's conscience may be his willingness to sacrifice something today for future generations whose words of thanks will not be heard.*
>
> — **Gaylord Nelson**
> United States Senator

Humans produce and emit in the environment up to tens of thousands of xenobiotics, whose fate is influenced by both their physico-chemical properties and those of the surrounding environment. Aquatic environment, in particular, has a crucial role in determining the fate of contaminants, as it is often their final sink, because of direct immision or hydrological/atmospheric processes. Within water, substances can be subject to long-range transport and can be transferred to other environmental compartments, such as biota. From water, some substances can bioaccumulate within organisms, reaching tissue concentrations that are several orders of magnitude higher than those measured in the environment.

More than forty years after the book of Rachel Carson (*Silent Spring*, [5]) documenting the detrimental effects of a widespread and not-regulated use of chemicals, in 2004, 131 Countries ratified and adopted the Stockholm Convention on Persistent Organic Pollutants (POPs). The main goal of the convention was to identify chemicals that, because of their lack of degradability, their tendency to bioaccumulate in food chains and their toxicity, can achieve harmful concentrations in upper trophic level organisms, such as human beings [6].

Bioaccumulation is recognized by the Stockholm Convention as a key feature of hazardous compounds. Hazardous POPs, in particular, are those chemicals that: *“(1) possess toxic characteristics, (2) are persistent, (3) are liable to bioaccumulate, (4) are prone to long-range transport and (5) are likely to cause significant adverse effects”*.
Moreover, even without acute/chronic effects detected from standard (eco)toxicity tests, bioaccumulation should be regarded as an hazard in itself. Some effects, in fact, may manifest in later phases of life, are multi-generational (e.g. endocrine disruption [7]) or can affect only high members of food webs [8].

Some of the most notorious examples of the potential deleterious effects of bioaccumulation are the near extinction of birds of prey due to the egg-shell thinning induced by dichlorodiphenyltrichloroethane (DDT) ([9], [10]) and the human methylmercury poisoning due to contaminated seafood in the Minamata Bay, Japan ([11], [12]).

This chapter introduces some fundamental concepts about the bioaccumulation process and its regulation. Moreover, it gives some insights into the field of QSAR (Quantitative Structure-Activity Relationship) modelling and its role in assessing the environmental fate of pollutants.

1.1 Bioaccumulation

1.1.1 Bioaccumulation processes

The hypothesis that man-made chemicals could bioaccumulate in food webs was first brought to the attention of the scientific community by Woodwell in 1967 [13]. His hypothesis was that ecological cycles can concentrate pollutants to levels at which they could be harmful to animals and humans, because of biomass transfer in food chains (Fig. 1.1).

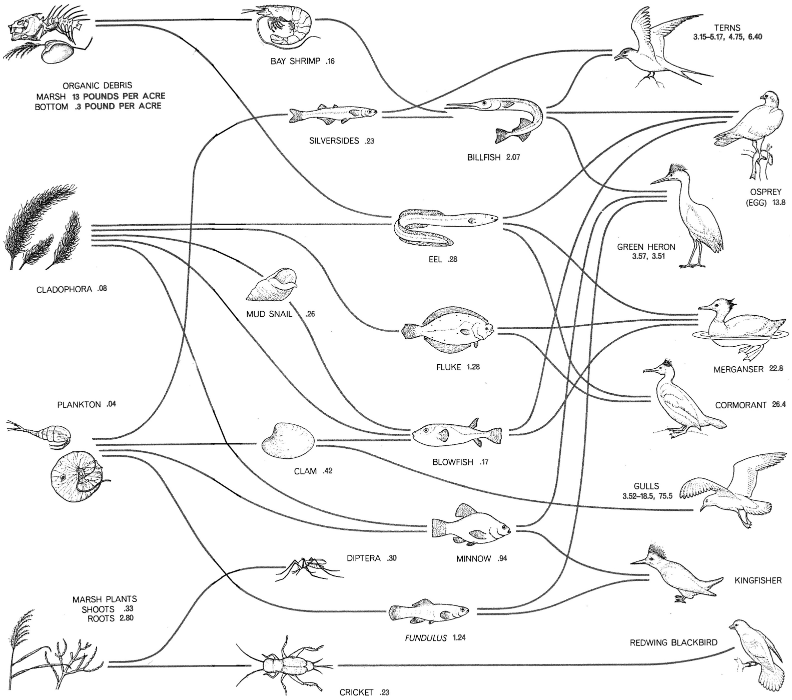


Fig. 1.1: Complex interactions between organisms of the food web; image from the original work of Woodwell, 1967 [13], which firstly hypothesised the bioaccumulation process. Numbers indicate the residues of DDT and its metabolites (ppm) in a Long Island estuary ecosystem.

A later study of Hamelink *et al.* in 1971 [14] observed the bioaccumulation of DDT in the food chain, but could not demonstrate that it was due to predator-prey mechanisms. The authors proposed that DDT bioaccumulation was the result of an exchange between water and fats, caused by the lipophilicity of DDT. This work was followed by a large number of other studies, in which fish were exposed to chemicals through water under laboratory conditions (e.g. [15]–[17]). The results showed excellent correlations between the octanol-water partition coefficient (K_{OW}) and the concentrations found within organisms. In this way, the bioaccumulation of chemicals was assumed to be an extremely predictable phenomenon.

However, later, field studies of the bioaccumulation of DDT and other persistent halogenated organic chemicals [18] showed that chemical concentrations in biota were greater than expected from lipid-water partitioning and also increased with increasing trophic level. These results evidenced that lipid-water partitioning alone is not able to explain the distribution of chemicals within biota and that additional processes could cause chemical transport from prey to predator. This chemical transport, differently from that driven by K_{OW}, involves the chemical transport against the thermodynamic gradient (from a low fugacity in the prey, to a higher fugacity in the predator), in a more complex way than expected. Nowadays, the complexity of bioaccumulation has been widely recognized, and the processes through which it occurs have been clearly defined and are briefly described below.

Bioconcentration Bioconcentration is the process of accumulation of water-borne chemicals in aquatic animals through non-dietary routes [19], such as skin or respiratory surfaces (e.g. lungs/gills). It occurs as a solubility-controlled partitioning between water and the animal. Bioconcentration ability of chemicals is quantified through the Bioconcentration Factor (BCF). BCF is determined in a laboratory experiment where the test organism is exposed to the chemical in the water but not in the diet. BCF is expressed as follows:

$$BCF = \frac{C_O}{C_W} \tag{1.1}$$

where C_O (g/kg wet weight) is the chemical concentration in the water-respiring organism and C_W (g/L) is the concentration in the water at the steady-state. The BCF is an estimate of a chemical's propensity to accumulate in aquatic animals. Typically fish are the targets of BCF assessment because of their importance in the food web (e.g. as a human food source) and the availability of standardized testing protocols.

Dietary bioaccumulation Dietary bioaccumulation is the process of chemical storage within organisms that is caused by the transport from prey to predator through diet [20]. Dietary bioaccumulation is fundamentally different from bioconcentration, since it involves chemical transport against the thermodynamic gradient. Often, dietary bioaccumulation is referred to as biomagnification, causing some ambiguities within the literature. In this dissertation, we will keep the definition of dietary bioaccumulation and use the term biomagnification with its broadest meaning (see next paragraph). The dietary bioaccumulation is quantified using the Biomagnification Factor (BMF), defined as follows:

$$BMF = \frac{C_O}{C_D} \tag{1.2}$$

where C_O (g/kg wet weight) is the chemical concentration in the water- or air-respiring organism and C_D (g/kg dry) is the concentration in the diet at the steady-state. The BMF can be laboratory-based or field-based. In the former case, organisms are exposed only through the diet, while in the latter case, they can be exposed through all the possible routes.

Bioaccumulation The term bioaccumulation is used to indicate the generic accumulation from the environment to organisms, by considering all the possible routes of uptake [21]. The bioaccumulation ability of chemicals is quantified using the Bioaccumulation Factor (BAF):

$$BAF = \frac{C_O}{C_W} \tag{1.3}$$

where C_O (g/kg wet weight) is the chemical concentration in the water- respiring organism and C_W (g/L) is the concentration in the water. The BAF is determined from field data and, thus, it accounts for both bioconcentration and dietary accumulation.

Biomagnification Biomagnification is the increase of the chemical concentration when increasing the trophic level (Fig. 1.2). It is the result of dietary bioaccumulation and bioconcentration at each level of the food web [22]. A relevant measure of biomagnification is the Trophic Magnification Factor (TMF) [23]. In the simplest formulation, an exponential increase of the chemical concentration with increasing the trophic level is assumed and the TMF is defined as follows:

$$C_{TL} = 10^{b \cdot TL}$$

$$\log C_{TL} = a + b \cdot TL \tag{1.4}$$

$$TMF = 10^b$$

where C_{TL} is the concentration at a given trophic level. In other words, because contaminant concentrations often increase exponentially through the food web, the TMF is calculated as the antilog of the regression slope b with base 10 (or e). Interested readers can find additional details in the work of Borgå *et al.* [23].

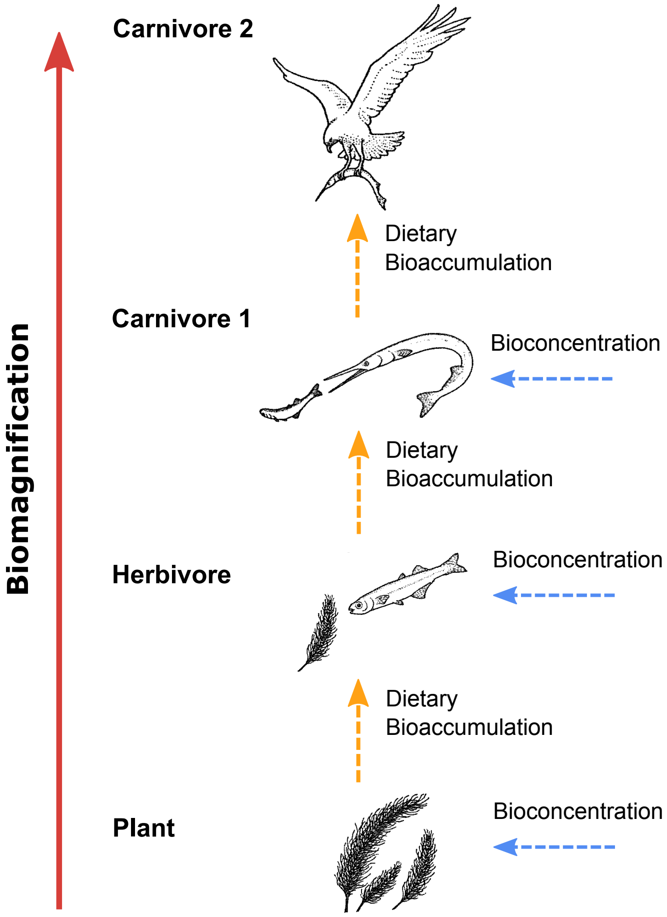


Fig. 1.2: Simplified scheme of the bioaccumulation processes in food webs. At each trophic level, bioconcentration or dietary bioaccumulation can occur, resulting in an increase of the chemical concentration with increasing trophic level (biomagnification).

1.1.2 Regulation of bioaccumulative chemicals

National and international regulations share a similar approach to identify bioaccumulative substances, for what concerns the set of requested criteria and hazard thresholds (Table 1.1).
All regulations identify bioaccumulative substances on the basis of the bioconcentration (BCF), the octanol-water partition coefficient (K_{OW}) and, only in the case of the Canadian regulation, using the field bioaccumulation (BAF).

Tab. 1.1: Summary of the criteria for regulatory bioaccumulation assessment (B = Bioaccumulative, vB = very Bioaccumulative).

Regulatory Agency	Criterion	Threshold	Judgement
European Union[a]	logBCF	≥ 3.3	B
	logBCF	≥ 3.7	vB
Environment Canada[b]	$\log K_{OW}$	≥ 5.0	B
	logBCF	≥ 3.7	B
	logBAF	≥ 3.7	B
United States[c]	logBCF	≥ 3.0	B
	logBCF	≥ 3.7	vB
United Nations[d]	$\log K_{OW}$	≥ 5.0	B
	logBCF	≥ 3.7	B

[a] REACH Regulation (EC 1907/2006)
[b] Canadian Environmental Protection Act (S.C. 1999, c. 33)
[c] Toxic Substances Control Act (TSCA) and Toxic Release Inventory (TRI)
[d] Stockholm Convention on Persistent Organic Pollutants, 2001

The use of the afore-mentioned criteria is connected to the availability of experimental data. Empirical BCF and BAF, for example, have been estimated to be available only for 4% and 0.2% of compounds, respectively [20], while more complex and realistic data are available only for even less chemicals. However, as pointed out by the recent work of Gobas *et al.* [24], more realistic data could be useful to fully account for the bioaccumulation behaviour of chemicals.

In regulatory frameworks, the QSAR approach is gaining increasing importance for the prediction of ecologic effects and of the environmental fate of chemical substances [25]. Moreover, the recent

European REACH Regulation (EC 1907/2006) recognized its central role in data-gap filling, testing prioritization and reduction of animals used for experimental studies. This is particularly relevant in the case of bioaccumulation-related measures. The determination of BCF, for instance, is very expensive (approximately 35,000 euro per single chemical) and requires the use of more than 100 animals for each standard study [26]. The theoretical bases of QSAR and some details about its role in environmental modelling and in REACH Regulation will be described in the next paragraphs.

1.2 Quantitative Structure-Activity Relationship (QSAR)

1.2.1 Theoretical Background

The cornerstone of the QSAR (Quantitative Structure-Activity Relationship) approach is the principle of similarity, which asserts that similar molecules are likely to exhibit similar biological properties [27]. On this basis, the QSAR approach exploits mathematical and statistical techniques to find an empirical relationship between the molecular structure and the property of interest, as follows:

$$P_i = f(x_i, x_i, ..., x_p)_i \qquad (1.5)$$

Where P_i, is the property of the *i*-th compound and $x_1,...,x_p$ are the predictors; *f* represents the mathematical relationship (i.e. model).
The independent variables are the so-called molecular descriptors, which can be defined as "the final result of a logic and mathematical procedure that transforms chemical information of a molecule, such as structural features, into useful numbers or the result of standardized experiments" [28]. The modelled property can be either physico-chemical (Quantitative Structure–Property Relationship, QSPR) or biological (QSAR); it can be continuous (regression modelling) or categorical (classification modelling).

Some of the key elements of the QSAR approach are (1) use of molecular descriptors, (2) selection of variables, (3) model validation and (4) applicability domain assessment. These aspects will be briefly discussed in the following paragraphs.

1. Molecular Descriptors

One of the most important aspects of QSAR modelling and related fields (e.g. virtual screening) is how to capture and convert the structural information of molecules into one or more meaningful numbers. Molecular descriptors are formally a mathematical representation of molecules obtained by a well-specified algorithm applied to a defined molecular representation or a well-specified experimental procedure [28].

Because of their numeric nature, molecular descriptors are the bridge between chemistry and quantitative sciences, such as mathematics, statistics and chemometrics. In this way, with the aid of different scientific disciplines, their use has permitted for the first time to link experimental knowledge to theoretical information arising from the molecular structure [29].

On the basis of the molecular representation and the algorithm they derive from, molecular descriptors can encode a wide spectrum of different chemical information. For the sake of simplicity, they are generally grouped conceptually according to the molecular representation they derive from [29]:

- *0D descriptors.* To this class belong all the descriptors for which no information about molecular structure and atom connectivity is required. Some examples of 0D descriptors are atom and bond counts, and sum or average of atomic properties. 0D descriptors can be easily calculated, naturally interpreted and do not require optimization of the molecular structure.

- *1D descriptors.* They are calculated from sub-structural information about the molecule. The most known 1D descriptors are counts of functional groups and substructure fragments, as well as atom-centred descriptors and fingerprints.

- *2D descriptors.* These descriptors are based on a graph representation of the molecule and encode graph–theoretical properties (e.g. adjacency, connectivity). They usually derive from a H-depleted molecular graphs and are sensitive to structural features such as size, shape, symmetry, branching, and cyclicity. They are divided into two categories: (1) topostructural descriptors, encoding only graph information, and (2) topochemical indices, which also encode specific chemical properties of atoms, e.g. mass or hybridization state.

- *3D descriptors.* 3D desciptors derive from a geometrical representation of the molecule, i.e. from the *x*–*y*–*z* Cartesian coordinates of its atoms. 3D descriptors have a high information content, but there are also several drawbacks: (1) the geometry optimization is required and the coordinates values can depend on the optimization method, (2) for highly flexible molecules, several minimum energy conformations are available, which can lead to very different descriptor values, (3) the molecular pose of binding can differ from the optimized geometry. For these reasons, in this work only 0- to 2D descriptors have been used.

Molecular descriptors are generally calculated through dedicated software and their number can be very large. For instance, Dragon 6 [30] can calculate up to 4,885 descriptors. Because of this, variable selection approaches play a fundamental role in identifying the relevant descriptors for QSAR modelling, as described below.

2. Variable selection

Nowadays, it is possible to calculate thousands of different molecular descriptors. However, according to Ockham's law of parsimony, it is reasonable to assume that only a small number of them are relevant in determining the modelled property. Furthermore, the increase in the number of model variables is known to generally improve the fitness to the training data, but it often causes a reduction of the predictive ability (overfitting). On the other hand, if the model is too simple, the bias will increase, and the model will not be able to capture important relationships between predictors and response, leading to underfitting. Thus, finding the optimal subset of variables is fundamental to maximize the model predictive ability and its robustness [31].

Variable selection (VS) techniques aim to explore/exploit the variable space in order to select the optimal subset of variables for building the model of interest. This is done by searching for the best trade-off between bias (i.e. model simplicity) and variance (i.e. data description). Throughout the years, many different methods have been proposed, which take inspiration from different fields, such as Darwin's theory of evolution (Genetic Algorithms [32]), the annealing process of metals [33], the movement of flock of birds [34] or of ants [35]. Because of the large number of available methods, only those that were used in the present work will be briefly described below.

All Subset Selection (ASM) It is the simplest approach, which consists in generating all the possible combinations of the p variables. This method – in principle – guarantees to find the best subset of variables, but it is computationally very consuming. Even when generating models up to a maximum size (V), the approach is still extremely demanding. For example, if $p = 130$ and $V = 20$, the total number of models to generate is 2.04×10^{23}. Assuming that we can compute 10,000 models per second (a reasonable estimate for current laptops), the time required to compute all the models would be 6.46×10^{11} years, which means we should have started long before the Big Bang to have the calculation completed by now [36]. For this reason, ASM was used only to explore the combination of small subsets of variables.

Genetic Algorithms (GA) GA ([32], [37]) are inspired from Darwin's theory of evolution. In analogy with biological systems, each chromosome (model) is a sequence of genes (variables) that evolves through two processes: (1) crossover, in which pairs of chromosomes generate offspring sharing some of the parent genes according to a crossover probability; and (2) mutation, in which some genes can change according to a mutation probability. Every time a new chromosome with a better fitness function than the already existing ones is generated, it enters the population and the worst model is discarded. In this way, chromosomes compete against each other and only the fittest survive, in analogy with Darwin's concept of "survival of the fittest".

Reshaped Sequential Replacement (RSR) RSR [36] was recently proposed as an improvement of the sequential replacement (SR) method proposed by Miller in 1984 [38]. RSR and SR share their replacement core, consisting in replacing each variable included in a model one at a time with each of the remaining variables and see whether a better model is obtained; this procedure is reiterated until convergence. According to RSR algorithm, in particular, a population of models is generated from a chosen minimum size up to a chosen maximum size. The model generation can be random or biased towards promising models. The models evolve through the replacement and several functions are implemented to decrease the probability of overfitting and increase the probability of converge to optimal models. RSR is characterized by a good compromise between exploration and computational time [39].

Several metrics can be used to select the best models during the selection of variables. The best practice is to use measures of model predictivity (e.g. in cross-validation) rather than fitting to the training data (see next paragraph).

3. Model validation

Because of the role of QSAR models in predicting the environmental and toxicological behaviour of chemicals, the reliability of their predictions is of primary concern. The so-called model validation consists in the quantitative assessment of their robustness and predictive power. Generally, two types of validation can be performed: (a) internal validation, (b) external validation. The dataset is commonly divided in the so-called training set, used to calibrate (and internally validate) the models and in the test set, used in a second stage only for the final validation.

Internal validation The internal validation is usually performed through the so-called cross-validation. In a typical cross-validation, the n objects of the training set are divided in G cancellation groups of equal size (s_g). Each *g*th group is, in turn, excluded from the model development phase and the model is calibrated using the $n - s_g$ remaining chemicals. The chemicals belonging to the *g*th cancellation group are used to test the ability of the calibrated model to predict their properties. This procedure is repeated until each chemical is excluded once. Different types of statistics can be then computed to quantify the prediction accuracy. When $s_g = 1$, the cross-validation is defined as Leave-one-out (LOO), while if $s_g > 1$, the cross-validation is Leave-more-out (LMO). It is generally accepted that, for large n values, the LOO procedure can be too optimistic and overestimate the real predictive ability of models. In the present dissertation, the well-established 5-fold cross-validation was used.

External validation External test objects can be arbitrarily selected from the pool of available chemicals or can be data whose response becomes available in a later stage. Their role is to obtain a measure of predictivity that is independent from the data/strategy used to calibrate the model. The external test set should be used only in the final stage to test the selected model(s). As for the case of cross-validation, the response of external objects is predicted and different metrics can be calculated to quantify the predictivity. In this dissertation, the parameters will be discussed on a case-by-case basis.

For what concerns training/test data splitting, several strategies have been proposed, such as cluster analysis [40], optimal design [41], and similarity/diversity algorithms [42]. However, a drawback is that external objects are selected using the information about the chemical space of the training set and thus, are not really external to the data used for model calibration [43]. For these reasons, in this work, the external validation sets were obtained by a random selection.
Finally, it is relevant to underline that the predictive ability can be quantified only for test compounds that fall within the model chemical space (i.e. applicability domain), as described in the next paragraph.

4. Applicability Domain assessment

QSAR models heavily rely on the chemical information of the training set to find a relationship between structure and property. The consequence is that, generally, reliable predictions are limited to to query chemicals that fall within the chemical space of the model, i.e. those that are structurally similar to the training compounds. The model interpolation space, where the property can can be reliably predicted, is defined applicability domain (AD). The AD assessment establishes whether the theoretical assumptions of the models are met for new chemicals to be predicted. A proper characterization of the AD, thus, plays a crucial role in (a) quantifying the real predictive ability of the model (i.e. only compounds within the AD should be considered to validate the model) and (b) assess the reliability of the predictions for new, non-tested chemicals.

The AD assessment is a non-trivial issue, as it strongly depends on the nature of the modelling approach and on the characteristics of the dataset. Hence, the AD assessment is generally determined on a case-by-case basis. A comprehensive survey can be found in the recent work of Sahigara *et al.* [44].

1.2.2 REACH Regulation and OECD principles

REACH (Registration, Evaluation, Authorisation and Restriction of Chemicals, EC 1907/2006) is the European regulation that addresses the production and use of chemical substances, and their potential impacts on human health and the environment.

Since the advent of REACH, QSAR methods have been in the spotlight of industries and regulatory agencies. REACH places on the same footing the experimental tests and the QSAR models, as far as the latter provide the same level of information. Vertebrate testing is regarded as the last resort, to be considered only after having gathered all existing information on physico-chemical, toxicological and ecotoxicological properties of a substance, including information generated by QSAR models. Moreover, the European Chemicals Agency (ECHA) stated that "*a QSAR is not only a model, but is associated with an underlying dataset. As a representation of this dataset, the model averages the uncertainty over all chemicals. Thus, it is possible for an individual model estimate to be more accurate than an individual measurement*" [45], further encouraging the usage of QSAR models.

In particular, QSAR models can be applied to [45]: (1) set testing priorities, (2) design experimental strategies, (3) improve the evaluation of existing test data, (4) gain mechanistic information, (5) fill data gaps for hazard and risk assessment. Point no. 5 includes the identification of PBT (Persistent Bioaccumulative and Toxic) or vPvB (very Persistent, very Bioaccumulative) substances, which is considered as a priority by REACH.

In order to facilitate the regulatory application of QSAR by governments and industry and to improve their acceptance, the OECD (Organization for Economic Collaboration and Development) has developed five principles for the validation of QSAR models [46]. These principles are indicated by the REACH as the systematic framework for describing and evaluating the characteristics of a QSAR model. They were agreed by OECD member countries at the 37^{th} Joint Meeting of the*Chemicals Committee and Working Party on Chemicals, Pesticides and Biotechnology* in November 2004. The principles aim to ensure a transparent validation process and an objective determination of the reliability of QSAR models and are outlined below.

1. A defined endpoint This principle aims to ensure clarity in the endpoint being predicted by a given model, since a given property could be determined by different experimental protocols and under different experimental conditions.

2. An unambiguous algorithm This principle aims to ensure transparency of the algorithm that generates the predictions. Without this information, the performance of a model cannot be independently established and it represents a barrier for regulatory acceptance.

3. A defined domain of applicability The need to define an applicability domain is connected to the fact that QSAR models are reductionist and, thus, inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions.

4. Appropriate measures of goodness-of–fit, robustness and predictivity This principle refers to the need of quantify the fitting to the training data, the robustness (in order to avoid overfitting) and its performance in predicting new data.

5. A mechanistic interpretation, if possible It is widely recognised that it is not always possible to provide a mechanistic interpretation of a QSAR model, or that there can be multiple interpretations. However, the possibility of a mechanistic association between the descriptors used and the predicted endpoint should be taken into consideration and documented.

The framework defined in these paragraphs constitutes the starting point of the present dissertation. Next chapter will give a brief overview of the aims and the structure of the project.

Research Framework 2

> *The best scientist is open to experience and begins with romance – the idea that anything is possible.*
>
> — **Ray Bradbury**
> Writer

This dissertation grounds on the state-of-the art of QSAR models for predicting the aquatic bioaccumulation of organic chemicals, aiming to address some of its current open problems. Bioconcentration and dietary bioaccumulation were addressed separately and this allowed to gather mechanistic knowledge about each of the processes. For each topic, some gaps, problems or deficiencies in previous literature were identified and then addressed by exploiting QSAR and chemometric approaches.

This chapter briefly describes the motivation of the project, its contributions to the field of knowledge and describes the thesis outline.

2.1 Problem Statement

The present work stems from some considerations about (1) the criteria requested for assessing the bioaccumulation and (2) the existing QSAR models for their prediction. These considerations are summarized below.

K_{OW} Because of the lack of experimental BCF/BAF data, a large number of compounds are screened according to their K_{OW} values. Despite K_{OW} is, in the majority of cases, a good predictor for bioaccumulation in aquatic organisms, it is unable to account for any biotransformation or elimination process of the chemical in the organism. Hence, using K_{OW} alone can lead to overestimation of the real bioaccumulation. Even more important, some contaminants could be stored within non-lipid tissues and this would lead to an underestimation of the real bioaccumulation potential. This could be the case, for instance, of Perfluorinated Alkyl Acids, which are hypothesized to have increased interactions with serum albumin, liver fatty acid-binding proteins and phospholipids bilayers [47].

BCF The critical aspects regarding K_{OW} reflect on BCF prediction. In fact, the majority of models for BCF prediction mainly rely on K_{OW} or related descriptors [48]. This potentially leads to the same errors of underestimation/overestimation of K_{OW}. Moreover, the BCF does not include dietary exposure and, in some cases, it can be a poor descriptor for biomagnification in food webs [20].

Other bioaccumulation criteria (e.g. BAF, BMF) Even when BCF data are available/predicted, they do not give any information about the bioaccumulation through diet. In the case of the Canadian Regulation, the BAF can be used as a screening criterion, but the availability of data (and, correspondingly, of models) is very limited. Dietary bioaccumulation data or other food web accumulation parameters could be extremely useful for gaining additional insights about the bioaccumulation properties of chemicals [24]. However, only a few QSAR models for small sets of molecules are available ([49], [50]).

2.2 Contributions

In the light of the previous considerations, the original contributions of this dissertation are:

1. The analysis of the accuracy of the K_{OW}-based prediction of BCF, with a focus on the mechanistic processes underlying the observed errors.

2. The development of a structure-based approach to detect compounds whose BCF could be overestimated/underestimated by K_{OW}; this gave the opportunity to gain insights into the mechanisms of bioconcentration.

3. The development of an expert system for BCF prediction in regulatory contexts, which combines the advantages of three existing QSAR models.

4. The comparison of bioconcentration (BCF) and dietary bioaccumulation data (BMF), with a focus on environmentally relevant classes of compounds.

5. The development of a QSAR model to predict the dietary bioaccumulation (BMF) of organic chemicals.

Particular attention was posed to (1) data curation, (2) model simplicity and (3) mechanistic interpretation of the results.

2.3 Research Boundaries

Fish was the model target because of its role in the trophic chain, the availability of standardized testing protocols and its being a benchmark for ecotoxicity tests in regulatory contexts. This reflects in a large availability of experimental data. Investigating the bioaccumulation in other aquatic organisms (e.g. algae) would be of interest, but it is limited by a lower availability of data. By focusing on fish, we were able to analyse the largest chemical space available.

In this work, only organic compounds were targeted. Inorganic compounds, organometallic or coordination compounds were excluded from the analysis despite the environmental relevance of some of them (e.g. methylmercury). This because many descriptors cannot be calculated for heavy atoms. The same problem concerns disconnected structures (i.e. mixtures and salts), for which descriptors can be calculated for one component only.

2.4 Thesis Structure

The dissertation is divided in two parts, reflecting the topics covered: (1) bioconcentration (Chapters 3-5); (2) dietary bioaccumulation (Chapters 6-7). The content of each chapter is outlined below.

Chapter 3

Nine existing models for BCF (four K_{OW}-based and five descriptor-based) are here compared in order to test whether the increased complexity is outweighed by an increased accuracy. To this end, experimental BCF data for 1056 compounds, along with experimental/predicted K_{OW} values, were collected and used for the comparison.

Chapter 4

This chapter presents a classification scheme to predict whether the BCF of a chemical can be predicted well by K_{OW} or may be over- or underestimated. The scheme consists of two QSAR classification trees, which are simple and interpretable. The scheme served to gather new insights into the mechanisms of bioconcentration.

Chapter 5

This chapter integrates the proposed classification scheme with existing BCF models for the regulatory prediction of BCF. The resulting expert system showed an improved accuracy on external data.

Chapter 6

In this chapter, a manually curated set of BMF values for 214 compounds is compared with BCF and K_{OW} values, highlighting that, for some chemical classes, the BCF alone may not be enough to predict the actual aquatic bioaccumulation.

Chapter 7

This chapter presents a set of QSAR models developed for BMF prediction. It provides details about the modelling techniques, the validation steps, the AD assessment and the mechanistic interpretation of the results.

Part II

Bioconcentration

3

QSAR models for bioconcentration: complexity *versus* accuracy

> *Frustra fit per plura quod potest fieri per pauciora.*
> (It is futile to do with more things what can be done with fewer.)
>
> — **William of Ockham**
> Philosopher

Complexity is often perceived by decision-makers as a limitation for QSAR models, which are sometimes seen as black boxes, and it can become a reason of scepticism. On one hand, the increase in the complexity (e.g. number of included variables, chosen machine-learning technique, use of non-linear methods) can increase the fitting to the data but, on the other, it tends to lower models transparency and interpretability, thus negatively affecting their application and acceptance, especially in regulatory contexts. For these reasons, increased model complexity should be balanced by a remarkable increase in the prediction accuracy and/or in the number of biological processes accounted for.

Aquatic bioconcentration, in particular, is mainly driven by a passive partitioning between water and lipids. Hence, most of the models rely on estimates of hydrophobicity, such as the octanol-water partition coefficient (K_{OW}), which mimics the partitioning from water to lipids (lipophilicity). For this reason, many K_{OW}-based models for BCF can be found throughout the scientific literature [48]. The first ones to be proposed relied solely on an empirical relationship between K_{OW} and BCF. Later, more complex QSAR models (based on more parameters and on a variety of different regression techniques), have been proposed; the majority of them still uses K_{OW} as principal predictor.

Despite lipophilicity is the major driving force of bioconcentration, other processes can significantly contribute, such as metabolism, elimination, and specific interactions with tissues different from lipids [51]. Chemicals that are metabolized into hydrophilic compounds, for example, can be eliminated faster and thus have BCF lower than predicted from K_{OW} ([52], [53]). Chemicals that establish specific interactions with non-lipid tissues can have BCF larger than predicted from K_{OW}, such as methylmercuric chloride, which has a low K_{OW} but a very high BCF (up to 1,000,000 in fish) due to its association with protein sulfhydryl groups [54].

The main aim of this chapter is to compare empirical K_{OW}-based equations and descriptor-based QSAR models in order to test whether:

1. The increase in the complexity of descriptor-based models is balanced by a corresponding increase in the prediction accuracy.

2. The increased complexity allows to take into account processes different from water-lipid partitioning.

To this end, we collected and manually curated a BCF dataset of 1056 compounds and compared 4 K_{OW}-based equations with 5 widely-used complex QSAR models.

3.1 BCF models

3.1.1 K_{OW}-based equations

K_{OW}-based equations use predicted or experimental K_{OW} as the independent variable for BCF prediction. Starting from the late 1970s, many K_{OW}-based equations for BCF (of different degree of complexity) have been proposed [48]. The first models to be developed were linear and calibrated on moderately hydrophobic compounds (logK_{OW} from 0 to 6), for which a proportional increase of logBCF when increasing logK_{OW} is generally noted. Higher order models were developed in later stages to address the so-called cut-off problem, that is, the tendency of highly hydrophobic substances (logK_{OW} > 6) to show a decreasing BCF with increasing K_{OW}. The reasons of the observed cut-off are still under debate and, while some authors ascribe it to reduced bioavailability in water (e.g. Wen *et al.* [55]), others attribute it to artefacts in BCF measurement [56].later

Among the linear models, two of the most widely-known equations were analysed, namely those of Veith *et al.* [16] and of Mackay [17] . Among the non-linear models, we chose: (a) the bilinear relationship of Bintein *et al*. [57], which resulted to give the best predictions when compared with a linear and a parabolic model by the authors; (b) the set of equations suggested by the European Technical Guidance Document (TGD) [58], which combines equations of different complexity, according to the range of K_{OW} of the chemical.

Models are graphically depicted in Fig. 3.1 and mathematical details can be found in Box 3.1.1.

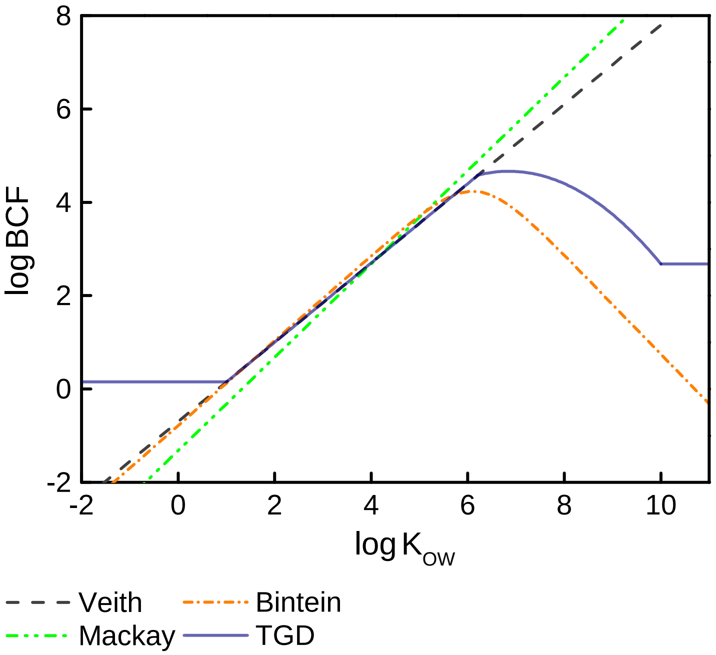


Fig. 3.1: Graphical representation of K_{OW}-based equations.

Veith

$$\log BCF = 0.85 \cdot \log K_{OW} - 0.70$$

Mackay

$$\log BCF = \log K_{OW} - 1.32$$

Bintein

$$\log BCF = (0.91 \cdot \log K_{OW}) - 1.98 \cdot \log(6.8 \cdot 10^{-7} \cdot K_{OW}) - 0.79$$

TGD

$$\begin{array}{ll} \log BCF = 0.15 & \log K_{OW} < 1 \\ \log BCF = 0.85 \cdot \log K_{OW} - 0.70 & \log K_{OW} \in [1, 6] \\ \log BCF = 0.20 \cdot (\log K_{OW})^2 + 2.74 \cdot \log K_{OW} - 4.72 & \log K_{OW} \in (6, 10] \\ \log BCF = 2.68 & \log K_{OW} > 10 \end{array}$$

Box. 3.1.1: K_{OW}-based equations.

3.1.2 Descriptor-based models

More complex QSAR models (i.e. descriptor-based) were developed in later stages, as a reflection of the improvements in machine-learning/modelling techniques and of the increasing number of available molecular descriptors [59]. These factors allowed for the development of more sophisticated and thoroughly validated models, usually based on a larger number of data.

Nowadays, many QSAR models for BCF exist [48], some of which have been implemented in freely-available software, targeting at non-expert usage in regulatory contexts. Among them, we chose the most widely-known ones, namely:

1. EPI Suite [60], developed by U.S. EPA (Environmental Protection Agency), which is considered the benchmark QSAR software for environmental endpoints. EPI Suite implements the Meylan BCFBAF model for BCF prediction.

2. VEGA [61], an open-source platform that integrates several literature-based QSAR models, developed by Istituto Mario Negri. VEGA contains three model for BCF prediction: Meylan, CAESAR and Read-Across. Moreover, the platform integrates an Applicability Domain assessment, quantified by the Applicability Domain Index (ADI). ADI ranges from 0 (compound outside the AD) to 1 (compound inside the AD) and takes into account factors such as the presence of similar molecules in the training set, the concordance of the prediction among similar molecules and the model descriptor range.

In addition, a recently-proposed *consensus* model [62] between VEGA CAESAR and VEGA Meylan was analysed. Details about the models can be found below and a graphical scheme can be found in Fig. 3.1.

Meylan models EPI Suite BCFBAF and VEGA Meylan models derive from the method of Meylan and co-authors [63]. Compounds are first classified as ionic or non-ionic: the formers are assigned a fixed BCF value according to their K_{OW}, while for the latter, BCF is predicted using K_{OW} and then corrected according to the presence of specific structural fragments (see Box 3.1.2). Ionic compounds include carboxylic and sulfonic acids, salts of sulfonic acids, and charged nitrogen compounds. All other compounds are classified as non-ionic. EPI Suite and VEGA Meylan differ slightly in their training set and in the model used to predict K_{OW}.

CAESAR model CAESAR ([64], [65]) is an hybrid QSAR model, based on the combination of two distinct sub-models, A and B. The sub-models share their principal descriptors, which are related to K_{OW}: (1) MlogP, the octanol-water partition coefficient calculated by Moriguchi model [66] and (2) BEHp2 [67], related to the atomic polarizability and thus to hydrophobicity. Moreover, they both comprise 2D autocorrelation descriptors ([68], [69]), which describe the distribution of a considered property along the molecular structure (GATS5v and MATS5v for model A and B, respectively) and descriptors referred to the presence and number of chlorine atoms (Cl-089 and SsCl for model A and B, respectively). Finally, model A comprises the descriptor AEige, a topological descriptor, while model B comprises the descriptor X0sol, considered as a total measure of the molecular electronegativity (the higher the electronegativity, the greater the separation of molecular charge and therefore the greater the hydrophilicity). The predictions are then combined according to the arithmetic mean of the BCF values provided by the sub-models (see Box 3.1.2).

VEGA *consensus* Gissi and co-authors [62] proposed a *consensus* combination between VEGA CAESAR and VEGA Meylan for regulatory purposes. Two criteria are used: the absolute difference between the two predictions (ΔBCF, in log units) and the ADI associated with each prediction.

1. If $\Delta\text{BCF} \leq 1$: the predictions are similar and considered as reliable, thus the highest BCF is chosen (conservative approach).

2. If $\Delta\text{BCF} > 1$: ADI is considered.

- ADI ≥ 0.7 for at least one model: the prediction with the largest ADI is used. In case of identical values, the largest predicted BCF is chosen.

- ADI < 0.7 for both the models. No prediction is provided.

Since the model is not available online, the predictions of the single VEGA models were combined in a MATLAB [70] workflow.

Meylan model

Non-ionic compounds

$$\begin{aligned}
&\log BCF = 0.50 && \log K_{OW} < 1 \\
&\log BCF = 0.66 \cdot \log K_{OW} + \sum CF && \log K_{OW} \in [1, 7] \\
&\log BCF = -0.49 \cdot \log K_{OW} + 7.55 + \sum CF && \log K_{OW} > 7
\end{aligned}$$

Ionic compounds

$$\begin{aligned}
&\log BCF = 0.50 && \log K_{OW} < 5 \\
&\log BCF = 1.00 && \log K_{OW} \in [5, 6) \\
&\log BCF = 1.75 && \log K_{OW} \in [6, 8) \\
&\log BCF = 1.00 && \log K_{OW} \in [8, 9) \\
&\log BCF = 0.50 && \log K_{OW} > 9
\end{aligned}$$

CAESAR model

$$\begin{aligned}
&\log BCF = 0.94 \cdot \log \overline{BCF} - 0.12 && \log \overline{BCF} \leq 1.36 \\
&\log BCF = 1.00 \cdot \log BCF_{min} && \log \overline{BCF} \in (1.36, 2.41] \\
&\log BCF = -1.05 \cdot \log \overline{BCF} && \log \overline{BCF} > 2.41
\end{aligned}$$

Box 3.1.2: Equations used to predict BCF by (i) Meylan model, *CF* represents the numerical correction factor of Meylan, corresponding to the presence of a given molecular sub-structure (detailed information can be found in EPI Suite BCFBAF/VEGA user guides); (ii) CAESAR model, $\log \overline{BCF}$ and $\log BCF_{min}$ represent respectively the arithmetic mean and the minimum predicted BCF values of CAESAR sub-models.

VEGA Read-Across VEGA Read-Across is based on the similarity principle, which states that similar molecules are likely to have similar properties. In particular, the model is based on a similarity index (SI), calculated on the basis of molecular extended fingerprints [71], some constitutional information (number/type of atoms and number/type of bonds) and a smaller contribution of functional and heteroatoms descriptors [72]. The experimental BCF values of the three compounds most similar to the target according to SI are used to predict the BCF as weighted average, using the SI as the weighting factor.

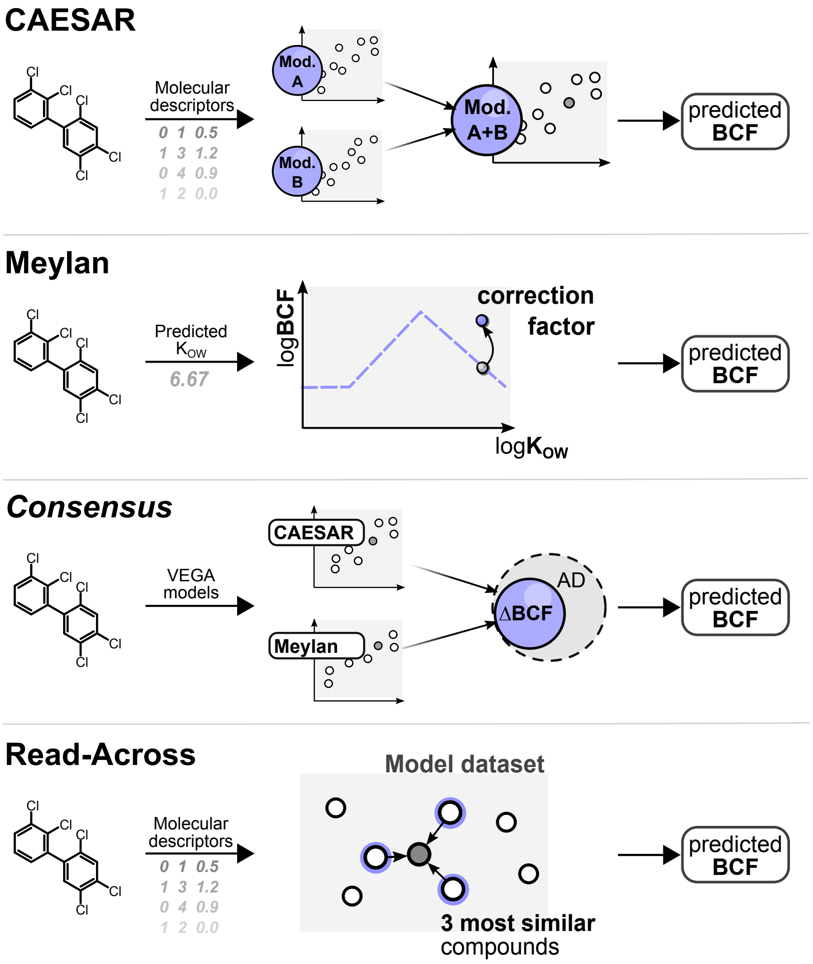


Fig. 3.2: Graphical scheme of the chosen descriptor-based models.

3.2 Data and curation

Two wet-weight BCF datasets were generated, namely: (1) a “merged" dataset (1011 compounds), obtained from the training data of the different tested models; (2) an “external" dataset (45 compounds), comprising new molecules external to all models under investigation (Fig. 3.3).

The merged dataset offered the advantage of using many compounds to quantify the model performance. The external dataset, which was manually-curated, offered the opportunity to test all the models on the same set of unknown compounds and obtain directly comparable statistics.

Some of the tested models were developed on wet-weight BCF data of a single species (e.g. CAESAR), while others used BCF data of different species. Since different authors reported no remarkable difference between interspecific and intraspecific variability on BCF data ([19], [63]) all the available data were used, regardless of the species. Details about data curation steps can be found in the next paragraphs and a simplified scheme is provided in Fig. 3.3.

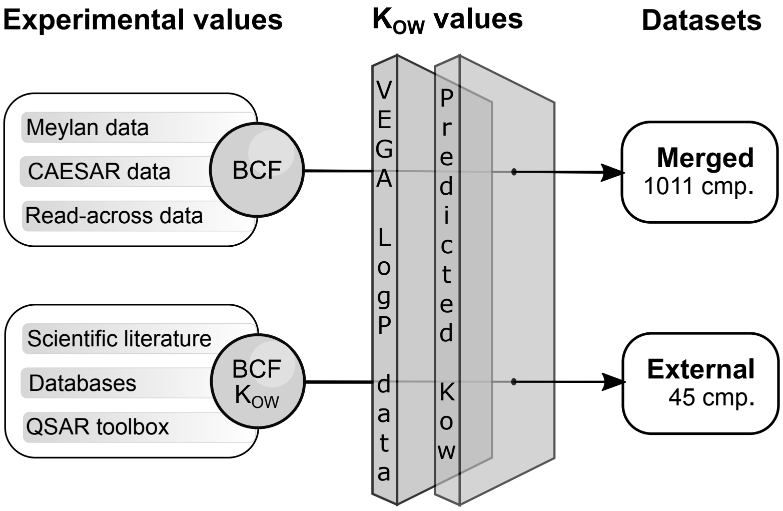


Fig. 3.3: Graphical scheme of data sources for merged and external datasets.

Merged set This dataset comprises 1011 compounds and was obtained by merging the BCF datasets of VEGA models (CAESAR, Meylan and Read-across). The workflow for data curation was the following:

1. Multiple values were merged according to CAS number.

2. The match between CAS and structure was verified through the Chemical Identifier Resolver [73]: in case of mismatch (50 compounds) the accordance between CAS and structure was manually checked on PubChem [74] and ChemSpider [75]. Only records referred to the correct CAS-structure pair were retained.

3. The arithmetic-mean of multiple BCF values for the same compound was calculated, since it was not possible to check all the original sources for the presence of duplicate records. This was done with the exception of compounds contained in both VEGA CAESAR and VEGA Read-across datasets, the latter containing an updated version of CAESAR dataset. In this case, only VEGA Read-across BCF values were retained.

Experimental K_{OW} values were obtained from VEGA logP dataset. If no experimental value was provided, a *consensus* K_{OW} was predicted, as suggested by Cronin and Livingstone [76]. In particular, the arithmetic mean was calculated from the predictions of: (1) Ochem ALOGPs [77]; (2) Dragon AlogP (or MlogP for the 5 Sn-containing compounds)[30]; (3) VEGA logP model (only predictions with ADI $>$ 0.7). Moreover, experimental K_{OW} values for 2 perfluorinated alkyl acids(badly predicted by existing K_{OW} models) were added.

External set This dataset consists of 45 compounds, external to all the models. Values were obtained for environmentally relevant classes of compounds: Synthetic Pyrethroids, Organophosphorous Compounds, Perfluorinated Compounds, Personal Care Products and Polychlorinated Biphenyls. Wet-weight BCF and K_{OW} data were obtained from: (a) the handbook of Mackay *et al.* [78]; (b) published papers; (c) QSAR toolbox [79]; (d) complex molecular database for Environmental Protection [80]; (e) EURAS BCF Gold Standard Database [81]; (f) LOGKOW database [82] ; (g) VEGA logP dataset [61].

The data curation workflow was the following:

1. BCF values were retained according to two rules: (1) only steady state values were accepted; (2) "common" species (e.g. fathead minnow, bluegill, rainbow trout) were preferred to "rare" species, if present.

2. For compounds with multiple values and a standard deviation of the BCF larger than 0.25, original articles were checked.

3. A Q-test [83] was performed at the 95% confidence level to remove outliers.

4. Median BCF value, which is a robust measure of central tendency of data, was calculated. In this case, in fact, unlike the merged dataset, every value came from a single experimental campaign.

5. For K_{OW}, recommended values, if present, were preferred, and the median value was calculated for each compound. For 5 compounds, no experimental K_{OW} value was found, thus a *consensus* prediction was performed as for the merged set.

6. SMILES (Simplified Molecular-Input Line-Entry System) were obtained from name, CASRN and InChI code, using the Chemical Identifier Resolver [73] of KNIME [84]. In case of mismatch, SMILES were manually checked on PubChem [74] and ChemSpider [75].

3.3 Results and discussion

Models were firstly evaluated for their global performance (Section 3.3.1) and then on compounds showing a significant deviation of the experimental BCF from the K_{OW}-driven BCF (Section 3.3.2).

3.3.1 Global Performance

For each model, its "complementary" set was built, namely the set containing the compounds of the merged set not used to build or validate that model. If the SMILES and/or CAS of a given record were present in the training/test sets but they were not correctly matched, the record was excluded from the complementary set of that model. For VEGA *consensus*, the complementary set was built taking into account the compounds belonging neither to CAESAR nor to Meylan datasets. For K_{OW}-based models (generally developed on small training sets or without a training set, as for TGD) the entire merged set was used. For VEGA models, only predictions with ADI $\geq$ 0.70 were considered. Model predictive ability was quantified using the Root Mean Square Error (RMSE), which can be expressed as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.1}$$

where y_i and $\hat{y}_i$ are the experimental and predicted logBCF values, respectively, and *n* is the number of compounds. RMSE was used to evaluate the performance of models on unknown compounds, i.e. those belonging to their complementary set and to the external set (Table 3.1). When available, also the statistics on training and test sets were reported for the sake of comparison.

On the complementary set, all descriptor-based models gave a lower RMSE than the K_{OW}-based equations. However, when VEGA *consensus* and VEGA Meylan are compared with TGD, only a small difference

Tab. 3.1: Global model performance according to the Root Mean Square Error (RMSE) on training, test, complementary and external sets. Number of compounds (*n*) and percentage of compounds outside the AD (% out) are also reported.

Model	**Training set**		**Test set**			**Complementary**			**External**	
	n	RMSE	*n*	RMSE	% out	*n*	RMSE	% out	RMSE	% out
Veith	-	-	-	-	-	1011	1.55	-	1.12	-
Mackay	-	-	-	-	-	1011	1.78	-	1.35	-
Bintein	-	-	-	-	-	1011	1.16	-	1.58	-
TGD	-	-	-	-	-	1011	1.05	-	0.94	-
EPI BCFBAF	516	0.52	146	0.60	-	349	0.94	-	1.33	-
VEGA Meylan	516	0.50	146	0.55	0.0	349	0.99	4.0	0.99	15.6
VEGA CAESAR	378	0.58	95	0.63	9.5	538	0.91	29.9	1.57	62.2
VEGA Read-across	686	0.50	173	0.60	0.6	152	0.81	1.3	0.91	0.0
VEGA *consensus*	-	-	-	-	-	253	0.99	2.8	1.31	0.0

in the prediction error can be noted. VEGA Read-across was the best model both for RMSE (0.81) and percentage of compounds outside its AD (1%). VEGA CAESAR predictions had a low RMSE (0.91), but a large percentage of compounds was outside the AD (30%). VEGA *consensus* did not improve the predictions with respect to VEGA Meylan, but had fewer compounds outside the AD (1.2% less).

Despite the external set comprised only 45 molecules, it was useful to evaluate the models on the same set of compounds. On the external set, only VEGA Read-across gave better predictions than the TGD and VEGA Meylan showed a RMSE similar to that of TGD. Again, for CAESAR a large percentage of compounds were outside the AD (62.2%). The RMSE of VEGA *consensus* had an intermediate value between VEGA Meylan and VEGA CAESAR, with all the compounds inside the AD.

3.3.2 Performance on critical compounds

The TGD model showed the lowest RMSE among the K_{OW}-based equations and, in most of the cases, its performance was comparable with that of complex models. Hence, it was used to represent the K_{OW}-driven bioconcentration. The majority of compounds were correctly estimated, while a large percentage was overestimated and a smaller percentage was underestimated (Fig. 3.4).

Since K_{OW} can be thought of as a proxy for lipid storage, the over- and underestimation can be ascribed to the additional processes that can influence the final BCF, such as biotransformation or additional storage within non-lipid tissues. In some cases, the biotransformation could lead to a bias in the measured BCF, even if the metabolites are bioaccumulative in themselves. This could be a critical aspect to consider when dealing with potentially metabolized compounds.

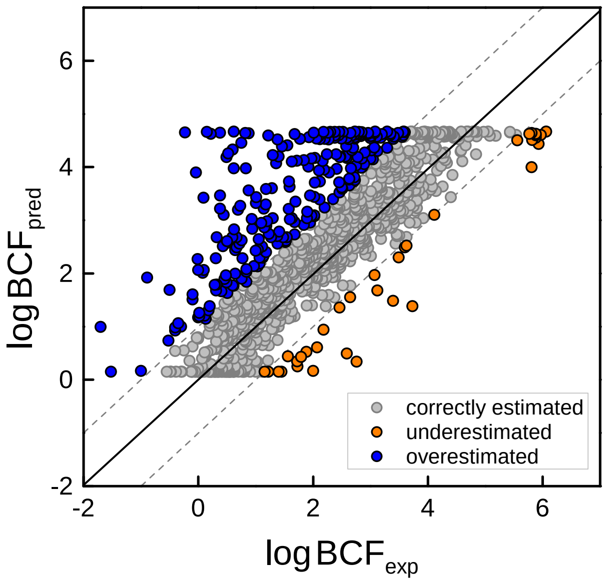


Fig. 3.4: Experimental BCF *vs* BCF predicted by TGD equation. The solid line represents the perfect fitting (logBCF$_{exp}$ = logBCF$_{pred}$); dashed lines represent a difference of ±1 log unit.

A systematic analysis of model performance on critical compounds was performed, by assigning each chemical to one of the following groups:

1. Correctly estimated compounds (789) if the experimental BCF lays within ±1 log unit from the TGD-BCF. As they are correctly predicted by K_{OW}, their bioconcentration can be ascribed to lipid storage.

2. Overestimated compounds (232), if the predicted BCF is significantly larger (> 1 log unit) than the experimental BCF. For these chemicals, we can suppose that metabolism or elimination processes occur, leading to a decreased BCF with respect to that based on affinity with lipids.

3. Underestimated compounds (35), if the predicted BCF is significantly smaller (more than 1 log unit) than the experimental

BCF. For these compounds, it is possible to suppose that storage in additional sites (e.g. proteins) leads to an increased BCF.

We aimed to understand how each model performs on critical compounds (over- and underestimated) in comparison with TGD. To this end, each model (with the exception of TGD, which lacks a proper training set) was evaluated for its performance on unknown data, i.e. complementary plus external data (Table 3.2 and Fig. 3.5).

Tab. 3.2: Model performance on critical compounds: Number of chemicals (n), Root Mean Square Error (RMSE), and percentage of compounds out of AD (% out) are reported.

Class	Model	n	RMSE	% out
Corr. estimated	TGD	789	0.47	-
	EPI Suite BCFBAF	254	0.70	-
	VEGA Meylan	254	0.59	5.5
	VEGA CAESAR	405	0.99	50.4
	VEGA Read-across	141	0.84	0.7
	VEGA *consensus*	194	0.77	3.6
Overestimated	TGD	232	1.98	-
	EPI Suite BCFBAF	116	1.25	-
	VEGA Meylan	116	1.43	3.4
	VEGA CAESAR	151	0.74	62.9
	VEGA Read-across	42	0.74	2.4
	VEGA *consensus*	85	1.38	0.0
Underestimated	TGD	35	1.43	-
	EPI Suite BCFBAF	24	1.89	-
	VEGA Meylan	24	1.63	20.8
	VEGA CAESAR	27	1.43	63.0
	VEGA Read-across	14	1.00	0.0
	VEGA *consensus*	19	1.53	0.0

On overestimated compounds, all models gave a lower RMSE than TGD. VEGA CAESAR and VEGA Read-across had the best performance and the latter has the broadest AD (only 2% outside the AD). Even in this case, VEGA *consensus* showed an intermediate performance between its sub-models, with all the compounds inside its AD.

On underestimated compounds, only VEGA Read-across had better performance than TGD. It is important, however, to highlight that 19 out of 35 underestimated compounds were used to train the model. This is an important aspect, since Read-Across is based on a similarity approach.

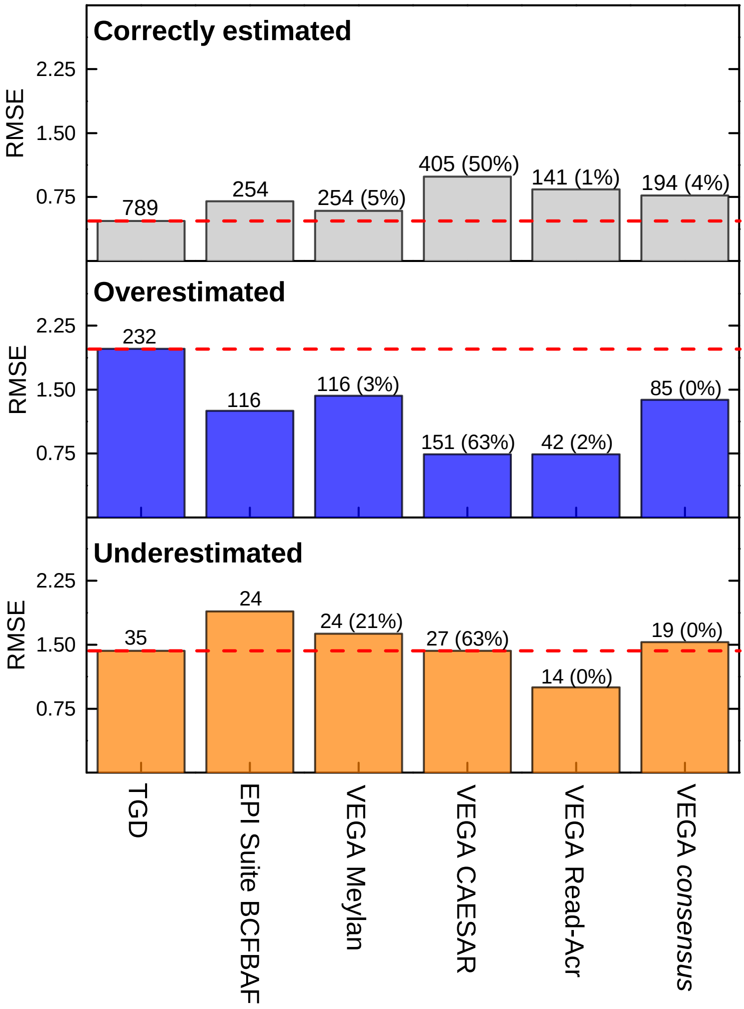


Fig. 3.5: RMSE of descriptor-based models in comparison with TGD RMSE (dashed line). Each bar is labelled by number of compounds used to calculate RMSE and percentage of compounds out of the AD. Values are reported in Table 3.2

The tendency of underestimation and overestimation of TGD (Fig. 3.6) is also present for descriptor-based models models, especially for underestimated chemicals, with the exception of VEGA Read-across. This could suggest that, even when the prediction accuracy improves with respect to TGD, the models take only partially into account processes different from water-lipid partitioning, especially those leading to an excess of bioconcentration.

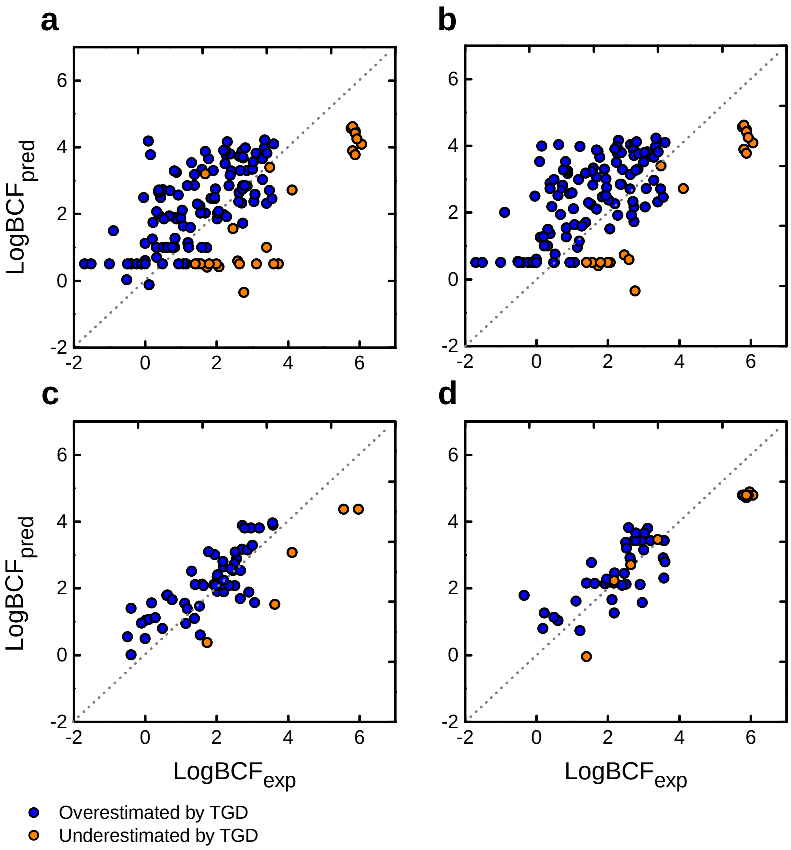


Fig. 3.6: Experimental *vs* predicted logBCF for compounds overestimated and underestimated by TGD (only compounds external to model training sets). Models: (a) EPI Suite BCFBAF; (b) VEGA Meylan; (c) VEGA CAESAR; (d) VEGA Read-across. Dashed lines represent the perfect fitting ($\text{logBCF}_{pred} = \text{logBCF}_{exp}$).

3.4 Conclusions

In this chapter, we compared the performance of four K_{OW} based models and five descriptor-based QSAR models on a dataset of wet-weight BCF data for 1056 compounds. Models were tested first for their global accuracy and then on critical compounds, i.e. those that are badly predicted when only K_{OW} is used.

Results highlighted a general improvement of the accuracy when complex models are used instead of K_{OW}-based equations. Moreover, the AD assessment allows to improve the reliability of the predictions. However, when the complex models were tested on critical compounds, the majority of them showed the same weaknesses of K_{OW}-based predictions. The exception was VEGA Read-across, which, however, is based on a local approach and thus can be applied only to molecules similar to those of the training set. Moreover, Read-Across lacks of a mechanistically interpretable model.

In conclusion, the complex models resulted to be generally more accurate than K_{OW}-based equations but not able to completely account for the processes different from lipid-water partitioning that affect the bioconcentration.

4

Mechanisms of bioconcentration: insights from QSAR classification trees

> *That's been one of my mantras – focus and simplicity. Simple can be harder than complex.*
>
> — **Steve Jobs**
> CEO Apple Inc.

As shown in the previous chapter, some compounds can be predicted well using K_{OW} only. For these compounds, it can be hypothesized that storage occurs within lipid tissues, while, for the others, additional processes occur. The aim of this chapter is to build a classification scheme to identify (1) compounds that can be predicted well using K_{OW}, (2) compounds that are underestimated by K_{OW} and (3) compounds that are overestimated by K_{OW}. Our hypothesis is that this behaviour is connected with different processes of bioaccumulation: (1) main storage within lipid tissues, (2) storage within non-lipid tissues, (3) metabolism/elimination. To our knowledge, the mechanisms that affect bioconcentration have never been investigated extensively in a QSAR setting, and this could give the opportunity to detect compounds with increased impact on biota.

Particular attention was posed to model simplicity and interpretability. To this end, CART (Classification and Regression Trees) [85] was chosen as machine-learning algorithm. CART is based on a recursive partitioning of data using one variable at a time: at each univariate split, data are divided in two mutually exclusive groups (as homogeneous as possible) according to their variable values, and then the splitting procedure is furtherly applied to each group separately. This procedure leads to a model that is graphically representable as a decision tree, where each node is a univariate split and leafs are the predicted classes for the objects that fall in that leaf. In addition to its simplicity and interpretability, CART technique is able to deal with non-linear relationships between variables, thus being particularly well suited for complex biological problems.

4.1 Materials and methods

4.1.1 Dataset

The BCF dataset of Chapter 3 consists of 1056 compounds, among which 654 had known experimental K_{OW} values. Since K_{OW} served to determine the classes, additional values were obtained from the dataset of 16,998 compounds curated by Mansouri [86]. Experimental values were retrieved according to structure and then the accordance between CAS numbers was checked on online databases ([74], [75]); only the correct records were retained. For compounds without experimental K_{OW} value, the new experimental value was added. For compounds with experimental K_{OW} values already retrieved (599), accordance between the values was checked. For 493 compounds, the same values were found. For compounds with multiple values, a test was performed on the standard error. In fact, the precision of the experimental determination of K_{OW} was found to decrease with increasing K_{OW} value [87], with an expected critical standard error (SE_{crit}), expressed as:

$$SE_{crit} = 0.20 - 0.09 \cdot \log K_{OW} + 0.14 \cdot (\log K_{OW})^2 \qquad (4.1)$$

We used this equation as a rule to check for anomalous K_{OW} data. SE_{crit} was compared with the observed standard error (SE_{obs}). If $SE_{obs} > SE_{crit}$ (46 compounds), values were manually checked on publicly available K_{OW} databases ([78], [88]–[90]) and wrong records were deleted. Finally, only compounds with known experimental K_{OW} were retained and the median K_{OW} was calculated. This led to a final dataset of 779 compounds with experimental BCF and K_{OW} values.

4.1.2 Molecular Descriptors

3,763 molecular descriptors (0- to 2D) were calculated using Dragon 6 [30] and reduced by excluding those: (a) with at least one missing value; (b) constant or near-constant; (c) with a standard deviation less than 0.01; (d) with a pairwise correlation larger or equal to 0.98 with other descriptors. In total, 1495 descriptors were retained.

4.1.3 Modelling strategy

Classification trees were grown using the following methodology:

1. Modelling a two-class problem (i.e. one class at a time against the others). This was done for class 2 and 3 and gave considerably better results than the three-class approach.

2. Using an optimization approach, by varying (a) the misclassification cost (from 0 to 0.90 with a step of 0.10) and (b) the splitting criterion (Gini diversity index and cross-entropy [85]).

3. Selecting the optimal tree complexity in cross-validation, by varying the minimum number of objects per leaf from 1 to 100 with a step of 10.

4. Implementing CART classification into a Genetic Algorithms setting. CART, in facts, selects the optimal subset of variables in a stepwise manner. However, when a large number of variables are available, not all of their possible interactions are exploited.

4.2 Results and discussion

Classes were determined using the TGD equation, as in Chapter 3. Each critical class (i.e. class 2 and class 3) was modelled in turn against the other ones. The developed classification trees offered mechanistic insights into the molecular features controlling the mechanisms of bioconcentration. The classification schemes were then combined in a *consensus* manner to obtain the final predictions (Fig. 4.1). Details can be found in the next paragraphs.

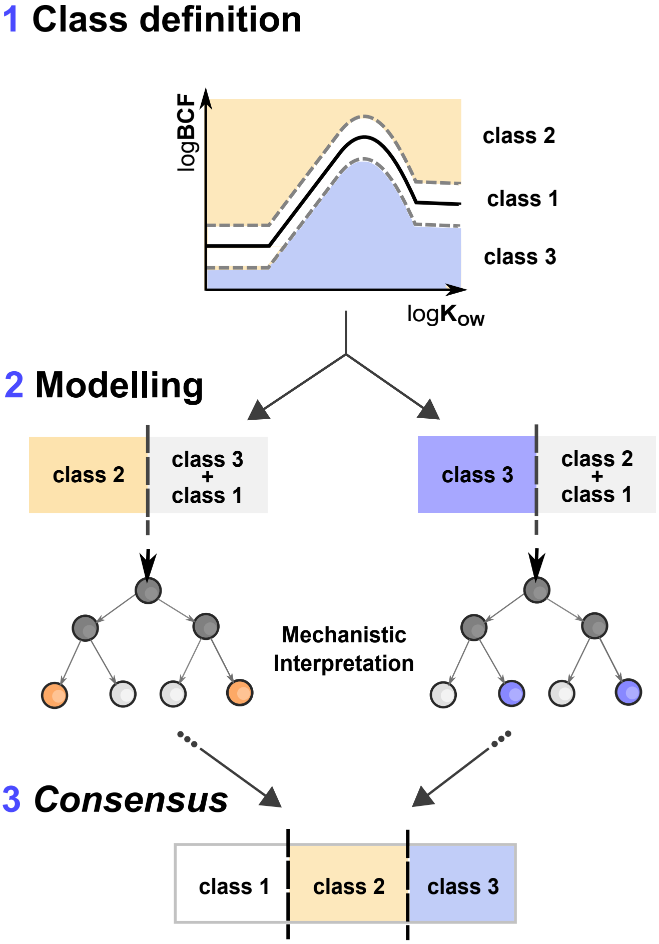


Fig. 4.1: Schematic workflow of the modelling approach.

4.2.1 Class definition

In analogy with the previous chapter, TGD equation was used as a proxy for lipid-driven bioconcentration, as it resulted the most accurate K_{OW}-based equation among the four tested. Again, our hypothesis was that compounds reliably predicted from K_{OW} mainly accumulate within lipids, while others could undergo additional processes.

In Chapter 3, we used a threshold of 1 log unit to divide compounds within three groups as preliminary analysis. In this case, a data-driven calibration of the threshold was performed. In particular, the threshold was chosen as twice of the 95th percentile of the standard error of multiple BCF values (0.24 log units). The chosen threshold was ±0.50 log units, which also resulted larger than the 99th percentile of the standard error (0.47 log units). Classes were, then, defined as follows:

1. Class 1 – Inert chemicals (460). Compounds whose experimental logBCF lays within ± 0.5 log units interval from predicted logBCF. They mainly bioconcentrate within lipid tissues and thus partitioning-related models can be used.

2. Class 2 – Specifically bioconcentrating chemicals (64); if experimental logBCF $\geq$ $\text{logBCF}_{\text{TGD}}$ + 0.5. For these compounds, additionally to lipid storage, specific interactions with tissues can be hypothesized, which lead to underestimation of BCF when K_{OW} or related parameters are used.

3. Class 3 – Less bioconcentrating chemicals (255 compounds): if experimental logBCF $\leq$ $\text{logBCF}_{\text{TGD}}$ - 0.5. The observed deficit of BCF can be connected with biotransformation, which leads to a faster elimination and/or to a bias in the measured BCF.

We do not rule out that for some class 1 compounds, metabolism and/or interactions with tissues occur, but deviations of less than 0.5 log units are not discernible from lipid-driven BCF. At the same time, some of the observed deviations could be due to model error or data uncertainty. Hence, where possible, we rationalized the obtained classifications through literature- and data-driven considerations.

The majority of compounds belong to class 1, while the 33% and only the 8% to class 3 and 2, respectively (Fig. 4.2). Class 2 chemicals lay mainly in logK_{OW} > 0 region and have a higher relative abundance of very bioaccumulative chemicals (logBCF $\geq$ 3.5), confirming that they are a class of concern.

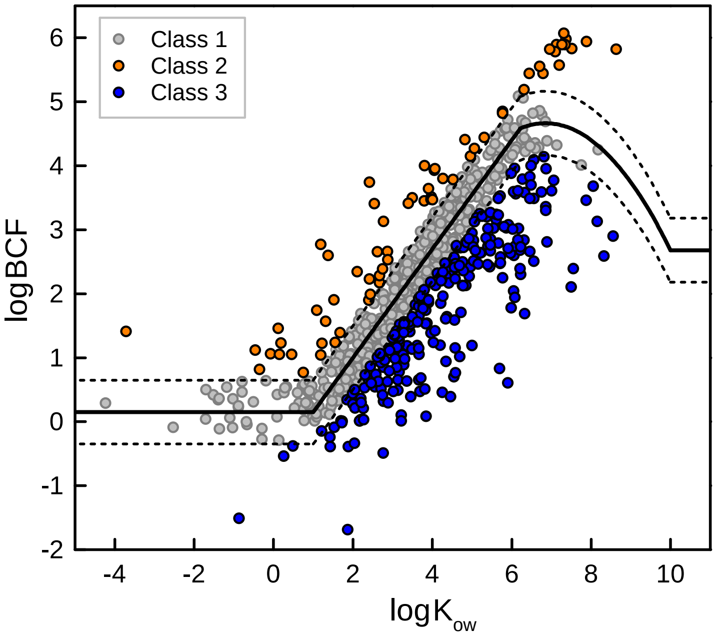


Fig. 4.2: K_{OW} *vs* BCF: solid line represents the predicted BCF across the range of K_{OW} according to TGD approach, while dashed lines represent a ±0.5 log units interval from estimated BCF. Compounds are coloured according to the assigned class.

The distribution within classes (Table 4.1) agrees with what known about some environmental pollutants:

- *Perfluorinated Alkyl Acids* (PFAAs) are hypothesized to have increased interactions with serum albumin, liver fatty acid-binding proteins and phospholipids bilayers ([47], [91]–[94]). Accordingly, they all were assigned to class 2. Perfluorooctane

Sulfonate showed the largest residual of this chemical class (2.35 log units).

- Some *Synthetic Fragrances*, *Pyrethroids* and *Organophosphorous compounds* are known to be metabolized in fish ([95], [96]). All of them are distributed within class 1 and 3, with the exception of glyphosate and dimethyl phenyl phosphate.

- *Polycyclic Aromatic Hydrocarbons* (PAHs) are effectively biotransformed into more hydrophilic compounds in fish liver and can be easily excreted ([97], [98]). Furthermore, the influence of metabolism on their final BCF has already been reported [99]; this agrees with their distribution in classes 1 (60%) and 3 (40%).

- Regarding *Polychlorinated Biphenyls* (PCBs), several studies reported a selective metabolic clearance in fish for a large number of congeners ([100], [101]) with no observed elimination for hexa-, hepta- and octa-CBs ([102], [103]). These considerations agree to some extent with data: 17 PCBs, in fact, belong to class 2 and 14 of them are congeners with 6 to 8 Cl atoms. This could suggest a general excess of BCF for PCBs, particularly evident for those that are not metabolized/eliminated. Alternatively, biases could be ascribed to BCF-determination methodology, as already hypothesized by Wang et al. [104]. Note that, for $\log K_{OW} > 6$, only PCBs (15 molecules) belong to class 2 and, with the exception of two, they are all congeners with 6 to 8 Cl atoms.

- *Ionogenic Organic Compounds* (IOCs) are a critical category of chemicals, as their uptake and elimination depend on hydrophobicity, degree of ionization, electrostatic interactions and steric factors of ionized and unionized forms; moreover, some IOCs may interact with phospholipids or other macromolecules [91]. Despite a deviation from K_{OW}-based BCF is conceivable, IOCs are about the 40% of the total, with almost equal relative distribution within the classes (Table 4.1). This means that no trend of overestimation/underestimation based on K_{OW} is directly associable with ionization. On this basis, ionizing compounds were retained for model development, in order to maximize the structural information available.

Tab. 4.1: Distribution within classes of some environmentally relevant chemicals. (PCBs = Polychlorinated Biphenyls, PAHs = Polycyclic Aromatic hydrocarbons)

Chemicals	**Class** 1	2	3	**Total**
Perfluorinated Alkyl Acids (PFAAs)	0	7	0	7
PCBs	27	17	1	44
Polybrominated Biphenyls (PBBs)	2	0	0	2
Synthetic Fragrances	1	0	2	3
Organophosphates	38	2	33	73
Synthetic Pyrethroids	4	0	12	16
PAHs	9	0	6	15
Ionogenic Compounds (IOCs)*	182	26	110	318

*Selected according to the presence of ionizing functional groups, as in the work of Meylan et al. [63].

4.2.2 Classification results

Model evaluation and validation Compounds were randomly split into a training set of 584 compounds (75%) and a test set of 195 (25%), preserving the proportion between the classes. Training set was used for variable selection, tree pruning (both with 5-fold cross-validation), and model calibration. Test set only served to validate the final pool of models. Model predictivity was quantified using Sensitivity (*Sn*), Specificity (*Sp*) and Non-Error Rate (*NER*), defined as follows:

$$Sn = \frac{TP}{TP + FN} \tag{4.2}$$

$$Sp = \frac{TN}{TN + FP} \tag{4.3}$$

$$NER = \frac{\sum_{i=1}^{G}(Sn_i)}{G} \tag{4.4}$$

where *TP*, *TN*, *FP* and *FN* are the number of true positives, true negatives, false positives and false negatives of each class, respectively; *G* is the number of classes. Tree pruning and GA selection were performed in cross-validation.

Variable selection Each critical class (i.e. 2 and 3) was modelled in turn against the remaining ones. In particular, for each optimization setting (paragraph 4.1.3), we carried out three step procedure: (1) GA selection on the whole pool of descriptors; (2) GA selection on the most frequently retained descriptors of step 1 (maximum 150); (3) generation of all the possible combinations of the most relevant descriptors of step 2 (maximum 15). As fitness-function, we chose the geometric mean between Sn and Sp, aiming at fostering the most balanced models.

Applicability Domain In order to predict test compounds, we characterized the model AD as an hyper-rectangle delimited by maximum and minimum values of each descriptor, by what is known as "bounding-box" approach [44]. Only compounds within the AD were predicted.

Model selection Best models were chosen according to: (1) performance in cross-validation and on the test set (in terms of Sn, Sp and NER), (2) complexity (the fewer nodes, the better) and (3) descriptor interpretability. This in order to produce predictive/robust models and gather new mechanistic insights about the structural features that characterize the classes.

Final models The selected models (Table 4.2) are very simple, comprising few univariate splits (i.e. nodes) and few variables. In both cases, only one test compound (mirex and carbon disulfide for class 2 and 3 trees, respectively) was outside the AD. Best performance was obtained on the most critical class. Class 2 compounds are, indeed, underestimated by K_{OW}. On the contrary, class 3 is the less critical, since it comprises compounds with reduced BCF. For this reason, the slightly lower performance of T3 is acceptable. The selected classification trees are analysed in depth and interpreted in the next paragraphs.

Tab. 4.2: Statistics of selected classification trees: characteristics of the model (k = number of nodes, p = number of descriptors, C = cost value for the target class) along with *Sn*, *Sp* and *NER* in fitting, cross-validation and on the test set. For the test set, number of compounds out of AD (n out) was also reported.

ID	Target Class	*k*	*p*	C		FIT	CV	TEST
T2	2	9	5	0.90	*NER*	0.85	0.75	0.75
					Sn	0.94	0.80	0.75
					Sp	0.76	0.71	0.75
T3	3	5	4	0.60	*NER*	0.73	0.72	0.78
					Sn	0.70	0.70	0.69
					Sp	0.76	0.75	0.68

4.2.3 Model analysis and interpretation

1. Specifically bioconcentrating chemicals (T2)

Fig. 4.3 depicts T2 tree, targeting at the classification of class 2 compounds. According to what observed on PCBs, first split regards those having 6 to 8 Cl atoms and $K_{OW} > 6$, which are assigned to class 2. T2 comprises additional 8 nodes based on 5 molecular descriptors resulting from GA selection. Descriptors are briefly described below.

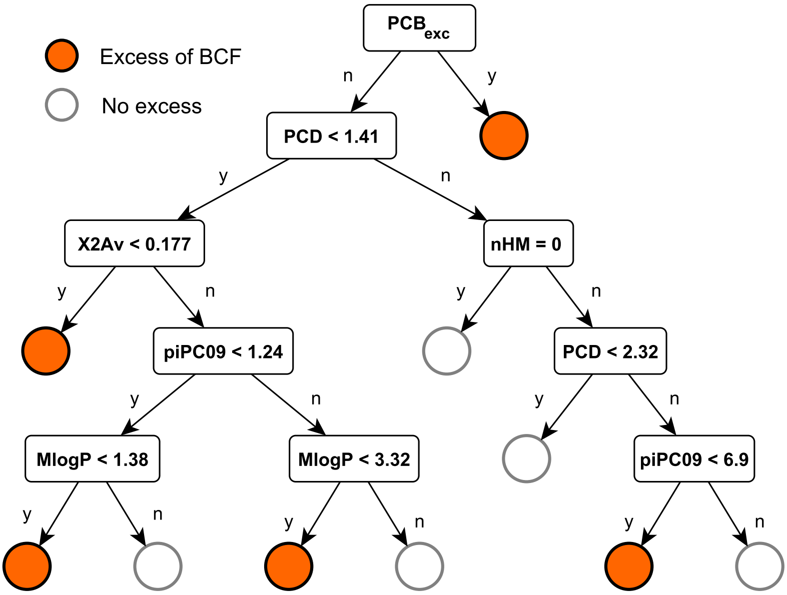


Fig. 4.3: Selected tree to discriminate class 2 compounds (orange) from other compounds (white). Square boxes denote univariate splits (i.e. nodes), while round boxes (i.e. leafs) denote the assigned class. PCB*$_{exc}$ refers to PCBs with 6 to 8 Cl atoms and $\log K_{OW} > 6$. Other nodes are labelled according to descriptor name.

PCD

PCD [30] is a path count descriptor, based on graph theory: the molecule is represented as a H-depleted molecular graph whose vertices are non-H atoms and edges are bonds, which can be characterized by paths (i.e., sequences of vertices without repetition). PCD is defined as:

$$PCD = \log\left(\frac{A + \sum_{^mP_{ij}} w_{ij}}{TPC}\right) \tag{4.5}$$

where

$$w_{ij} = \prod_{b=1}^{m} \pi_b^* \tag{4.6}$$

A is the number of vertices in the H-depleted molecular graph; $^mP_{ij}$ denotes a path of length m ($m \in [0, 10]$), from the *i*th to the *j*th vertex; w_{ij} is the path weight, calculated by multiplying the conventional bond order (π_b^*)[1] of all m edges of the path $^mP_{ij}$; TPC is the total number of paths of any length (from 0 to 10). PCD mainly relates to bond type/number and it tends to increase with increasing the number of multiple bonds (unsaturation) (Table 4.3).

X2Av

X2Av is the average valence connectivity index of order 2 [105], calculated as follows:

$$X2Av = \frac{\sum_{j=1}^{K}(\prod_{i=1}^{3} \delta_i^V)}{K} \tag{4.7}$$

where j runs over all the K 2nd order paths of the molecular graph. Each path is weighted by the product of the valence vertex degree (δ^V) of the 3 vertices involved in the path, which

[1] equal to 1, 2, 3 for single, double and triple bonds, respectively, and 1.5 for aromatic bonds

depends on the number of valence electrons (Z^v) and hydrogen atoms (h) bonded to an atom, as follows:

$$\delta^V = Z^v - h \tag{4.8}$$

X2Av accounts for the presence of heteroatoms in the molecule as well as double and triple bonds (Table 4.3). This index decreases when increasing (a) the density of adjacent triplets of vertices with many valence electrons (e.g. F – C – F); (b) the number of cycles (high K); (c) the density of unsaturated/aromatic bonds.

nHM

nHM is the number of heavy atoms (i.e., halogens, P, S, Si and Sn in this dataset).

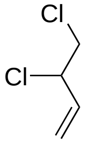
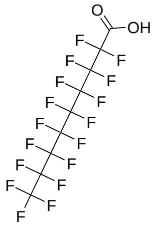
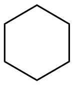
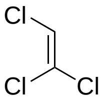
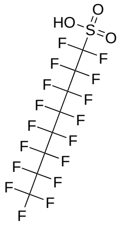
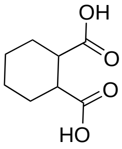
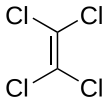
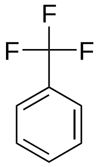
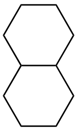
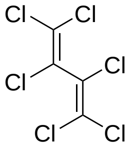
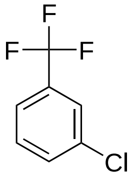
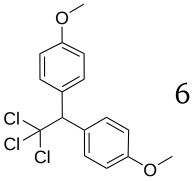
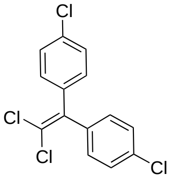
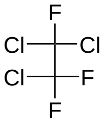
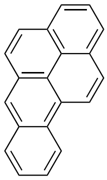
piPC09

piPC09 is the count of all the paths of length 9 in the H-depleted molecular graph. It is influenced by molecular size and tends to be higher for polycyclic aromatic molecules, which are characterized by higher bond orders and more paths than aliphatic molecules (Table 4.3).

MlogP

MlogP is the logK$_{OW}$ predicted by Moriguchi model [66], which is based on a group contribution approach.

Tab. 4.3: Examples of PCD, X2Av and piPC09 descriptor values for some molecules of the dataset, sorted in ascending order.

PCD		X2Av		piPC09	
Structure	Value	Structure	Value	Structure	Value
	0.21		0.09		0
	0.32		0.12		2.30
	0.34		0.13		2.57
	0.65		0.15		6.06
	2.15		0.23		9.32

Mechanistic Interpretation To the left branch of T2 belong molecules with a small number of multiple bonds (PCD < 1.41) and among them, those with X2Av < 0.177 can show an excess of BCF. Part of the information encoded within PCD and X2Av is overlapping and opposite, since low values of PCD correspond to few multiple bonds, while small values of X2Av to high density of multiple bonds. However, PCD is more influenced by molecular dimension than X2Av is. In this leaf, therefore, lie small molecules (i.e. low PCD), with high density of multiple bonds, in particular aromatic rings. Aromatic rings could be responsible of the excess of BCF, as they are involved in important biological intermolecular interactions, such as bonding between aromatic amino-acid side chains of proteins and (hetero)aromatic rings of small ligands [106]. In this leaf, we also find linear molecules, characterized by a high density of adjacent triplets/couples of heteroatoms and few multiple bonds. In this case, the excess could be ascribed to increased molecular flexibility (due to abundance of rotatable bonds), causing a structural rearrangement within organisms, able to maximize the interactions between heteroatoms and tissues [107]. To this leaf belong 25% of class 2 compounds, e.g. all PFAAs, some carbamates and other N/O/Cl rich compounds, meaning that the structural features encoded by PCD and X2Av are relevant for BCF excess. Among the 20 small monocyclic aromatic compounds of this leaf, only the meta-substituted anilines and 2-aminopyridine show an excess of BCF.
Concerning the left branch (PCD < 1.41 and X2Av ≥ 0.177), the underestimation by TGD is more likely for low MlogP values. This could be related to errors of the TGD equation itself or to preferential storage within organism water phases (e.g. blood) ([104], [108]).

To the right part of the tree (PCD ≥ 1.41) belong large molecules with many multiple bonds and a low number of heteroatoms.

This branch is characterized by a small number of BCF-excess compounds (19 out of 378) and this could be related with: (a) molecular rigidity due to double bonds abundance; and (b) the stabilizing effect of neighbouring carbon atoms, which could limit the structural interaction with macromolecular targets and tissues. In particular, molecules with PCD $\geq$ 1.41 and no heavy atoms or 1.41 $\leq$ PCD $\leq$ 2.32 do not show an excess of BCF; these nodes have a purity of 100% and 98.2%, respectively. The remaining molecules (PCD $\geq$ 2.32 and nHM $>$ 0) all have two or more aromatic rings and 77 out of 88 contain Cl atoms. They could show an excess of BCF if their piPC09 $<$ 6.9. All the molecules of this node (except 3) having (a) at least one O-Cl at topological distance of 4, or (b) more than 3 circuits, have piPC09 $\geq$ 6.9 and do not show an excess of BCF.

2. Less bioconcentrating chemicals (T3)

T3 (Fig. 4.4) comprises 5 nodes and 4 molecular descriptors, which are briefly described below.

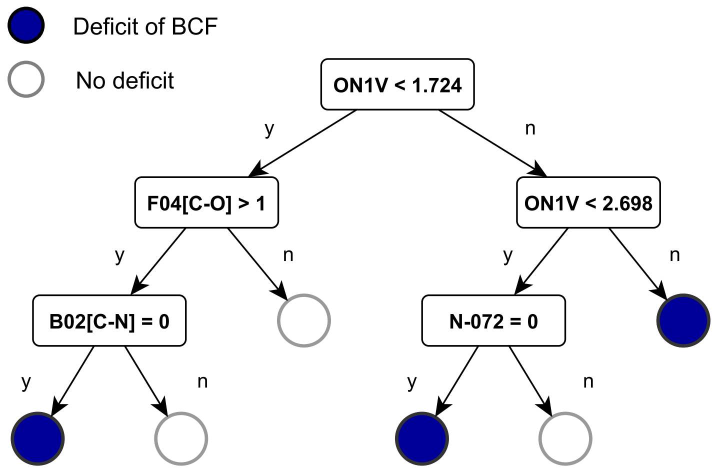


Fig. 4.4: Selected tree to discriminate class 3 compounds (blue) from other compounds (white). Square and round boxes denote univariate splits (i.e. nodes) and the assigned class (i.e. leaf), respectively. Nodes are labelled according to molecular descriptors acronyms, whose description is given in the text.

ON1V

ON1V is the overall first-order modified Zagreb index [109], defined as:

$$ON1V = \sum_{b=1}^{B} (\delta^V_{b(1)} \cdot \delta^V_{b(2)})^{-1} \tag{4.9}$$

where B is the number of bonds, is δ^V the valence vertex degree (Eq. 4.8) and $b(1)$ and $b(2)$ are the atoms connected by the bth bond. ON1V increases when increasing the number of carbon atoms (Fig. 4.5) and reflects molecular dimension, branching and presence of heteroatoms (Table 4.4).

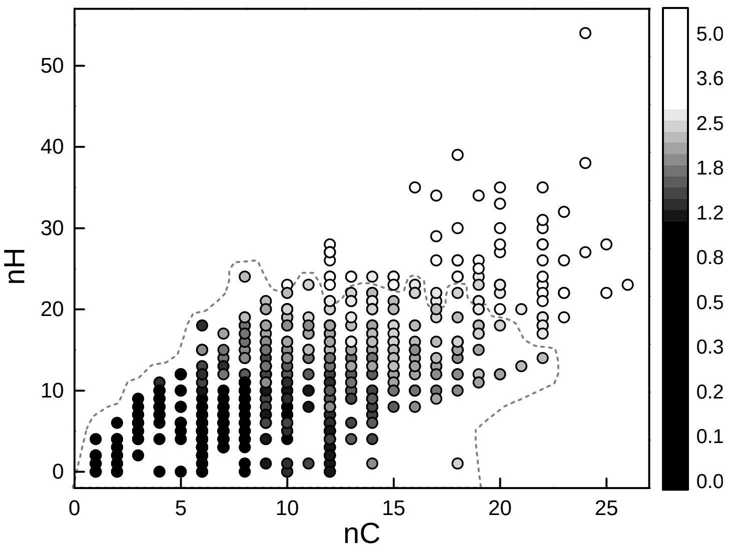


Fig. 4.5: Variation of ON1V (colour map) according to the number of carbon (nC) and hydrogen (nH) atoms. The lighter, the greater ON1V values; dashed line delimits the region of ON1V < 2.698. Note that nH is not influent in ON1V calculation (as it derives from an H-depleted graph). However, nH represents the degree of branching, aromaticity and ciclicity, which tend to decrease the total number of hydrogen atoms bonded to carbon atoms.

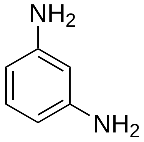
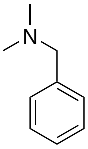
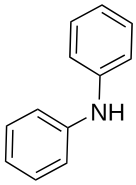
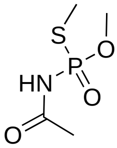
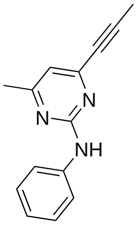
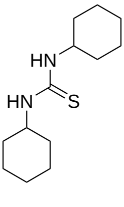
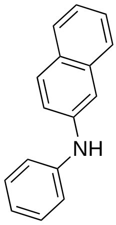
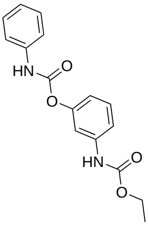
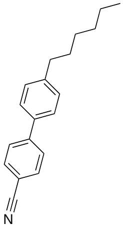
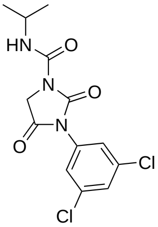
F04[C-O] and *B02[C-N]*

F04[C-O] and B02[C-N] are 2D atom pairs descriptors [110]. F04[C-O] counts the occurrences of connected C and O atoms at a topological distance of 4; B02[C-N] is equal to 1 when there is at least one pair of C and N atoms separated by 2 bonds, and 0 otherwise.

N-072

N-072 is an atom-centered fragment [111], which counts the occurrence of RCO-N<, RCS-N and >NCX=X (X being any electronegative atom) in the molecule (Table 4.4)

Tab. 4.4: Examples of ON1V and N-072 descriptor values (in ascending order).

ON1V		**N-072**	
Structure	Value	Structure	Value
	0.776		0
	1.347		1
	1.728		2
	1.750		2
	2.911		3

Mechanistic Interpetation Large and branched compounds with few heteroatoms (ON1V ≥ 2.698) may show a reduced BCF. This leaf, in fact, contains 75% of class 3 molecules (56), while only 25% of class 1 (14) and 2 (5) compounds. The effect of molecular size on bioconcentration has been being an object of debate for the last decades. In fact, while several works assert the influence of size on bioconcentration reduction ([112]–[114]), others ascribe the observed deviations to uncertain BCF data [115] and/or to a decreased bioavailability due to sorption to particles of highly hydrophobic chemicals [116]. In our case, however, 22% of class 3 compounds have ON1V ≥ 2.698 across the whole range of K_{OW}, ostensibly supporting an effect of molecular size and branching on the reduced bioconcentration. As already hypothesized [113], effective diameter controls membrane permeability and, thus, large and branched compounds could have a limited diffusion through cell membranes, resulting in BCF values smaller than expected.

Among the molecules with ON1V < 2.698, the presence of C and O at lag 4 is often associated with a deficit of BCF. The presence of oxygen atoms has already been related to increased metabolic rates in fish [117] and this is in agreement with our model. It is important to note however that, among the 310 molecules with F04[C-O] = 0, 46 belong to class 3. Among the molecules with F04[C-O] > 1, those with B02[C-N] = 0 may show a deficit of BCF. These molecules are characterized by an high abundance of aromatic oxygen atoms, phosphate esters, aliphatic ethers and aromatic/aliphatic ketones, fragments related to increased biotransformation rates [117]. Among the compounds with B02[C-N] = 1, half (56) contain at least one aromatic nitro-group and 46/56 belong to class 1, while the remaining to class 3 with small residues (< 0.90 log units). Hence, it can be stated that these structural features are related with small or no biotransformation. This is in contrast with the study of Arnot and co-authors [117]. It is important to note, however, that this structural feature regards only compounds with ON1V < 1.724 and F04[C-O] > 1.

The terminal split of the right branch (N-072), in analogy with the left side of the tree, is based on an N-related descriptor. Even in this case, molecules without N-072 fragments may be metabolized, but no shared structural characteristics are evident. All PAHs lying in this leaf (5 out of 6) are correctly classified. T3 misclassifies all PFAAs.

4.2.4 *Consensus* model

T2 and T3 trees perform well on the individual class level; however, in order to allow for model application, one has to be able to classify a given compound into one of the three classes. To this end, we tested the combination of trees in a *consensus* manner, i.e. by assigning a compound to a class if and only if both models agree on it. In case of conflict (i.e. compound predicted as both belonging to class 2 and 3), no class was assigned and the compound was discarded. This allowed also to assign compounds to class 1, i.e. when they were predicted as both non-excess and non-deficit (Fig. 4.6)

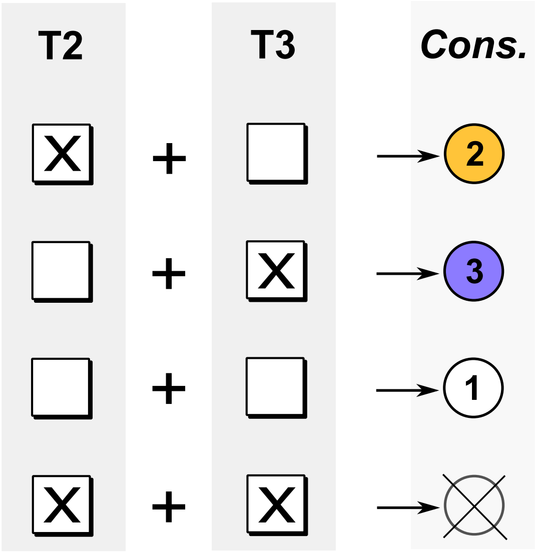


Fig. 4.6: Simplified scheme of the *consensus* model. If a compound is assigned to class 2 by T2 and not to class 3 by T3, it is predicted as belonging to class 2. If it is assigned to class 3 (T3) and not to class 2, it is assigned to class 3. If it is predicted as not belonging to class 2 nor to class 3, it is assigned to class 1. If the predictions are in disagreement, the compound is discarded.

The resulting *consensus* model (Table 4.5) shows an increased *Sp* for class 3 compounds, meaning that it identifies well the compounds external to this class. This is a prominent characteristic, as class 3 is characterized by a decreased bioaccumulation potential. On class 2 compounds, *Sp* increases with respect to T2, with a slight decrease of *Sn*. On the training set, this is mainly caused by the misclassification of PFAAs by T3 and their consequent non-prediction, while on the test set, by the rejection of two class 2 compounds (correctly predicted by T2). Despite class 1 was not modelled, the *consensus* model shows acceptable statistics, especially when considering *Sp* values.

Tab. 4.5: Statistics of the *consensus* classification for each class on training and test set: *NER*, *Sn*, *Sp* and number of non-predicted compounds (*np*) are reported.

Class	**Training**				**Test set**			
	NER	*Sn*	*Sp*	*np*	*NER*	*Sn*	*Sp*	*np*
1	0.69	0.62	0.76	20	0.62	0.57	0.66	9
2	0.87	0.90	0.84	17	0.76	0.71	0.81	2
3	0.74	0.65	0.83	28	0.66	0.58	0.74	16

4.3 Conclusions

This chapter presented a scheme to identify compounds that are: (1) mainly stored within lipids (predicted well by K_{OW}), (2) affected by additional interactions with non-lipid tissues (underestimated by K_{OW}), or (3) metabolized and/or eliminated (overestimated by K_{OW}). The scheme is based on two QSAR classification trees, whose salient features are simplicity, easy applicability, and interpretability.

Particular attention was given to the mechanistic interpretation of the molecular descriptors, which was integrated with literature-based biological and chemical knowledge. This allowed gathering new insights into the structural features connected with the mechanisms of bioconcentration.

CART was chosen for its simplicity at the expense of high model performance. Complex modelling strategies and/or complex molecular descriptors could lead to higher performances, but with a loss of interpretability and applicability. Because we aimed to investigate the mechanisms of bioconcentration, reaching high-performance classification by losing model/descriptor interpretability was beyond the scope of this work. Nonetheless, despite the simplicity of the selected trees and descriptors, the performances were adequate.

The best classification performance was shown for compounds with potential increased interactions with tissues (class 2). Structural features connected with their increased BCF are molecular flexibility and heteroatom density, or the presence of aromatic rings in small molecules. The tree for metabolized/eliminated compounds (class 3) showed a high capability of rejecting false positives. This characteristic allows for a cautionary approach, because these compounds show a reduced BCF due to metabolic clearance. Key structural features related to the deficit of BCF resulted are the molecular branching/dimension, and the presence of aromatic oxygen atoms, phosphate esters, aliphatic ethers and aromatic/aliphatic ketones. Importantly, the trees can be combined into a *consensus* classification scheme, which can serve to assess the reliability of K_{OW}-based predictions of BCF.

5 Expert system to predict BCF for regulatory purposes

> *Knowledge without application is like a book that is never read.*
>
> — **Anonymous**

Chapter 4 offered a theoretical basis to investigate the mechanisms that influence the bioconcentration. However, a prominent feature of QSAR models is their potential to save time and money and minimize the use of animal testing. Thus, the application of QSAR models beyond the boundaries of the pure research is a field of interest of both industry and regulatory agencies.

This chapter aims to apply what known on the defined classes (e.g. that class 1 compounds are predicted well by K_{OW}) to the regulatory prediction of BCF within an expert system architecture. An expert system is a formalized integration of several models based on the knowledge about their drawbacks and advantages. Expert systems have been successfully applied to many QSAR challenges, such as the prediction of carcinogenic potential [118], of estrogen receptor binding [119] and of mechanisms of toxic action [120]. In particular, the classification scheme was used as the starting basis to choose the optimal model for each class. Four BCF models - the TGD approach, VEGA CAESAR, VEGA Meylan and VEGA Read-Across - were analysed for their advantages and drawbacks on each predicted class, with the aim of developing the optimal expert system.

As the final target was the regulatory application, the OECD principles were taken into account along with other factors that influence the application and acceptance of QSAR models, namely the model simplicity, transparency and ease of use [121].

5.1 Materials and Methods

5.1.1 Datasets

In analogy with Chapter 3, each model was tested on (1) its training and test sets and (2) on the external data. The starting point was the BCF dataset of Chapter 3, consisting of 1056 compounds, which was the source of external data for each model.

5.1.2 Class prediction

Class was predicted for each compound using the *consensus* classification scheme proposed in Chapter 4. No class was assigned to compounds that: (a) were out of the AD of at least one classification tree or (b) were predicted with disagreement by the two trees (i.e. as belonging to both classes 2 and 3).

5.2 Results and Discussion

5.2.1 Approach refinement

The classification scheme of Chapter 4 was calibrated using experimental K_{OW} data. However, in order to fully quantify the model predictive ability, we wanted to test its performance when predicted K_{OW} is used. Moreover, often no experimental K_{OW} is available for new, non-tested chemicals.

The first step was to test software-based models for K_{OW} for their prediction accuracy. To this end, we chose three of the most-widely known models for K_{OW} prediction:

- KOWWIN, based on the Atom/Fragment contribution method proposed by Meylan and Howard [122];
- AlogP, based on the hydrophobicity contribution of 120 atom types as proposed by Ghose and Crippen [123];
- MlogP, which consists of a regression equation based on 13 structural parameters [66].

Amongst the available tools for K_{OW} prediction, we chose VEGA platform [124] because it includes an applicability domain assessment, which is essential for model application in regulatory contexts. In analogy with BCF models (see Paragraph 3.1.2), the reliability of the prediction is expressed by the Applicability Domain Index (ADI), which ranges from 0 (prediction unreliable) to 1 (reliable prediction). Moreover, recently, VEGA K_{OW} models resulted to be the most accurate among the several models tested [125] because of their AD assessment.

K_{OW} values were predicted for Chapter 4 dataset, comprised of manually curated experimental K_{OW} values for 779 compounds. Only compounds inside the model AD (ADI > 0.75) were considered. A *consensus* between the predicted K_{OW} was also performed, by using the arithmetic mean of the K_{OW} values of the three models (only those for which the compound was inside the AD). Predicted and experimental values were compared and the prediction accuracy was quantified using the Root Mean Squared Error (Eq. 3.1). RMSE represents the average model error and is in the same measuring units of the experimental response. Results are reported in Table 5.1.

Tab. 5.1: Performance of K_{OW} models on the training set of 779 compounds. (n = number of compounds, n_{in} = number of compounds within the AD.)

K_{OW} model	n	n_{in}	RMSE
VEGA MlogP	779	543	0.59
VEGA AlogP	779	654	0.60
VEGA KOWWIN	779	654	0.45
VEGA *Consensus*	779	730	0.55

Amongst the tested models, VEGA KOWWIN had the highest accuracy (lowest RMSE) and it was chosen as the benchmark model for K_{OW} prediction.

5.2.2 Analysis of the method

Once VEGA KOWWIN was selected as the optimal K_{OW} model, it was used to predict K_{OW} for all the available molecules. The predicted values were used as the input for the TGD equation (Box 3.1.1). Only reliable K_{OW} predictions (ADI > 0.75) were retained. The CART scheme was then used to assign each compound to a class, as explained in Section 5.1.2. The accuracy of the TGD approach was then tested on (1) the training and test sets used to develop the classification scheme (779 compounds) and (2) an external dataset comprised of the remaining 277 compounds (Table 5.2).

Tab. 5.2: Statistics of TGD model on the training and the external sets for all compounds, for each class and for non-predicted compounds (np); n = number of compounds, n_{in} = number of compounds within the AD of KOWWIN.

Set	Class	n	n_{in}	RMSE
Training	all	584	495	0.82
	1	246	230	0.52
	2	112	91	0.74
	3	161	125	1.13
	out	65	49	1.12
Test	all	195	159	0.84
	1	77	64	0.54
	2	30	27	0.71
	3	61	47	0.98
	out	27	21	1.27
External	all	277	104	1.18
	1	54	27	0.54
	2	30	11	0.97
	3	149	57	1.46
	out	44	9	0.77

For all the sets, the TGD approach is more accurate on class 1 compounds than on:(a) all compounds and (b) on class 2 and 3 compounds. Moreover, the RMSE for class 1 is comparable on all the sets, indicating that the approach is stable towards unknown molecules. The results show that: (a) the proposed approach works also when predicted K_{OW} is used instead of experimental K_{OW}; (b) the underlying hypothesis (i.e. that class 1 compounds can be predicted well by K_{OW}) is satisfied also for molecules not used to train the classification scheme.

5.2.3 Comparison with benchmark BCF models

To further test the predictive ability of the proposed scheme, the predictions were compared with VEGA BCF models of Chapter 3: (a) VEGA CAESAR, (b) VEGA Meylan, (c) VEGA Read-across. Also in this case, we chose VEGA platform because of its applicability domain assessment. Each model was evaluated on its training and test sets and on external compounds. As for Chapter 3, only predictions within the AD were considered, but in this case a more restrictive threshold was chosen, i.e. ADI > 0.75, accounting only for highly reliable predictions. Prediction accuracy was quantified through the RMSE (Table 5.3).

In most of the cases, VEGA models perform better on class 1 than on all compounds and this is coherent with what observed previously (Chapter 3). Moreover, on external compounds, the TGD approach is more accurate on class 1 than the benchmark models are on all compound. This suggests that the classification scheme could be used as a filter to choose the molecules to be predicted by K_{OW} only without the need of an increased complexity.

On a class-basis, one can note that no model always outperforms the others. Nonetheless, some models have a similar behaviour on all the datasets, keeping a constantly low RMSE: TGD on class 1 (RMSE from 0.52 to 0.54), Meylan on class 2 (RMSE from 0.43 to 0.52) and Read-Across on class 3 (from 0.46 to 0.55). In all other cases, the RMSE has a larger variance. On class 1 compounds, the TGD approach was chosen as the optimal model, since its simplicity outweighs the slightly better performance of the other models on their training/test sets. For the other classes, the model comparison was refined through a multi-criteria decision-making strategy, also taking into account the percentage of compounds inside the AD.

Tab. 5.3: Statistics of the benchmark models on each predicted class; n = number of compounds external to the model, n$_{in}$ = number of compounds within the model AD.

Model	Class	Train			Test			External		
		n	n_{in}	RMSE	n	n_{in}	RMSE	n	n_{in}	RMSE
CAESAR	all	378	334	0.43	95	86	0.63	583	162	0.79
	1	166	160	0.39	49	41	0.43	164	71	0.69
	2	64	53	0.42	10	8	0.56	101	29	1.04
	3	124	103	0.50	33	25	0.41	217	53	0.78
	out	24	18	0.46	3	1	0.70	101	9	0.78
Meylan	all	516	389	0.41	146	101	0.45	394	46	0.78
	1	222	195	0.37	73	61	0.44	88	11	0.79
	2	71	55	0.43	34	20	0.51	73	12	0.51
	3	171	104	0.46	29	17	0.46	175	14	0.94
	out	52	35	0.39	10	3	0.35	58	9	0.78
Read-Across	all	686	584	0.50	173	146	0.53	197	79	0.68
	1	253	230	0.45	62	61	0.47	65	39	0.76
	2	112	94	0.46	25	18	0.52	39	9	0.59
	3	233	195	0.53	63	51	0.55	78	28	0.46
	out	88	65	0.60	23	16	0.68	15	3	1.32

5.2.4 Expert-based system

Read-Across and Meylan models have a stable performance on class 2 and class 3, respectively. However, despite their having a low RMSE on these classes, they do not always outperform the other models in terms of RMSE and percentage of compounds within the AD. Thus, a refined model comparison was performed by applying the recently proposed wR-Hasse technique [126], a modified version of the well-established Hasse Diagrams [127].

Hasse Diagrams order the alternatives according to their variable values: When one alternative has better (or equal) values for all its criteria than another one, the two alternatives are comparable and an order can be set, otherwise the objects are incomparable and the ordering is not possible. The relationships are represented through a graph, whose vertices are the alternatives and edges are the ordering relationships. Hasse Diagrams are very powerful but often ineffective on datasets with many criteria, where a large number of incomparabilities may be observed. The recently proposed wR-Hasse overcomes this issue by setting an order when an alternative is better than another one for at least a given percentage of criteria (threshold). In this way, the number of incomparabilities can be sensibly reduced. Moreover, unlike the original version, the criteria can be weighed according to their relevance. When a weighting scheme is adopted, one object dominates over the other when the weights of the criteria for which it is better sum up to the threshold value.

The RMSE was considered as the most relevant parameter but also the percentage of compounds within the AD was taken into account (the higher, the better). Moreover, the performance on unknown data (test and external sets) was considered more relevant than that on training data. Thus, a total weight of 0.80 was assigned to the RMSE and 0.20 to the percentage of compounds within the AD. Within each group, variables referred to the training set were weighted 1/6 of the total weight, while those referred to the test and the external set were weighted 2/6 and 3/6 of the total, respectively (Table 5.4).

Tab. 5.4: Chosen weights according to variable type and considered set. (%in = percentage of compounds within the AD)

Type	Total Weight	Set	Weights
%in	0.20	Training	0.03
		Test	0.07
		External	0.10
RMSE	0.80	Training	0.13
		Test	0.27
		External	0.40

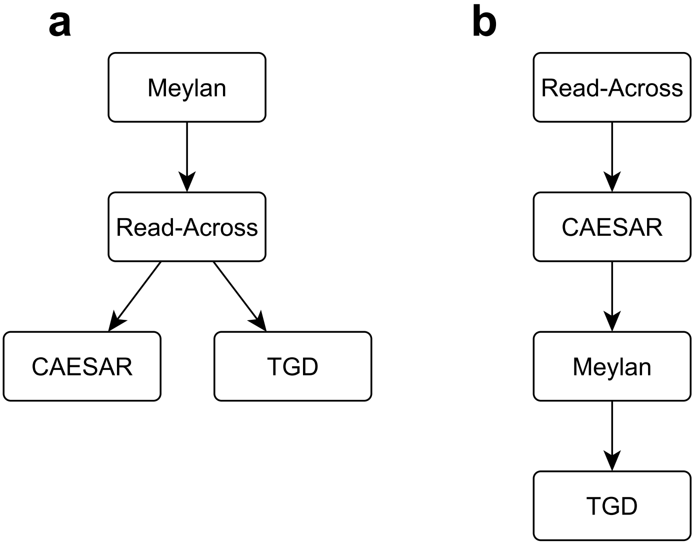


Fig. 5.1: wR-Hasse Diagrams for the selected BCF models on: (a) class 2 compounds (threshold = 0.80), (b) class 3 compounds (threshold = 0.60).

The wR-Hasse technique was applied separately on the results of class 2 and 3, using the weights of Table 5.4. The threshold was chosen on a case-by-case basis, as the largest one leading to an effective representation (Fig. 5.1). As highlighted by the resulting diagrams, the optimal models are VEGA Meylan and VEGA Read-Across for class 2 and 3, respectively. These results confirm what was already observed when considering model stability.

According to these considerations, the expert system was built as follows (Fig. 5.2): (1) TGD for class 1, (2) VEGA Meylan for class 2, (3) VEGA Read-Across for class 3. The resulting expert system had a satisfactory performance on all the datasets, with a higher accuracy towards unknown chemicals than the individual models (Table 5.5).

Tab. 5.5: Statistics of the proposed expert system on training, test and external data. Number of compounds for each set (n), number of compounds inside the AD (n_{in}) and Root Mean Squared Error (RMSE) are reported.

Set	Class	Model	n	n_{in}	RMSE
Training	1	TGD	247	230	0.52
	2	Meylan	71	55	0.43
	3	Read-Across	233	195	0.53
	all	Expert System	551	480	0.52
Test	1	TGD	77	64	0.54
	2	Meylan	34	20	0.51
	3	Read-Across	63	51	0.55
	all	Expert System	174	135	0.54
External	1	TGD	54	27	0.54
	2	Meylan	73	20	0.45
	3	Read-Across	78	36	0.51
	all	Expert System	205	83	0.51

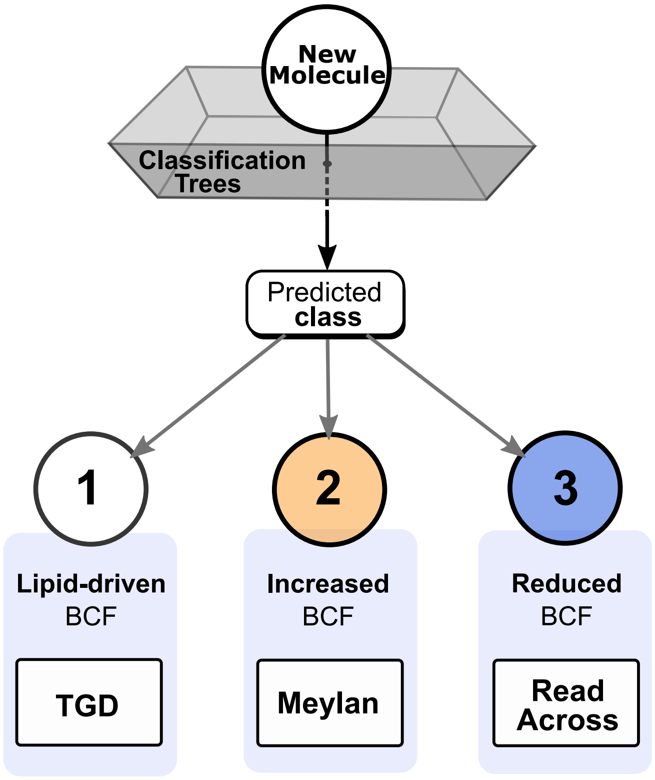


Fig. 5.2: Simplified scheme of the final expert system

5.3 Conclusions

In this chapter, the classification scheme of Chapter 4 was used for BCF prediction within an expert system setting.

The simplicity of a K_{OW}-based approach (TGD) was combined with the advantages of two more sophisticated VEGA models (Read-Across and Meylan), which are freely available and integrate the OECD principles for QSAR validity, such as the applicability domain assessment.

Our major goal was to provide a simple and transparent expert system, with a satisfactory performance on unknown data, because QSAR models are needed at most for data gap filling. The resulting expert system had an increased performance on external data with respect to the analysed models, with an improvement of RMSE up to 0.67 log units.
Since reaching high performance was secondary to keeping the model as simple and as transparent as possible, the TGD equation was chosen for compounds with lipid-driven bioconcentration, whose BCF can be reliably predicted by K_{OW} only.

The expert system can be applied within REACH framework, as (1) it models a defined endpoint through an unambiguous algorithm, (2) it is properly validated, (3) integrates an applicability domain assessment. Moreover, a mechanistic explanation is available for each class.

As the use of QSAR also depend on the model availability and transparency, the approach will be soon implemented as a KNIME [84] workflow.

Part III

Dietary Bioaccumulation

6 Biomagnification factor: critical comparison with K_{OW} and BCF

> *It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.*
>
> — **Sir Arthur Conan Doyle**
> Writer

The BCF is the criterion requested by regulatory agencies for bioaccumulation assessment. Despite this, it is known that it can be sometimes a poor descriptor for biomagnification within food webs, as it does not include dietary exposure [20]. A study of Arnot and Gobas compared fish BCF and BAF data, reporting that, in some cases, the BAF can be more than 1 order of magnitude greater than BCF due to storage through diet [20]. Thus, the authors recommend to consider other diet-including parameters, such as BAF, BMF or TMF, especially for regulatory purposes.

Despite these evidences, only few analyses can be found in the scientific literature that compare the BCF with other aquatic bioaccumulation criteria and this is ostensibly due to data availability.

In light of these considerations, this chapter aims to critically understand in which cases the dietary exposure information is relevant in addition to the BCF to fully account for the bioaccumulation profile of chemicals.
We chose the Biomagnification Factor (BMF) as the measure of dietary bioaccumulation, which can be determined in a laboratory setting, thus being less affected by environmental variability and less expensive to determine than other criteria such as TMF or BAF. Moreover, the BMF is the complementary value of the BCF (only dietary exposure *vs* only non-dietary exposure) and it allows to analyse the two bioaccumulation routes separately.

A dataset of laboratory fish BMF values for 214 organic chemicals was collected and manually curated. BMF values were compared with K_{OW} and BCF for: (a) their bioaccumulation assessment on the basis of screening thresholds, and (b) their agreement in quantifying the bioaccumulation behaviour of chemicals. Finally, relevant classes of chemicals were analysed, highlighting interesting and different patterns of bioaccumulation on a case-by-case basis, which can be reconducted to different biological mechanisms.

6.1 BMF Dataset

BMF database Laboratory BMF values were retrieved from the database recently developed by Arnot and Quinn [128]. The database is comprised of 3,032 aquatic endpoint values for 19 fish species and 477 chemicals. Literature data were screened for their quality and their compliance with OECD testing guidelines [129]. Each value is provided with a reliability score, i.e. H (High) > M (Medium) > L (Low), assigned on the basis of several factors, such as diet type, feeding rates, exposure duration, experimental design, growth correction and analysed tissues. Records are univocally identified with name, CAS and SMILES.

Structure curation Since SMILES are fundamental for descriptor calculation and model development, particular attention was given to their curation, despite they were provided by the authors. Thus, for each compound, SMILES were obtained from CAS and name using KNIME CIR node [73]. All the records leading to different structures from the used identifier were removed. If the SMILES was not retrieved by name, CAS, or both, they were manually searched from benchmark online databases: ChemSpider [75], Scifinder [130] and PubChem [74]. Eventually, the SMILES obtained by CAS and name were compared with those provided by the authors. Further mismatches (e.g. due to specified/unspecified stereochemistry) were manually resolved, if possible. Records with conflicting structures were removed.

Experimental BMF curation Once the final set of CAS, name and SMILES was obtained, multiple BMF values were merged according to CAS number. Single-valued compounds with a lowly reliable (L) BMF were removed. For the other compounds, an outlier analysis was performed by checking for anomalous standard deviations. In particular, the pooled standard deviation was calculated using H and M records (equal to 0.67) to be regarded as the reference, since it was comparable to that obtained by H records only (0.53). If for a given compound the standard deviation was above the calculated threshold and contained L records (29 compounds), they were removed and the standard deviation was recalculated. If again the values were above the threshold (6 compounds), also M records were deleted. If the standard deviation was above again, the chemical was excluded from the preliminary analysis phase. Record with only L values and a standard deviation larger than the threshold were also removed. For each compound, the median BMF value was used as reference value, as it is a robust estimator. The final dataset contained 214 compounds.

6.2 Results and Discussion

For the BMF dataset compounds, K_{OW} and BCF data were obtained from the datasets of Chapters 3 and 4. Data gaps were filled using: (1) the expert system of Chapter 5 for BCF and (2) VEGA KOWWIN for K_{OW}, for the reasons stated previously. Compounds out of the AD of the BCF model or of KOWWIN model were not considered, obtaining a total of 168 compounds with known K_{OW}, BCF and BMF values (Fig. 6.1).

As already noticed, K_{OW} can often overestimate or underestimate the BCF, because of biotransformation and non-lipid storage. The same could happen if K_{OW} is used to assess the dietary bioaccumulation. Finally, if for a given compound the main bioaccumulation route is through diet, also the BCF will lead to the underestimation of the real bioaccumulation process. That being said, the criteria were firstly compared for their coherence in detecting bioaccumulating compounds and then for their correlation. Finally, they were compared on specific chemical classes of compounds, to test whether relevant trends can be seen.

6.2.1 Classification of bioaccumulative compounds

This paragraph aims to test whether the bioaccumulation criteria are coherent in identifying bioaccumulative compounds. For the sake of simplicity, the REACH criteria on BCF and K_{OW} were

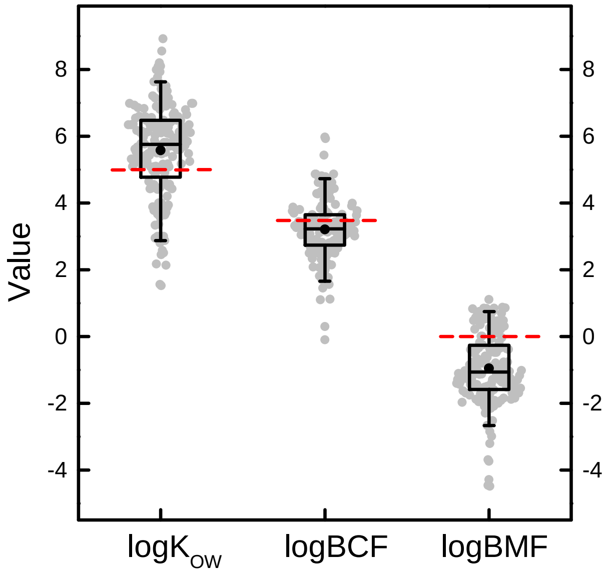


Fig. 6.1: Distribution of K_{OW}, BCF and BMF values. Boxplots show median, 1st and 3rd quartiles (solid lines), mean (black dots) and 5th - 95th percentiles (whiskers). Dashed red lines represent the thresholds generally used for bioaccumulation screening ($\log K_{OW}$ = 5, logBCF = 3.3, logBMF = 0.)

used, namely logBCF $\geq$ 3.3 and $\log K_{OW} \geq 5$. For BMF, the generally accepted threshold is BMF $\geq$ 1 (logBMF $\geq$ 0) [128]. The mentioned thresholds were applied to the dataset, highlighting that the number of compounds identified as bioaccumulative varies according to the criterion looked at, being $n_{KOW} > n_{BCF} > n_{BMF}$ (Table 6.1).

Moreover, a K_{OW}-based assessment identifies more false negatives for compounds that bioconcentrate (11) than for those that accumulate through the diet (Table 6.2). The opposite occurs for false positives, which are larger for compounds bioaccumulating through the diet.

The BCF identifies 41 compounds that are potentially bioaccu-

Tab. 6.1: Classification matrix of the bioaccumulation assessment according to the criteria. *Thr* = threshold used, n = number of compounds, % = percentage of compounds. Compounds are considered as Bioaccumulative if $\log K_{OW} \geq 5$, $\log BCF \geq 3.3$ or $\log BMF \geq 0$.

Criterion	Thr	non Bioaccumulative		Bioaccumulative	
		n	%	n	%
$\log K_{OW}$	5	46	27	122	73
logBCF	3.3	92	55	76	45
logBMF	0	131	78	37	22

mulative without accumulating through the diet (Table 6.2), while the opposite occurs only for 2 compounds.

Tab. 6.2: Contingency tables of the bioaccumulation assessment according to different criteria (K_{OW} and BCF). The thresholds used for bioaccumulation screening are $\log K_{OW} = 5$, $\log BCF = 3.3$, $\log BMF = 0$. N = negative (non-Bioaccumulative), P = positive (Bioaccumulative).

$\log K_{OW}$	logBCF		logBMF		logBCF	logBMF	
	N	P	N	P		N	P
N	35	11	45	1	N	90	2
P	57	65	86	36	P	41	35

In other words, this first analysis shows that the K_{OW}-based assessment of bioaccumulation has more pitfalls for compounds above the bioconcentration threshold than for those that accumulate through diet. At the same time, the bioconcentration factor seems a good criterion for detecting compounds that bioaccumulate.

These considerations are founded on a threshold-based approach only, which has several limitations: (1) it gives only partial information about the bioaccumulation process, because it neglects the continuous values of the criteria; (2) small variations of the chosen threshold can lead to different outcomes, especially for borderline compounds. For these reasons, the next paragraph focuses on comparing the continuous values of the criteria.

6.2.2 Continuous values comparison

This paragraph compares the experimental BMF values with the BCF and K_{OW} values (Fig. 6.2).

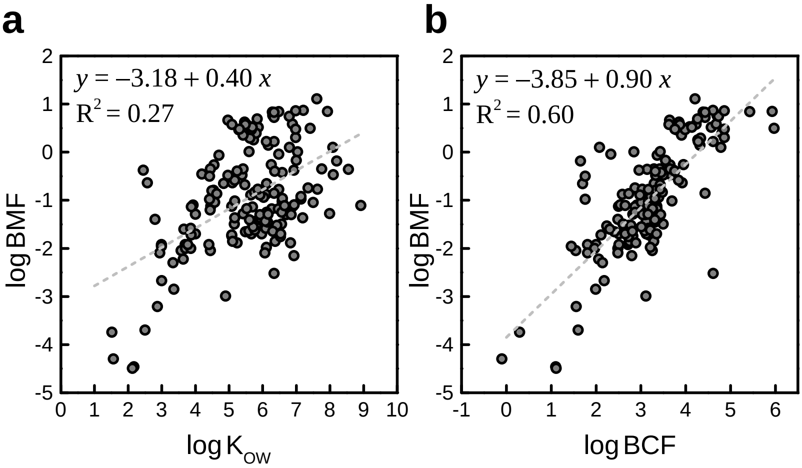


Fig. 6.2: Comparison between BMF data and: (a) K_{OW}, (b) BCF. Dashed lines represent the ordinary least squares model, whose equation and parameters are reported in the graph.

Despite K_{OW} and BMF are coherent in the threshold-based bioaccumulation assessment, they are not much correlated (Fig. 6.2a). Therefore, the estimation of the BMF based on K_{OW} would lead to underestimations/overestimations of several orders of magnitude.

The fit improves when the BCF is used (Fig. 6.2b), but also in this case, some compounds would be largely under- or overestimated in a regression-based approach. These observations suggest that the bioaccumulation through diet can be uncoupled from the bioconcentration process, despite their being reasonably correlated. For this reason, in the next paragraph, the bioaccumulation criteria were compared on specific chemical classes.

6.2.3 Cluster Analysis

The previous sections showed that, in some cases, K_{OW} and BCF are not able to fully account for the bioaccumulation processes and that some compounds may have a BMF larger than the estimates. This paragraph analyses the bioaccumulation of specific classes of chemicals in a multi-variate manner. In particular, a Self-Organizing Map (SOM) [131] was trained using K_{OW}, BCF and BMF as the input variables.

SOMs mimic the action of a neural network of neurons, where each neuron accepts different signals from neighbouring neurons and processes them in a self-organising manner. The SOM map is a squared space, consisting of a grid of N^2 neurons. Each neuron contains p elements (weights), p being the number of variables. The weights of each neuron are randomly initialised between 0 and 1 and updated on the basis of the input samples. Each weight represents the contribute of the variable in determining the value of that neuron. In each training step, samples are projected on the network, one at a time, and assigned to the most similar neuron on the basis of the Euclidean distance. Then, the weights of each neuron are updated taking into consideration the values of the introduced sample. This is reiterated for all objects and all the training epochs. In this way, objects will reorganize within the map and cluster according to their similarity. The data structure can be easily visualised and interpreted and the role of the experimental variables can be elucidated through the Kohonen weights.

A non-toroidal SOM was trained for 20 epochs on the 168 compounds described by the 3 parameters, with a size of 12 × 12 neurons to potentially allow each compound to occupy one neuron (Fig. 6.3).

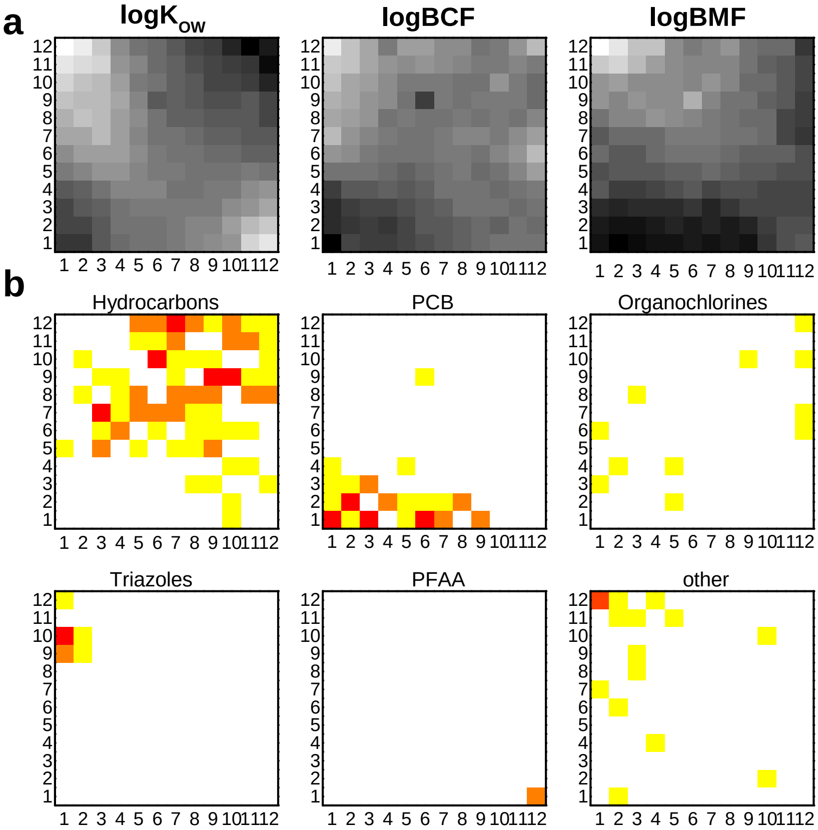


Fig. 6.3: Self Organizing Map using K_{OW}, BCF and BMF as descriptors: (a) neuron weights for the variables, from 0 (white) to 1 (black); (b) distribution of the compounds on the map; the most represented classes of compounds of the dataset were plotted separately. Color map represents the number of compounds in each neuron (yellow = one, orange = two, red = three).

Data structure can be understood through the weights (Fig. 6.3a), which can be directly compared with the object positions in the map (Fig. 6.3b), as follows:

- The *bottom-left* side of the map is characterized by high values of all the variables. Compounds that lie here bioaccumulate both trough dietary (high BMF) and non-dietary routes (high BCF), mainly because of lipid storage (high K_{OW}). In these neurons lie the PCBs. They distribute in the region of highest BMF values, even those with relatively low BCF.

- The *top-left* side contains compounds with low K_{OW}, BCF and BMF values. These compounds are not lipophilic and not bioaccumulative. In this part of the map lie all the triazoles, which are rapidly metabolised in fish [132]. Here also lie some herbicides (napropamide, diflufenican) and one synthetic musk fragrance (musk xylene).

- To the *top-right* neurons belong compounds with high K_{OW} values but smaller BCF and BMF values. In analogy with Chapter 4, these compounds are probably metabolised, resulting in a faster elimination (or in a bias in the observed concentration). The BMF values are larger than BCF, suggesting that the dietary bioaccumulation could be more relevant in this case. To these neurons belong the majority of the hydrocarbons, especially PAHs. This is coherent with the observations of Chapter 4.

- To the *bottom-right* part of the map belong compounds that have BCF and BMF larger than K_{OW}. Perfluorinated Alkyl Acids (PFAA) lie in this region, meaning that the hypothesized non-lipid storage could be relevant also for dietary routes. Note that the weight of BMF is higher than that of BCF in this region.

- The *central* part of the SOM hosts compounds with intermediate behaviour, which have similar (and moderate) values of K_{OW}, BCF and BMF.

6.3 Conclusions

This chapter compared fish BMF with K_{OW} and BCF for their identification of bioaccumulative compounds. A general agreement was found in the threshold-based screening.

However, the correlation of the continuous K_{OW}, BCF and BMF values resulted sub-optimal. In some cases, in fact, the prediction of the BMF based on K_{OW}/BCF would lead to underestimation or overestimations of several orders of magnitude of the real dietary bioaccumulation and, as a consequence, of the actual biomagnification potential. Given these considerations, specific classes of compounds were analysed with the aid of a self-organizing map. Results highlighted that for some classes of compounds, e.g. PFAAs and PCBs, the dietary bioaccumulation could be even more relevant than the bioconcentration.

In conclusion, this chapter highlighted the relevance of considering also the BMF in addition to the BCF in order to account for the potential to biomagnify in the food web. Unlike the BCF, however, not many BMF data are available and, to the best of our knowledge, no QSAR model developed on heterogeneous chemicals currently exists. To this end, we set out to model the BMF in a QSAR setting, the model to be used as a side tool for the bioaccumulation assessment. Procedure, results and mechanistic insights are fully described in the next chapter.

7 Modelling the dietary bioaccumulation

> *The purpose of models is not to fit the data but to sharpen the questions.*
>
> — **Samuel Karlin**
> Mathematician

The previous chapter showed that, for some chemicals, the dietary bioaccumulation could be more relevant than the bioconcentration. Thus, the BMF can be useful to detect hazardous compounds. To our knowledge, no global models to predict the BMF exist and only a few target specific chemical classes (e.g. [50]). Therefore, the dataset of Chapter 6 was used to calibrate a QSAR model in compliance with the OECD principles for QSAR validation, to allow for its regulatory application.

After a screening phase to find the optimal regression and variable selection techniques, the most promising settings were used to obtain two QSAR models, one local and one linear. The models were combined in a *consensus* manner, increasing the prediction performance. Particular attention was posed to AD assessment, model validation and mechanistic interpretation. The models revealed that, while some of the structural features underlying the dietary bioaccumulation are shared with the bioconcentration process, others are different; this explains the presence of preferential bioaccumulation routes.

This chapter firstly describes the optimization phase, and then gives in depth details about the models, with a special focus on the explanation and the mechanistic interpretation of the selected molecular descriptors.

7.1 Materials and methods

7.1.1 Dataset

The BMF dataset of Chapter 6 was used, consisting of 214 organic compounds. Since some compounds had spikes in their response values (Fig. 7.1a), BMF was log-transformed to have an optimal value distribution (Fig. 7.1b).

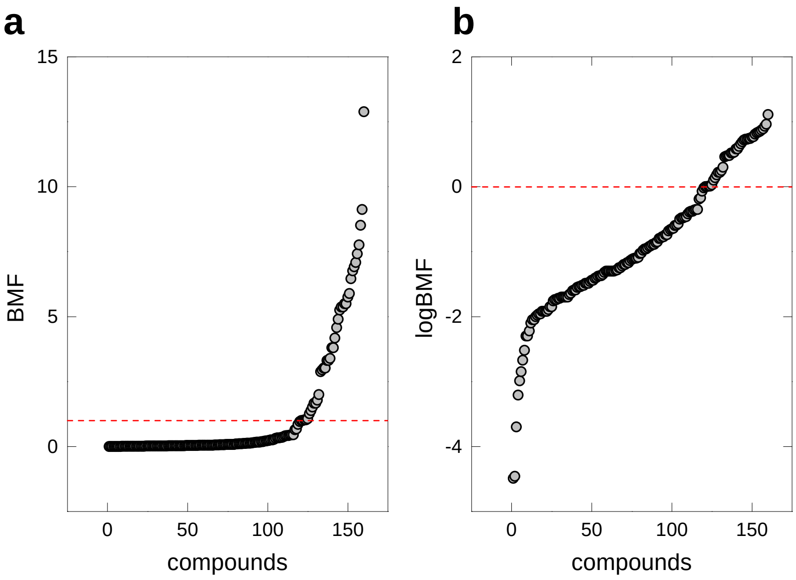


Fig. 7.1: Distribution of experimental values: (a) BMF, (b) logBMF. Dashed lines represent the threshold for compounds that bioaccumulate through diet (BMF = 1, logBMF = 0).

7.1.2 Molecular Descriptors

Dragon 6 [30] molecular descriptors were calculated from 0- to 2D and reduced by excluding those: (a) with at least one missing value; (b) constant or near-constant; (c) with a standard deviation less than 0.001; (d) with a pairwise correlation larger or equal to 0.95 with other descriptors. The final number of descriptors was 788, divided in 18 logical groups, based on the type of encoded chemical information (e.g. connectivity, presence of functional groups, molecular properties).

7.1.3 Variable selection

Each variable selection method is, in general, characterized by a peculiar capability to explore/exploit the variable space. Moreover, on a given set of variables, different regression methods can lead to highly different prediction outcomes.

On these bases, two variable selection strategies were chosen to identify the most relevant molecular descriptors, namely the Genetic Algorithms (GA) [32] and the Reshaped Sequential Replacement (RSR) [36]. In a recent study [39], these methods resulted to have a very different exploration ability, sharing, at the same time, the capability of finding highly relevant subsets of variables. They were combined with several regression methods.

7.2 Results and Discussion

7.2.1 Modelling strategy

Model evaluation and validation

Compounds were randomly split into a training set of 160 compounds (75%) and a test set of 54 compounds (25%).

The training set was used to choose the best regression approach, select the variables and obtain the final models. As measure of model predictivity, the Q^2 in cross-validation was used:

$$Q^2_{cv} = 1 - \frac{\sum_{i=1}^{n_{tr}} (y_i - \hat{y}_{i/i})^2}{\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2} \tag{7.1}$$

where n_{tr} is the number of training compounds, $\hat{y}_{i/i}$ is the value of the *i*th object predicted by the model in which the *i*th object was not taken into consideration, while y_i is the real response; $\bar{y}_{tr}$ is the arithmetic mean of the response (over all the training data). A 5 fold Venetian blind resampling technique was used to obtain comparable and consistent Q^2_{cv} for the different calculated models. The Q^2_{cv} differs from the more known R^2, which only quantifies the ability of the model to describe the training data and is defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7.2}$$

where $\hat{y}_i$ is the calculated response of the *i*th object when it is used for model calibration and the other parameters are those of Eq. 7.1.

The test set was used only to validate the final pool of models and not used for model selection/calibration [133]. Compounds of the test set were predicted using the final models and the predictivity was quantified through the Q$^2_{ext}$ parameter proposed by Consonni *et al.* [134]:

$$Q^2_{ext} = 1 - \frac{\left(\sum_{j=1}^{n_{ts}} (y_j - \hat{y}_j)^2\right) / n_{ts}}{\left(\sum_{i=1}^{n_{tr}} (y_i - \bar{y}_{tr})^2\right) / n_{tr}} \tag{7.3}$$

where n_{ts} is the number of objects of the test set; y_j and $\hat{y}_j$ are the real and the predicted response of the *j*th test compound, respectively; n_{tr} and n_{ts} are the number of training and test compounds, respectively.

R$^2_{cv}$, Q$^2_{cv}$ and Q$^2_{ext}$ range from 0 (poor performance) to 1 (best performance). When these values are high and similar, they indicate model stability and predictivity towards new compounds.

Strategy optimization

In order to identify the best strategy for model development, two variable selection methods (GA and RSR) were combined with 6 regression techniques, as shown in Table 7.1.

Tab. 7.1: Regression methods and variable selection techniques used in the exploratory phase.

Regression method	ID	GA	RSR
Ordinary Least Squares	OLS	●	●
Principal Component Regression	PCR		●
Partial Least Squares	PLS	●	
Least Abs. Shrinkage and Sel. Operator	LASSO	●	
Nearest Neighbour Regression	NNR	●	●
Binned Nearest Neighbour Regression	BNNR	●	

The tested regression methods were chosen on the basis of two criteria:

- *Simplicity and interpretability*. Since one of the major goals was to understand the influence of molecular descriptors on the dietary bioaccumulation, only easily interpretable regression approaches were considered. For this reason, high-performing but very complex strategies such as support vector regression [135] were not considered.

- *Implementation within the variable selection algorithm*. GA and RSR have a very different behaviour, the former being very fast, while the latter being more exploitative but slower. For this reason, some regression techniques are more suitable for one method than they are for the other or *vice versa*. Because of the preliminary character of this phase, only the methods already implemented were considered, with the exception of LASSO and BNNR, which for this study were firstly implemented within the GA architecture. Because they did not outperform existing methods, they were not implemented also within RSR strategy.

Regression techniques Some details about the tested regression techniques can be found below:

- *Ordinary Least Squares* (OLS). The Ordinary Least Squares approach is the most intuitive and well-known regression technique [136]. Given a set of p variables, the response ($\hat{y}$) is predicted as follows:

$$\hat{y}_i = b_0 + \sum_{j=1}^{p} b_j \cdot x_{ij} \tag{7.4}$$

  where b_0 is the intercept, b_j is the regression coefficient for the *j*th variable and x_{ij} is the value of the *j*th variable for the *i*th object. The coefficients are determined by minimizing the residuals sum of squares: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

- *Principal Component Regression* (PCR) shares a similar logic of OLS [136], but the regression inputs (Eq. 7.4) are obtained by linearly combining the original *p* variables into *M* new orthogonal variables explaining the largest data variance (Principal Components).

- *Partial Least Squares* (PLS). PLS technique also constructs a set of linear combinations of the variables (latent variables), but unlike PCR, it uses **y** in addition to **X** for their construction [137].

- *Least Absolute Shrinkage and Selection Operator* (LASSO). LASSO [138] shrinks the coefficient of OLS (Eq. 7.4) through a parameter λ, as follows:

$$\hat{y}_i = b_0 + \sum_{j=1}^{p} (b_j + \lambda |b_j|) \cdot x_{ij} \tag{7.5}$$

When $\lambda = 0$, the model is an OLS model, while $\lambda = 1$ corresponds to the maximum shrinkage.

- *Nearest Neighbour Regression* (NNR) is a local approach that uses the *k* most similar compounds (nearest neighbours) to predict the response of unknown objects [139] as the arithmetic mean of the neighbours' response or the weighted mean (using the similarity as the weighting factor). In this study, the dissimilarity was quantified through the Manhattan distance, which resulted to have the best average performance among 110 metrics compared [140] and can be expressed as follows:

  $$D_{st} = \sum_{j=1}^{p} |s_j - t_j| \tag{7.6}$$

  where *s* and *t* are the two objects under analysis, j runs over the *p* variables and s_j (t_j) is the value of the *j*th variable for s (t). Both the unweighted and the weighted versions were tested.

- *Binned-Nearest Neighbour Regression* (BNNR) derives from the classifier recently proposed by Todeschini *et al.* [141]. BNNR takes inspiration from NNR, but the number of neighbours varies for each object, according to predefined similarity intervals (bins), optimized through a parameter α. All the neighbours falling into the largest similarity bin of a new object are considered for the prediction of its response as the arithmetic mean. Also in this case, the Manhattan distance was used.

All the regression parameters (*k*, α, λ, number of latent variables and of components) were optimized in cross-validation.

Variable selection GA were run 100 times with 100 steps of evolution each, as proposed by Leardi *et al.* [142]. The most frequently selected variables (up to a maximum of 15) were used for model calibration. For RSR, models from 2 to 15 variables were generated. The best models for each combination between VS method and regression technique were chosen as the optimal compromise between model complexity (i.e. number of included variables) and Q^2_{cv}. Results were then compared to identify the most promising methods for the endpoint of interest.

Preliminary Results The most promising strategies resulted to be GA coupled with NNR and RSR in combination with linear models (OLS, PCR). OLS was preferred to PCR for its simplicity, given their comparable performance. Because of the complementary nature of the best performing variable selection/regression methods (GA *vs* RSR and local *vs* linear), they were both used in the phase of model development.

7.2.2 Model Development

Variable selection The optimal combinations between variable selection and regression methods obtained in the previous phase were used for model development, as follows:

- Genetic Algorithms were run in combination with NNR in a two-step procedure: (1) on each logical block of descriptors, (2) on the best descriptors of each block (up to a maximum of 150). Finally, an All Subset Selection was performed on the best descriptors of phase 2 (15).

- RSR was combined with OLS; a population of 3 models for each dimension (from 2 to 10 variables) was generated.

RSR was run (1) on each logical block of descriptors, (2) on the best descriptors of phase 1.

Model Selection The models were first screened according to a set of criteria: (1) Q^2_{cv} (the higher, the better), and (2) number of variables (the lower, the better). Selected models were then validated on the test set and chosen according to (1) Q^2_{ext} (the more similar to Q^2_{cv}, the better), and (2) descriptor interpretability (the higher, the better). Unlike Chapter 4, only a few satisfactory models included simple descriptors; thus, attention was given to descriptor interpretation. The selected models are reported in Table 7.2. In addition, the arithmetic mean of the predictions of the two models was calculated. The obtained *consensus* model had better performance than the sub-models alone (Table 7.2). The models will be discussed in depth in the next paragraphs.

Tab. 7.2: Performance of the selected models in fitting, cross-validation and on the test set. VS = Variable Selection, Reg = Regression Technique, p = number of variables.

ID	VS	Reg	p	FIT		CV		TEST	
				RMSE	R^2	RMSE	Q^2	RMSE	Q^2
M1	GA	NNR	4	0.52	0.77	0.55	0.74	0.54	0.75
M2	RSR	OLS	7	0.53	0.76	0.55	0.74	0.57	0.72
Consensus				0.47	0.80	0.50	0.78	0.46	0.82

7.2.3 Model Analysis

1. Model M1

M1 is a NNR model comprised of 4 variables. The optimal k value, optimized in cross-validation, was 9. For each compound, the 9 most similar objects were used to predict the response in a weighted manner, according to their distance, as follows:

$$\hat{y}_s = \sum_{t=1}^{k} w_t \cdot y_t \tag{7.7}$$

$$w_t = \frac{\sum_{t=1}^{k} D_{st}}{0.1 + D_{st}} \tag{7.8}$$

where D_{st} is the Manhattan distance between the compounds *s* and *t* (Eq. 7.6), and y_t and w_t are the experimental response and the weighting parameter of the *t*-th neighbour, respectively; $\hat{y}_s$ is the predicted response of *s*.

The model was analyzed by a multi-dimensional scaling, which projects the distances between chemicals on a bi-dimensional plane [143]. Compounds with similar BMF values lie close to each other in the descriptors space (Fig. 7.2). This explains why a local approach gives good results.

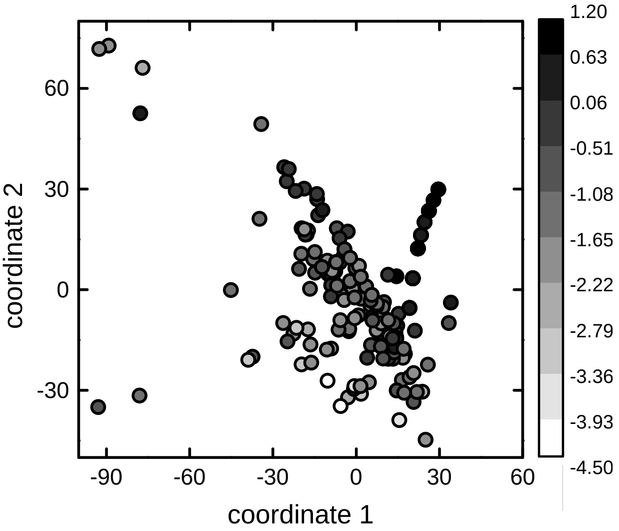


Fig. 7.2: Multi-Dimensional scaling on M1 descriptors, using the Manhattan distance. Compounds are coloured according to logBMF.

M1 has all interpretable and simple descriptors, namely: MlogP2, nBT, F06[C-C], B02[N-O], which are briefly described below.

MlogP2

MlogP2 is the squared MlogP descriptor (Section 4.2.3), which is the logK_{OW} predicted by the Moriguchi model [66].

nBT

nBT is a constitutional descriptor that quantifies the total number of bonds. It is mainly related to molecular dimension (the higher the dimension, the higher nBT), but also to molecular cyclicity and to the presence of multiple bond/heteroatms (Fig. 7.3).

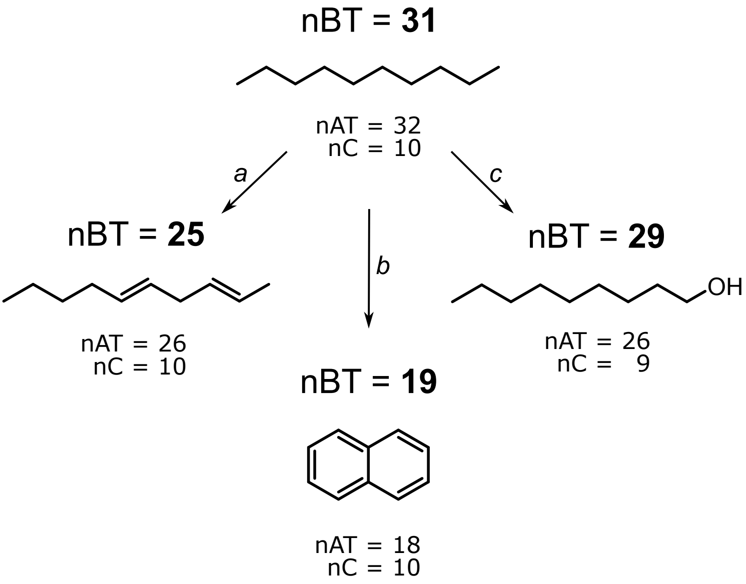


Fig. 7.3: Factors that influence nBT descriptor, given a molecule with 10 non-H atoms: (a) presence of multiple bonds, (b) molecular cyclicity, (c) presence of heteroatoms. Total number of atoms (nAT) and of carbon atoms (nC) are also reported.

B02[N-O]

B02[N-O] [110] is equal to 1 if there is at least one pair of N and O atoms separated by 2 bonds.

F06[C-C]

F06[C-C] is a 2D atom pairs descriptor [110], which counts the occurrence of C atoms separated by 6 bonds. It depends on molecular dimension, branching and presence of heteroatoms (Fig. 7.4).

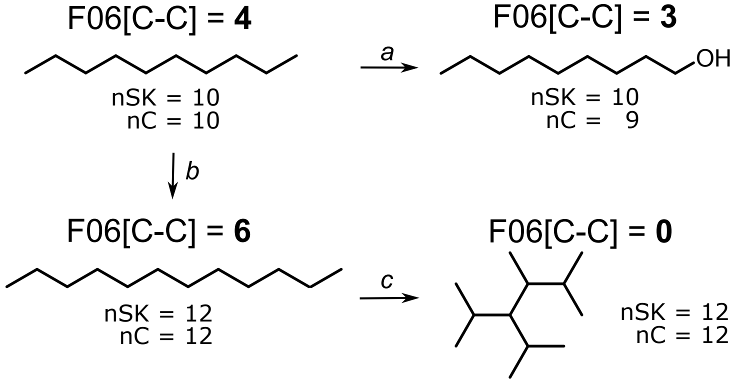


Fig. 7.4: Effect on F06[C-C] descriptor of: (a) presence of heteroatoms, (b) molecular dimension, (c) molecular branching. Total number of non-H atoms (nSK) and of carbon atoms (nC) are also reported.

Mechanistic Interpretation An easy model interpretation was allowed by a Principal Component Analysis (PCA [144]), a multivariate technique for data visualization and dimensionality reduction. PCA linearly combines the variables into new orthogonal ones (Principal Components, PCs), such that the first PC explains the largest data variance, the second one (orthogonal to the first) the second largest variance and so on. In this way, one can observe and interpret the relationship among the variables, the objects and the PCs.

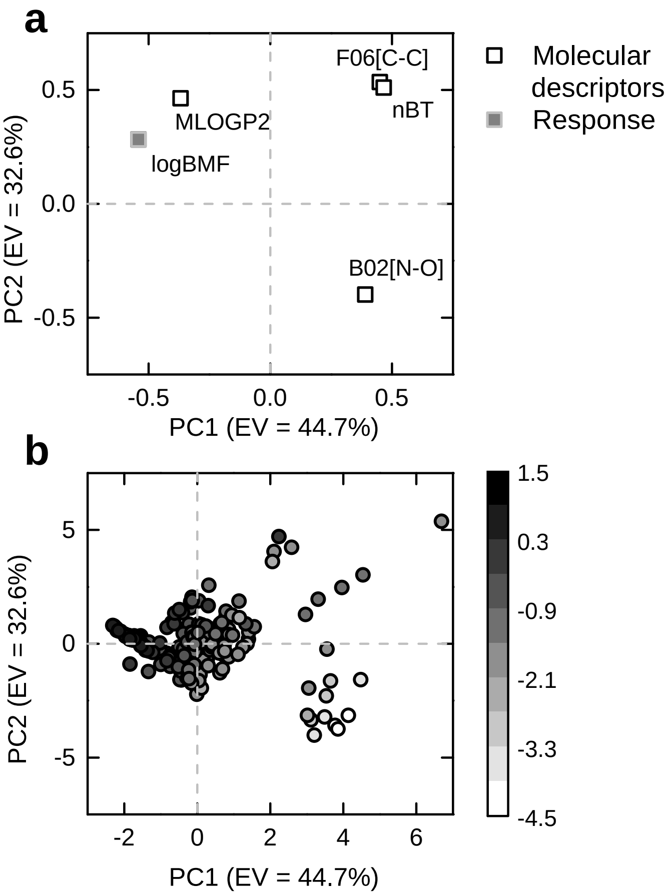


Fig. 7.5: PCA on M1 descriptors + experimental response: (a) loading plot, (b) score plot. Compounds are coloured according to logBMF (from white to black).

To perform the PCA, the compounds were described by the 4 model variables plus the logBMF, which was added to stretch the results along the direction of experimental response variation. The 80% of the variance was explained by the first two PCs, which were thus analysed in depth. The other PCs contain only marginal information, as they explain less than 13% of variance each.

The contribution of each variable to each PC is quantified by the loadings (Fig. 7.5a), the higher (in absolute value), the greater the contribution. In particular:

- MlogP2 is relevant for both the PCs. It increases from the bottom-right to the top-left part of the PC1-PC2 space and it is the variable mostly correlated with logBMF.

- F06[C-C] and nBT are high for high values of PC1 and PC2, i.e. they increase from the bottom-left to the top-right part. This corresponds to an increase of branching/cyclicity, number of heteroatoms and of multiple bonds, while molecular size dimension increases in the opposite direction;

- Non-null B02[N-O] correspond to positive values of PC1 and negative values of PC2. Thus, compounds with B02[N-O] will lie in this portion of the PC space.

The logBMF decreases from the top left to the bottom right part of the score plot, in the same direction of MlogP2 variation. As for the bioconcentration, the dietary bioaccumulation is directly proportional to the lipophilicity, since affinity with lipids is responsible for uptake/storage within lipid tissues. However, unlike the bioconcentration, the logBMF is proportional to the squared K_{OW} value. This means that the compounds with very small negative MlogP and those with very large positive MlogP will have similar BMF values. In other words, also compounds that are very hydrophilic can have high BMF values and this can be ascribed to a preferential storage within organism water phases, such as blood ([104], [108]).

The variation of nBT and F06[C-C] (Fig. 7.5b) indicates that the logBMF tends to be higher for small molecules, with a high density of heteroatoms/multiple bonds and a high degree of branching/cyclicity. This suggests that increasing molecular size reduces the dietary bioaccumulation, as already hypothesized for the bioconcentration ([112], [113], Chapter 4). The increase of BMF with increasing heteroatoms/multiple bonds density may be connected to increased interactions with tissues.

Finally, the presence of at least one pair of N and O at lag 2 (B02[N-O] = 1) leads to a remarkable decrease of the BMF. All the compounds with B02[N-O] = 1 have very low BMF values (logBMF < -1.92), in agreement with what found by Arnot and co-authors [117], which identified several fragments with N and O at lag 2 as responsible for a significant increase of fish biotransformation rates.

Applicability Domain Assessment One of the most common ways to define the model AD is to use a distance-based approach, i.e. to delimit the model chemical space through a threshold on the distance of new compounds from the training data. Chemicals falling within the delimited space are hypothesised to be sufficiently similar to the model data to meet its theoretical assumptions, while too diverse compounds will be probably unreliably predicted.

M1 is based on a distance approach in itself and, thus, its settings were used for AD assessment. The average Manhattan distance of each compound from its k (9) neighbours was used to quantify its similarity to the bulk of the training data. The threshold was optimized in cross-validation as the best compromise between Q^2_{cv} and number of non-predicted compounds (Fig. 7.6). The chosen threshold was 4.3, which (1) improved the statistics on the test set and (2) led to better results than the maximum and minimum distance values (Table 7.3).

Tab. 7.3: Applicability Domain assessment for M1 according to a threshold on average (Aver), minimum (Min) and maximum (Max) distance from the *k* neighbours; *d** = selected threshold.

Criterion		**CV**			**TEST**		
Rule	*d**	RMSE	Q^2_{cv}	%out	RMSE	Q^2_{ext}	%out
None	–	0.55	0.74	–	0.54	0.75	–
Aver	4.3	0.53	0.75	3.00	0.48	0.80	7.41
Min	0.8	0.53	0.76	8.13	0.48	0.81	14.81
Max	2.1	0.53	0.75	2.50	0.51	0.77	1.85

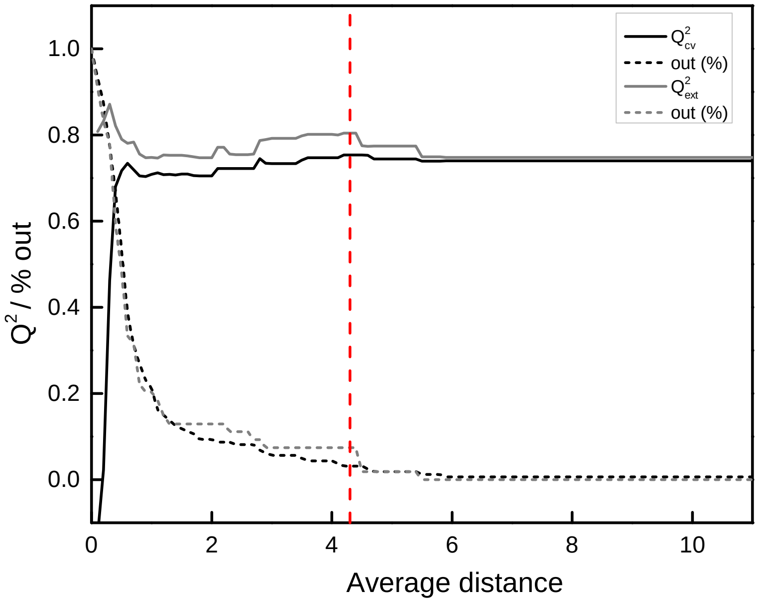


Fig. 7.6: Model M1: AD assessment. Black lines denote Q^2_{cv} and percentage of training compounds outside the AD (% out) according to the average Manhattan distance from the *k* neighbours (threshold). The vertical line represents the chosen threshold (4.3). Statistics on the test set (grey lines) are also reported.

2. Model M2

M2 is an OLS model selected by RSR and it is comprised of 7 molecular descriptors. The model can be expressed as follows:

$$\begin{aligned}\log BMF = &+ 1.66 \cdot \mathit{X0Av} - 0.33 \cdot \mathit{X1Per} - 0.11 \cdot \mathit{SaaaC} \\ &+ 0.89 \cdot \mathit{VE1\_B(m)} + 0.04 \cdot \mathit{MLOGP2} \\ &- 1.26 \cdot \textit{B02[N-O]} - 1.42 \cdot \textit{B03[N-Cl]} - 4.60\end{aligned} \tag{7.9}$$

M2 shares two of its descriptors with M1 (MlogP2 and B02[N-O]). This suggests the relevance of these variables to model the BMF, as they were found independently by the two approaches. The other descriptors are briefly described below.

X0Av

X0Av is the average valence connectivity index of order 0 [105], calculated as follows:

$$\mathit{X0Av} = \frac{\sum_{j=1}^{K} (\delta_j^V)^{1/2}}{K} \tag{7.10}$$

In analogy with X2Av (Eq. 4.7), j runs over all the K paths of order 0, i.e. on the K non-hydrogen atoms; δ^V is the atomic valence connectivity index (Eq. 4.8). X0Av is the mean atomic valence connectivity index of the molecule.

This descriptor accounts for molecular shape, presence of heteroatoms, as well as of double and triple bonds. In particular, X0Av tends to increase when increasing molecular and branching and decreases when increasing the number of heteroatoms and multiple bonds (Fig. 7.7). Being an average value, it is not much sensitive to molecular dimension.

X1Per

X1Per is a perturbation connectivity index [145] that uses the perturbation delta value (δ^p) instead of δ^V, as follows:

$$X1Per = \sum_{j=1}^{K} \prod_{i=1}^{n} (\delta_i^p)_j^{1/2} \tag{7.11}$$

where δ_i^p is the perturbation delta value, which is the valence vertex degrees δ^V modified by the atomic environment, as follows:

$$\delta_i^p = \delta_i^V + 0.1 \cdot \sum_{j=1}^{nSK} a_{ij} \cdot \delta_j^V \tag{7.12}$$

a_{ij} is the element of the adjacency matrix (equal to one only for adjacent vertices and zero otherwise) and nSK is the number of non-H atoms. Thus, the perturbation term of an atom is the sum of the valence vertex degrees of its first neighbours. In analogy with X0Av, X1Per is sensitive to heteroatoms, shape and presence of multiple bonds. In particular, it decreases when increasing the branching, the number of heteroatoms and of multiple bonds and increases when increasing the molecular dimension (Fig. 7.7).

SaaaC

SaaaC is the sum of the electrotopological state of the atoms of type --C(--)-- [146]. For each atom, the electrotopological state is defined as follows:

$$S_i = I_i + \Delta_i = I_i + \sum_{j=1}^{nSK} \frac{I_i - Ij}{(d_{ij}+1)^k} \tag{7.13}$$

where I_i is the intrinsic state of the *i*th atom (Eq. 7.14) and Δ_i is the perturbation of all other atoms on the *i*th atom; d_{ij} is the topological distance between the *i*th and the *j*th atoms; nSK is the number of non-hydrogen atoms in the molecule.

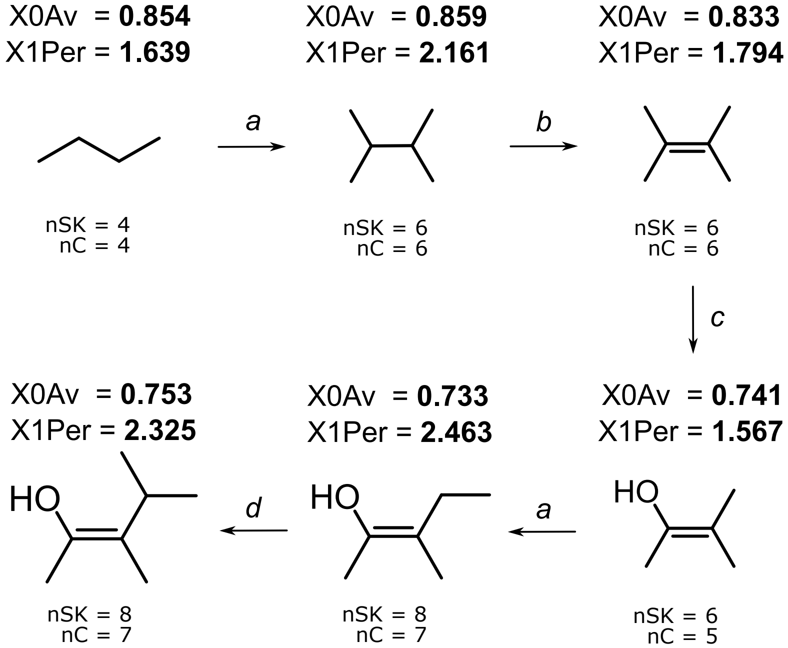


Fig. 7.7: Variation of X0Av and X1Per according to: (a) increasing molecular size, (b) presence of multiple bonds, (c) presence of heteroatoms, (d) branching.

The exponent k tunes the influence of distant atoms; here $k = 2$. The intrinsic state (I_i) is defined as follows:

$$I_i = \frac{(2/L_i)^2 \cdot \delta_i^V + 1}{\delta_i} \tag{7.14}$$

where L_i is the principal quantum number, δ^V is the number of valence electrons (valence vertex degree, Eq. 4.8) and δ is the number of sigma electrons (simple vertex degree) of the ith atom in the H-depleted molecular structure. SaaaC is influenced by the presence of multiple bonds and heteroatoms and has non-zero values for highly branched and/or polycyclic compounds.

VE1_B(m)

VE1_B(m) is a 2D matrix based [147] descriptor, derived from a Burden matrix weighted by mass (**B**(m)). **B**(m) derives from a H-depleted molecular graph as follows: (1) the diagonal elements are atomic carbon-scaled masses (m_i/m_C); (2) the off-diagonal elements corresponding to pairs of bonded atoms are the square roots of conventional bond orders[1]; (3) entries corresponding to terminal bonds are augmented by 0.1; all other matrix elements are set at 0.001. The eigenvalues of the matrix are then computed to calculate the descriptor, as follows:

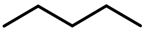
$$VE1\_B(m) = \sum_{i=1}^{nSK} |l_i| \tag{7.15}$$

where l_i is the *i*th coefficient of the last eigenvector of **B**(m) and nSK is the number of non-H atoms. VE1_B(m) depends in a complex way from molecular size, shape, presence of heavy heteroatoms and of multiple bonds. In particular, (1) the presence of heavy heteroatoms will increase the value of the diagonal elements of B(m), (2) molecular size will increase the size of B(m), (3) the other features will increase the off-diagonal values (Fig. 7.8).

B02[N-O] and *B03[N-Cl]*

B02[N-O] and *B*03[N-Cl] are 2D atom pairs descriptors [110], the former is equal to 1 if there is at least one pair of N and 0 atoms separated by 2 bonds, and 0 otherwise; the latter is 1 if there is at least one pair N-Cl separated by three bonds.

[1] equal to 1, 2, 3 for single, double and triple bonds, respectively, and 1.5 for aromatic bonds

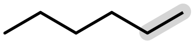


B(m) =

Atoms	C	C	C	C	C
C	1	1.1	0.001	0.001	0.001
C	1.1	1	1	0.001	0.001
C	0.001	1	1	1	0.001
C	0.001	0.001	1	1	1.1
C	0.001	0.001	0.001	1.1	1

VE1_B(m) = **2.172**

a.

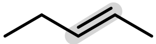


B(m) =

Atoms	C	C	C	C	C	C
C	1	1.1	0.001	0.001	0.001	**0**
C	1.1	1	1	0.001	0.001	**0**
C	0.001	1	1	1	0.001	**0**
C	0.001	0.001	1	1	1	**0**
C	0.001	0.001	0.001	1	1	**1.1**
C	**0.001**	**0.001**	**0.001**	**0.001**	**1.1**	**1**

VE1_B(m) = **2.368**

b.

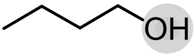


B(m) =

Atoms	C	C	C	C	C
C	1	1.1	0.001	0.001	0.001
C	1.1	1	1	0.001	0.001
C	0.001	1	1	**1.414**	0.001
C	0.001	0.001	**1.414**	1	1.1
C	0.001	0.001	0.001	1.1	1

VE1_B(m) = **2.111**

c.

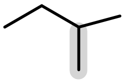


B(m) =

Atoms	C	C	C	C	O
C	1	1.1	0.001	0.001	0.001
C	1.1	1	1	0.001	0.001
C	0.001	1	1	1	0.001
C	0.001	0.001	1	1	1.1
O	0.001	0.001	0.001	1.1	**1.332**

VE1_B(m) = **2.152**

d.



B(m) =

Atoms	C	C	C	C	C
C	1	1.1	0.001	0.001	0.001
C	1.1	1	1	0.001	0.001
C	0.001	1	1	1.1	1.1
C	0.001	0.001	1.1	1	**0.001**
C	0.001	0.001	1.1	**0.001**	1

VE1_B(m) = **2.137**

Fig. 7.8: Factors that influence **B**(m) matrix and the corresponding VE1_B(m) descriptor: (a) molecular size, (b) multiple bonds, (c) heteroatoms, (d) branching.

Mechanistic Interpretation In the case of OLS models, the contribution of the descriptors can be easily interpreted through the standardized regression coefficients (Fig. 7.9), which quantify (a) the relevance of each descriptor (the higher the absolute value, the more relevant), (b) the proportionality with the modelled response (direct proportionality for positive coefficient values and inverse proportionality for negative values).

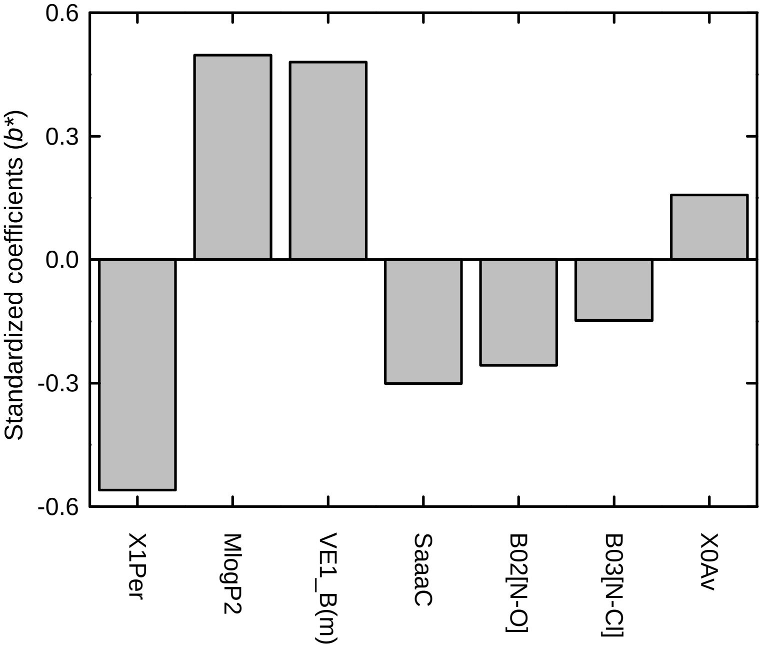


Fig. 7.9: Standardized regression coefficients of M2 descriptors, sorted according to their relevance (the higher the absolute value, the more relevant).

The most relevant descriptor is X1Per, which is inversely proportional to the logBMF (negative coefficient). X1Per decreases when increasing the molecular branching/cyclicity, the number of heteroatoms and of multiple bonds. According to the coefficient, these factors lead to increased logBMF values. This is in agreement with what noted for model M1. The second most relevant descriptor is MlogP2, leading to the same considerations of the previous paragraph.

Despite VE1_B(m) and X1Per have a similar chemical meaning (i.e. they decrease when increasing molecular branching and number of multiple bonds), they have opposite standardized coefficient. VE1_B(m) is more sensitive to molecular linearity and to the presence of atoms with high mass, as it is based on a

Burden matrix weighted by mass. All the molecules have similar VE1_B(m) values, with median and standard deviation equal to 3.587 and 0.575, respectively, but this descriptor has spikes for linear siloxanes and perfluorinated alkyl acids (PFAAs). The former also have high values of X1Per, while the latter have low X1Per values. This means that for siloxanes, high VE1_B(m) values are counterbalanced by high X1Per values, while PFAAs have low X1Per values and, correspondingly, high logBMF. These considerations suggest that the structural features encoded within X1Per and VE1_B(m) are relevant to understand the bioaccumulation behaviour of PFAAs.

The increase of SaaaC value leads to a decrease of logBMF. Only 31 molecules have SaaaC larger than 0 and all have at least two aromatic rings. 28 out of 31 are polycyclic aromatic hydrocarbons (PAHs), and this in agreement with what observed in Chapters 4 and 6. In other words, as for the bioconcentration [99], the fish metabolism affects their final BMF.

Non-null values of N-O at lag 2 and of N-Cl at lag 3 lead to a remarkable decrease of the logBMF, of 1.26 and 1.42 log units, respectively (Eq. 7.9). Similarly to B02[N-O], the relevance of the fragment B03[N-Cl] could be related to a destabilizing effect of N and Cl at topological distance of 3, which could lead to a faster biotransformation.

Part of the information encoded within X0Av overlaps with X1Per, but their coefficients are opposite. In particular, they differ for their sensibility to: (a) molecular size, which leads to an increase of X1Per but not of X1Av, and (b) branching, which increases X0Av and decreases of X1Per. This confirms that the logBMF tends to increase with increasing branching and agrees with what noticed on M1 descriptors.

Applicability Domain Assessment As for M1, a distance-based approach was used to assess the AD. The distance of a given chemical from the center of an OLS model can be quantified by the leverage. The leverage matrix (**H**) is calculated as follows:

$$\mathbf{H} = \mathbf{X}^{\mathsf{T}}(\mathbf{X}_{tr}^{\mathsf{T}}\mathbf{X}_{tr})^{-1}\mathbf{X}^{\mathsf{T}} \tag{7.16}$$

where $\mathbf{X}_{tr}$ is the model matrix; **X** can be the model matrix or a matrix of external data. The diagonal elements of **H** ($h_{i,i}$) are the leverage values of the *i*th chemical. There is no general rule for the determination of the threshold leverage. A warning value is generally three times the average of the training set leverages [44]. Here, a data-driven threshold was optimized from 1.5 to 20 times the average leverage. The lowest threshold giving the highest Q^2_{cv} was chosen (10.5 times the average leverage) and the Q^2_{ext} increased from 0.72 to 0.74, by discarding one test compound (Fig. 7.10).

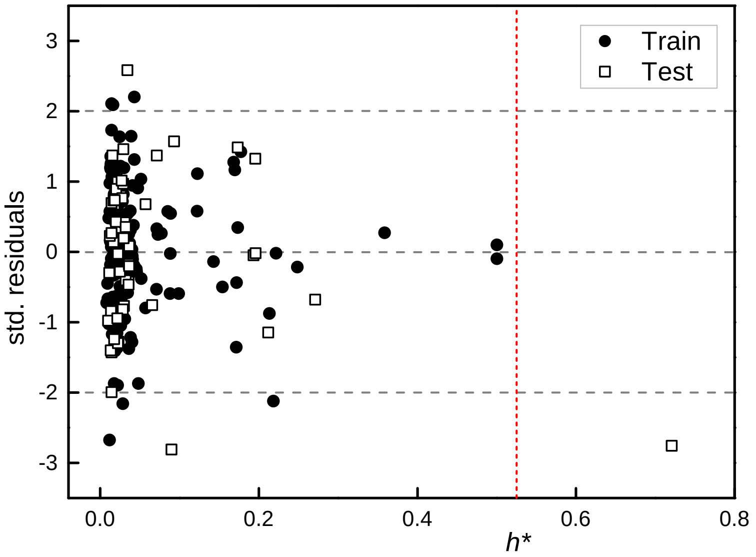


Fig. 7.10: Williams plot for model M2. Standardized residuals in prediction are compared with the leverage. The vertical line is the optimized warning leverage (h^* = 0.525). Horizontal lines are $\pm\sigma$.

3. *Consensus* model

The *consensus* predictions were obtained as the arithmetic mean of the values predicted by M1 and M2. The *consensus* counterbalances the weaknesses of the single sub-models, resulting in an increased performance (Fig. 7.11).

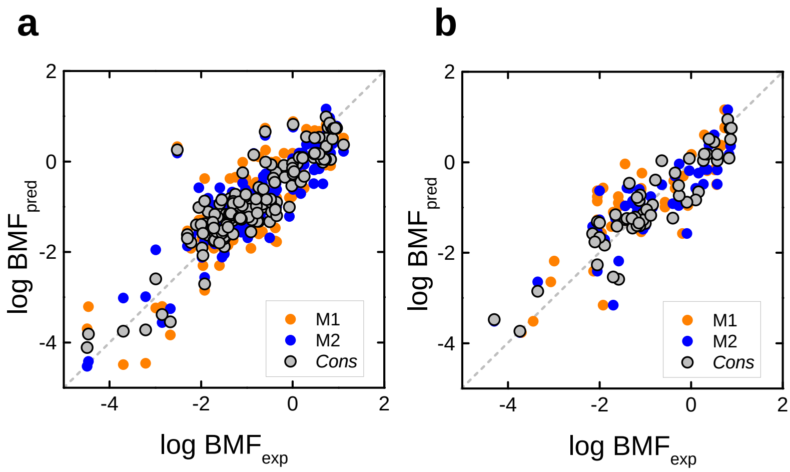


Fig. 7.11: *Consensus* model predictions, experimental *vs* predicted BMF. Predictions are compared with those of the sub-models (M1 and M2) on: (a) training set compounds, (b) test set compounds. Dashed lines represent the perfect fit (experimental = predicted).

The AD of the sub-models can be used to assess a global AD for their *consensus*. In general, one can think that:

- Compounds within the AD of both the models will be the most reliably predicted by the *consensus*(Table 7.4a).

- Compounds within at least one of the two ADs will be less reliably predicted, and the response to be used is that of the model with the appropriate AD (Table 7.4b).

- The predictions of compounds outside both models ADs should not be considered.

The performance of the *consensus* improves on the test set when only the compounds inside the AD of both models are considered. On the training set, no improvement occurs, as all the compounds are inside the AD of M2. The predictions on the compounds within the AD of at least one model are slightly worse, as they rely only on the predictions of one model (Table 7.4b). Thus, the chosen option was the strictest, i.e. to retain compounds within the AD of both the models.

Tab. 7.4: AD assessment for the *consensus* model. Statistics are reported for (a) compounds within the AD of both the models, (b) for the compounds within the AD of at least one model, in comparison with those without AD assessment (c). RMSE, R^2, Q^2 and percentage of compounds outside the AD (*%o*) are reported.

	FIT			**CV**			**TEST**		
	RMSE	R^2	*%o*	RMSE	Q^2	*%o*	RMSE	Q^2	*%o*
a	0.48	0.80	2	0.50	0.78	3	0.43	0.84	9
b	0.47	0.80	0	0.50	0.79	0	0.47	0.81	0
c	0.47	0.80	–	0.50	0.78	–	0.46	0.82	–

7.3 Conclusions

This chapter presented a set of QSAR models to predict the fish BMF of organic chemicals. The models were thoroughly validated and assessed for their applicability domain, in order to allow for a reliable application, especially in regulatory contexts. The models have a satisfactory performance and a mean error comparable with the experimental BMF error, confirming their potential to be used as additional tool for bioaccumulation assessment.

Special attention was posed to understanding the selected molecular descriptors, in order to derive insights into the bioaccumulation through dietary routes.

The most important descriptor was MlogP2, i.e. the squared K_{OW} predicted by the model of Moriguchi [66]. This is the biggest difference with BCF models, which mainly rely on K_{OW}. The fact that MlogP2 is the principal descriptor highlights that also compounds with very negative K_{OW} values can accumulate through diet. It means that dietary bioaccumulation can occur: (a) for very hydrophobic compounds, probably within lipid tissues, and (b) for very hydrophilic compounds, probably through the organism aqueous phases, such as blood.

In analogy with the BCF, the molecular size seems to be a limiting factor for the BMF and it is probably linked to a reduced membrane permeation. However, unlike BCF, the branching leads to an increase of accumulation through diet for small molecules.

Some N-containing structural fragments (B02[N-O] and B03[N-Cl]) resulted to be relevant for a decreased BMF, probably due to increased biotransformation rates, as well as the presence of many condensed aromatic rings.

Finally, the increase of the number of heteroatoms and multiple bonds leads to an increased bioaccumulation through diet, potentially due to increased interactions with tissues.

Part IV

Conclusions

Conclusions

This thesis melded some well-established QSAR and chemometric techniques with some newly developed tools to target the *in silico* prediction of aquatic bioaccumulation. The aim was to detect, analyse and address some of the current open problems in the field. Fish were chosen as the model organism because of their key role in the food chain (e.g. as a food source for humans) and the availability of multi-species data.

The project was structured in two parts, addressing (1) the bioconcentration of chemicals, which, despite its being extensively modelled, still has several limitations, (2) the dietary bioaccumulation of chemicals, for which there is a general lack of models.

The analysis of nine benchmark models for BCF revealed that, in most of the cases, only lipid-driven bioconcentration is predicted well and that other mechanisms affecting the bioconcentration could be neglected. This offered the opportunity to investigate the mechanisms of bioconcentration, by developing a data-driven classification scheme. The developed tool was then used to combine the advantages of existing models on each mechanistic class, in order to maximize the accuracy of BCF prediction towards unknown data.

The comparison of BCF and BMF data highlighted that, in some cases, the BMF could be relevant in assessing the biomagnification potential of chemicals within the food chain. For this reason, the BMF was modelled in a QSAR setting, in compliance with the OECD principles for QSAR validity and regulatory acceptance. To the best of our knowledge, this is the first BMF model for heterogeneous sets of chemicals.

The mechanistic interpretation of the selected molecular descriptors allowed to investigate and rationalize the structural features that may be responsible for the biomagnification of organic chemicals within the food web. This could offer a theoretical basis for predicting the environmental fate of emerging contaminants, such as Perfluorinated compounds. Salient features of the developed approaches are simplicity and interpretability, which can allow for a widespread and transparent application, especially for regulatory purposes.

The future perspectives will be to refine the developed models further. One possibility will be to combine the classification scheme of Chapter 4 with some metabolism-related models, in order to take into account the BCF of the metabolites, if known. Another perspective will be to combine the bioconcentration and the dietary bioaccumulation assessments. For instance, the approach of Chapters 4 and 5 could be used to understand which bioaccumulation route is more relevant for a given compound. This could improve the understanding of the biomagnification process and of associated hazards, and the regulatory assessment of bioaccumulation.

Bibliography

[1] U.S. Environmental Protection Agency (US EPA), *A framework for a computational toxicology research program*, 2003.

[2] H. Raunio, "In silico toxicology–non-testing methods", *Frontiers in pharmacology*, vol. 2, 2011.

[3] R. Kavlock, G. Ankley, J. Blancato, *et al.*, "Computational toxicology–a state of the science mini review.", *Toxicological sciences: An official journal of the Society of Toxicology*, vol. 103, no. 1, p. 14, 2008.

[4] A. P. Worth and M. Balls, *Alternative (non-animal) methods for chemicals testing: Current status and future prospects*. Frame, 2002.

[5] R. Carson, *Silent spring*. Houghton Mifflin Harcourt, 1962.

[6] United Nations Environment Program, "Final act of the conference of plenipotentiaries on the stockholm convention on persistent organic pollutants, 44.", UNEP/POPS/CONF/4UNEP Stockholm, Sweden/Geneva, Switzerland, 2001.

[7] H. J. Geyer, G. G. Rimkus, I. Scheunert, *et al.*, "Bioaccumulation and Occurrence of Endocrine-Disrupting Chemicals (EDCs), Persistent Organic Pollutants (POPs), and Other Organic Compounds in Fish and Other Organisms Including Humans", en, in *Bioaccumulation – New Aspects and Developments*, ser. The Handbook of Environmental Chemistry 2J, B. Beek, Ed., Springer Berlin Heidelberg, 2000, pp. 1–166.

[8] R. van der Oost, J. Beyer, and N. P. Vermeulen, “Fish bioaccumulation and biomarkers in environmental risk assessment: A review”, *Environmental Toxicology and Pharmacology*, vol. 13, no. 2, pp. 57–149, 2003.

[9] D. A. Ratcliffe, “Decrease in eggshell weight in certain birds of prey”, *Nature*, vol. 215, pp. 208 –210, 1967.

[10] G. Woodwell, “Broken eggshells: The miracle of DDT was short-lived, but it helped launch the environmental movement.”, *Science 84*, vol. 5, no. 9, pp. 115–117, 1984.

[11] A. Kudo, Y. Fujikawa, S. Miyahara, *et al.*, “Lessons from Minamata mercury pollution, Japan—after a continuous 22 years of observation”, *Water Science and Technology*, vol. 38, no. 7, pp. 187–193, 1998.

[12] K. Eto, “Minamata disease”, *Neuropathology*, vol. 20, no. s1, pp. 14–19, 2000.

[13] G. M. Woodwell, *Toxic substances and ecological cycles*. WH Freeman, 1967, vol. 1066.

[14] J. L. Hamelink, R. C. Waybrant, and R. C. Ball, “A proposal: Exchange equilibria control the degree chlorinated hydrocarbons are biologically magnified in lentic environments”, *Transactions of the American Fisheries Society*, vol. 100, no. 2, pp. 207–214, 1971.

[15] W. B. Neely, D. R. Branson, and G. E. Blau, “Partition coefficient to measure bioconcentration potential of organic chemicals in fish”, *Environmental Science & Technology*, vol. 8, no. 13, pp. 1113–1115, 1974.

[16] G. D. Veith, D. L. DeFoe, and B. V. Bergstedt, “Measuring and Estimating the Bioconcentration Factor of Chemicals in Fish”, *Journal of the Fisheries Research Board of Canada*, vol. 36, no. 9, pp. 1040–1048, 1979.

[17] D. Mackay, “Correlation of bioconcentration factors”, *Environ. Sci. Technol.*, vol. 16, no. 5, pp. 274–278, 1982.

[18] J. P. Connolly and C. J. Pedersen, “A thermodynamic-based evaluation of organic chemical accumulation in aquatic organisms”, *Environmental science & technology*, vol. 22, no. 1, pp. 99–103, 1988.

[19] M. G. Barron, “Bioconcentration. Will water-borne organic chemicals accumulate in aquatic animals?”, *Environ. Sci. Technol.*, vol. 24, no. 11, pp. 1612–1618, 1990.

[20] J. A. Arnot and F. A. Gobas, “A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms”, *Environmental Reviews*, vol. 14, no. 4, pp. 257–297, 2006.

[21] T. Katagi, “Bioconcentration, Bioaccumulation, and Metabolism of Pesticides in Aquatic Organisms”, en, in *Reviews of Environmental Contamination and Toxicology*, ser. Reviews of Environmental Contamination and Toxicology 204, D. M. Whitacre, Ed., Springer New York, 2010, pp. 1–132.

[22] B. C. Kelly, M. G. Ikonomou, J. D. Blair, A. E. Morin, and F. A. P. C. Gobas, “Food Web–Specific Biomagnification of Persistent Organic Pollutants”, en, *Science*, vol. 317, no. 5835, pp. 236–239, 2007.

[23] K. Borgå, K. A. Kidd, D. C. Muir, *et al.*, “Trophic magnification factors: Considerations of ecology, ecosystems, and study design”, en, *Integrated Environmental Assessment and Management*, vol. 8, no. 1, pp. 64–84, 2012.

[24] F. A. Gobas, W. de Wolf, L. P. Burkhard, E. Verbruggen, and K. Plotzke, “Revisiting Bioaccumulation Criteria for POPs and PBT Assessments”, en, *Integrated Environmental Assessment and Management*, vol. 5, no. 4, pp. 624–637, 2009.

[25] M. T. D. Cronin, J. D. Walker, J. S. Jaworska, *et al.*, “Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances.”, *Environ Health Perspect*, vol. 111, no. 10, pp. 1376–1390, 2003.

[26] ILSI Health and Environmental Sciences Institute, *Workshop on bioaccumulation assessments, dutch congress centre, the hague, the netherlands*, 2006.

[27] M. A. Johnson and G. M. Maggiora, *Concepts and applications of molecular similarity*. Wiley, 1990.

[28] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)*. John Wiley & Sons, 2009, vol. 41.

[29] R. Todeschini, V. Consonni, and P. Gramatica, "Chemometrics in QSAR", in *Comprehensive Chemometrics*, S. Brown, R. Tauler, and B. Walczak, Eds., vol. 4, Oxford: Elsevier, 2009, pp. 129–172.

[30] Talete srl, *Dragon*, version 6.0, 2012.

[31] L. Yu, K. K. Lai, S. Wang, and W. Huang, "A bias-variance-complexity trade-off framework for complex system modeling", in *Computational Science and Its Applications-ICCSA 2006*, Springer, 2006, pp. 518–527.

[32] J. H. Holland, *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. MIT press, 1992.

[33] J. M. Sutter, S. L. Dixon, and P. C. Jurs, "Automated descriptor selection for quantitative structure-activity relationships using generalized simulated annealing", *Journal of chemical information and computer sciences*, vol. 35, no. 1, pp. 77–84, 1995.

[34] J. Kennedy, "Particle swarm optimization", in *Encyclopedia of Machine Learning*, Springer, 2010, pp. 760–766.

[35] M. Goodarzi, M. P. Freitas, and R. Jensen, "Ant colony optimization as a feature selection method in the QSAR modeling of anti-HIV-1 activities of 3-(3, 5-dimethylbenzyl) uracil derivatives using MLR, PLS and SVM regressions", *Chemometrics and intelligent laboratory systems*, vol. 98, no. 2, pp. 123–129, 2009.

[36] M. Cassotti, F. Grisoni, and R. Todeschini, "Reshaped Sequential Replacement algorithm: An efficient approach to variable selection", *Chemometrics and Intelligent Laboratory Systems*, vol. 133, no. 0, pp. 136 –148, 2014.

[37] D. E. Golberg, “Genetic algorithms in search, optimization, and machine learning”, *Addion wesley*, vol. 1989, 1989.

[38] A. J. Miller, “Selection of Subsets of Regression Variables”, English, *Journal of the Royal Statistical Society. Series A (General)*, vol. 147, no. 3, pp. 389–425, 1984.

[39] F. Grisoni, M. Cassotti, and R. Todeschini, “Reshaped Sequential Replacement for variable selection in QSPR: Comparison with other reference methods”, *Journal of Chemometrics*, vol. 28, no. 4, pp. 249–259, 2014.

[40] A. Golbraikh, M. Shen, Z. Xiao, *et al.*, “Rational selection of training and test sets for the development of validated QSAR models”, *Journal of computer-aided molecular design*, vol. 17, no. 2-4, pp. 241–253, 2003.

[41] P. Gramatica, P. Pilutti, and E. Papa, “Validated QSAR prediction of OH tropospheric degradation of VOCs: Splitting into training-test sets and consensus modeling”, *Journal of chemical information and computer sciences*, vol. 44, no. 5, pp. 1794–1802, 2004.

[42] M. Snarey, N. K. Terrett, P. Willett, and D. J. Wilton, “Comparison of algorithms for dissimilarity-based compound selection”, *Journal of Molecular Graphics and Modelling*, vol. 15, no. 6, pp. 372–385, 1997.

[43] V. Consonni, D. Ballabio, and R. Todeschini, “Comments on the Definition of the Q2 Parameter for QSAR Validation”, *J. Chem. Inf. Model.*, vol. 49, no. 7, pp. 1669–1678, 2009.

[44] F. Sahigara, K. Mansouri, D. Ballabio, *et al.*, “Comparison of Different Approaches to Define the Applicability Domain of QSAR Models”, en, *Molecules*, vol. 17, no. 5, pp. 4791–4810, 2012.

[45] European Chemicals Agency (ECHA), “Guidance on information requirements and chemical safety assessment chapter R.6: QSARs and grouping of chemicals”, 2008.

[46] OECD, *Guidance document on the validation of (Q) SAR models*, 2007.

[47] M. W. Woodcroft, D. A. Ellis, S. P. Rafferty, *et al.*, “Experimental characterization of the mechanism of perfluorocarboxylic acids’ liver protein bioaccumulation: The key role of the neutral species”, en, *Environmental Toxicology and Chemistry*, vol. 29, no. 8, pp. 1669–1677, 2010.

[48] M. Pavan, T. I. Netzeva, and A. P. Worth, “Review of Literature-Based Quantitative Structure–Activity Relationship Models for Bioconcentration”, en, *QSAR & Combinatorial Science*, vol. 27, no. 1, pp. 21–31, 2008.

[49] T. Ivanciuc, O. Ivanciuc, and D. J. Klein, “Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative super-structure activity relationships (qssar)”, *Molecular diversity*, vol. 10, no. 2, pp. 133–145, 2006.

[50] K. Mansouri, V. Consonni, M. K. Durjava, *et al.*, “Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by qsar modeling”, *Chemosphere*, vol. 89, no. 4, pp. 433–444, 2012.

[51] ECETOC, “The Role of Bioaccumulation in Environmental Risk Assessment: The Aquatic Environment and Related Food Webs”, *Technical report 67, Brussel, Belgium*, 1995.

[52] W. de Wolf, E. S. Yedema, W. Seinen, and J. L. Hermens, “Bioconcentration kinetics of chlorinated anilines in guppy, Poecilia reticulata”, *Chemosphere*, vol. 28, no. 1, pp. 159–167, 1994.

[53] D. C. Muir, B. R. Hobden, and M. R. Servos, “Bioconcentration of pyrethroid insecticides and DDT by rainbow trout: Uptake, depuration, and effect of dissolved organic carbon”, *Aquatic Toxicology*, vol. 29, no. 3–4, pp. 223–240, 1994.

[54] R. E. Reinert, L. J. Stone, and W. A. Willford, “Effect of temperature on accumulation of methylmercuric chloride and p, p’ DDT by rainbow trout (salmo gairdneri)”, *Journal of the Fisheries Board of Canada*, vol. 31, no. 10, pp. 1649–1652, 1974.

[55] Y. Wen, J. He, X. Liu, J. Li, and Y. Zhao, "Linear and nonlinear relationships between bioconcentration and hydrophobicity: Theoretical consideration", *Environmental Toxicology and Pharmacology*, vol. 34, no. 2, pp. 200–208, 2012.

[56] M. T. Jonker and S. A. van der Heijden, "Bioconcentration factor hydrophobicity cutoff: An artificial phenomenon reconstructed", *Environmental science & technology*, vol. 41, no. 21, pp. 7363–7369, 2007.

[57] S. Bintein, J. Devillers, and W. Karcher, "Nonlinear Dependence of Fish Bioconcentration on n-Octanol/Water Partition Coefficient", *SAR and QSAR in Environmental Research*, vol. 1, no. 1, pp. 29–39, 1993.

[58] European Commission, "Technical Guidance Document (TGD) on risk assessment in support of Commission Directive 93/67/ EEC on risk assessment for new notified substances and Commission Regulation (EC) No 1488/94 on risk assessment for existing substances and Directive 98/8/EC of the European parliament and of the council concerning the placing of biocidal products on the market", *The European Community, Brussels, Belgium*, 2003.

[59] R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics (2 volumes)*. Wiley-VCH, 2009.

[60] U.S. Environmental Protection Agency (US EPA), *BCFBAF*, version 3.01, 2012.

[61] Istituto di Ricerche Farmacologiche Mario Negri, *VEGA noninteractive client*, version 1.0.8, 2013.

[62] A. Gissi, A. Lombardo, A. Roncaglioni, *et al.*, "Evaluation and comparison of benchmark QSAR models to predict a relevant REACH endpoint: The bioconcentration factor (BCF)", *Environmental Research*, vol. 137, pp. 398–409, 2015.

[63] W. M. Meylan, P. H. Howard, R. S. Boethling, *et al.*, "Improved method for estimating bioconcentration/ bioaccumulation factor from octanol/water partition coefficient", en, *Environmental Toxicology and Chemistry*, vol. 18, no. 4, pp. 664–672, 1999.

[64] C. Zhao, E. Boriani, A. Chana, A. Roncaglioni, and E. Benfenati, “A new hybrid system of QSAR models for predicting bioconcentration factors (BCF)”, *Chemosphere*, vol. 73, no. 11, pp. 1701–1707, 2008.

[65] A. Lombardo, A. Roncaglioni, E. Boriani, C. Milan, and E. Benfenati, “Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish”, en, *Chemistry Central Journal*, vol. 4, no. Suppl 1, S1, 2010.

[66] I. Moriguchi, S. Hirono, I. Nakagome, and H. Hirano, “Comparison of Reliability of log P Values for Drugs Calculated by Several Methods”, *Chemical & pharmaceutical bulletin*, vol. 42, no. 4, pp. 976–978, 1994.

[67] F. R. Burden, “Molecular identification number for substructure searches”, *J. Chem. Inf. Comput. Sci.*, vol. 29, no. 3, pp. 225–227, 1989.

[68] R. C. Geary, “The contiguity ratio and statistical mapping”, *The Incorporated Statistician*, vol. 5, no. 3, pp. 115–146, 1954.

[69] P. A. Moran, “Notes on continuous stochastic phenomena”, *Biometrika*, pp. 17–23, 1950.

[70] The MathWorks Inc., *Matlab*, version 7.1.4.0, (R2012a), Natick, Massachusetts, 2012.

[71] D. Rogers and M. Hahn, “Extended-connectivity fingerprints”, *Journal of chemical information and modeling*, vol. 50, no. 5, pp. 742–754, 2010.

[72] M. Floris, A. Manganaro, O. Nicolotti, *et al.*, “A generalizable definition of chemical similarity for read-across”, *Journal of Cheminformatics*, vol. 6, no. 1, pp. 1–7, 2014.

[73] NCI/CADD Group, *Chemical Identifier Resolver*, 2013.

[74] NCBI, *The PubChem Project, www.pubchem.ncbi.nlm.nih.gov*, 2013.

[75] Royal Society of Chemistry, *ChemSpider, www.chemspider.com*, 2013.

[76] M. T. D. Cronin and D. Livingstone, *Predicting Chemical Toxicity and Fate*, en. CRC Press, 2004.

[77] VCCLAB, *Virtual Computational Chemistry Laboratory*, 2005.

[78] D. Mackay, W.-Y. Shiu, K.-C. Ma, and S. C. Lee, *Handbook of Physical-Chemical Properties and Environmental Fate for Organic Chemicals, Second Edition*, en. CRC Press, 2010.

[79] European Chemicals Agency (ECHA), *QSAR toolbox*, version 3.1, 2013.

[80] HURO/0901/037/2.2.2, *Complex Molecular Database for environmental protection*, 2011.

[81] CEFIC LRI, *EURAS bioconcentration factor (BCF) Gold Standard Database*, 2006.

[82] Sangster Research Laboratories, *LOGKOW Database*, 2013.

[83] D. B. Rorabacher, "Statistical treatment for rejection of deviant values: Critical values of Dixon's" Q" parameter and related subrange ratios at the 95% confidence level", *Analytical Chemistry*, vol. 63, no. 2, pp. 139–146, 1991.

[84] M. R. Berthold, N. Cebron, F. Dill, *et al.*, "KNIME: The Konstanz Information Miner", Springer, 2007.

[85] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

[86] K. Mansouri, V. Consonni, M. K. Durjava, *et al.*, "Assessing bioaccumulation of polybrominated diphenyl ethers for aquatic species by QSAR modeling", *Chemosphere*, vol. 89, no. 4, pp. 433–444, 2012.

[87] M. Chessells, D. W. Hawker, and D. W. Connell, "Critical evaluation of the measurement of the 1- octanol / water partition coefficient of hydrophobic compounds", *Chemosphere*, vol. 22, no. 12, pp. 1175–1190, 1991.

[88] L. M. L. Nollet and L. S. P. D. Gelder, *Handbook of Water Analysis, Second Edition*, en. CRC Press, 2007.

[89] Guidechem Chemical Network, *http://www.guidechem.com/*.

[90] Government of Canada - Environment Canada, *Health Canada reports*, eng, 2006.

[91] J. M. Armitage, J. A. Arnot, F. Wania, and D. Mackay, "Development and evaluation of a mechanistic bioconcentration model for ionogenic organic chemicals in fish", en, *Environmental Toxicology and Chemistry*, vol. 32, no. 1, pp. 115–128, 2013.

[92] P. D. Jones, W. Hu, W. De Coen, J. L. Newsted, and J. P. Giesy, "Binding of perfluorinated fatty acids to serum proteins", en, *Environmental Toxicology and Chemistry*, vol. 22, no. 11, pp. 2639–2649, 2003.

[93] H. J. Lehmler and P. M. Bummer, "Mixing of perfluorinated carboxylic acids with dipalmitoylphosphatidylcholine", *Biochim. et Biophysica Acta (BBA) - Biomembranes*, vol. 1664, no. 2, pp. 141–149, 2004.

[94] W. Xie, G. D. Bothun, and H.-J. Lehmler, "Partitioning of perfluorooctanoate into phosphatidylcholine bilayers is chain length-independent", *Chemistry and Physics of Lipids*, vol. 163, no. 3, pp. 300–308, 2010.

[95] G. G. Rimkus, "Polycyclic musk fragrances in the aquatic environment", *Toxicology Letters*, vol. 111, no. 1–2, pp. 37–56, 1999.

[96] J. de Bruijn and J. Hermens, "Uptake and elimination kinetics of organophosphorous pesticides in the guppy (poecilia reticulata): Correlations with the octanol/water partition coefficient", en, *Environmental Toxicology and Chemistry*, vol. 10, no. 6, pp. 791–804, 1991.

[97] J. Lech and M. Vodicnik, "Biotransformation", *Fundamentals of Aquatic Toxicology: Methods and Applications. Hemisphere Publishing Corporation Washington DC. 1985. p 526-557, 16 fig, 6 tab, 50 ref.*, 1985.

[98] D. T. H. M. Sijm and A. Opperhuizen, "Biotransformation of Organic Chemicals by Fish: Enzyme Activities and Reactions", en, in *Reactions and Processes*, ser. The Handbook of Environmental Chemistry 2 / 2E, Springer Berlin Heidelberg, 1989, pp. 163–235.

[99] G. Jonsson, R. K. Bechmann, S. D. Bamber, and T. Baussant, “Bioconcentration, biotransformation, and elimination of polycyclic aromatic hydrocarbons in sheepshead minnows (Cyprinodon variegatus) Exposed to Contaminated Seawater”, en, *Environmental Toxicology and Chemistry*, vol. 23, no. 6, pp. 1538–1548, 2004.

[100] A. H. Buckman, C. S. Wong, E. A. Chow, *et al.*, “Biotransformation of polychlorinated biphenyls (PCBs) and bioformation of hydroxylated PCBs in fish”, *Aquatic Toxicology*, vol. 78, no. 2, pp. 176–185, 2006.

[101] M. J. Melancon and J. J. Lech, “Isolation and identification of a polar metabolite of tetrachlorobiphenyl from bile of rainbow trout exposed to14c-tetrachlorobiphenyl”, en, *Bull. Environ. Contam. Toxicol.*, vol. 15, no. 2, pp. 181–188, 2013.

[102] J. de Boer, F. van der Valk, M. A. T. Kerkhoff, P. Hagel, and U. A. T. Brinkman, “An 8-Year Study on the Elimination of PCBs and Other Organochlorine Compounds from Eel (Anguilla anguilla) under Natural Conditions”, *Environ. Sci. Technol.*, vol. 28, no. 13, pp. 2242–2248, 1994.

[103] J. de Boer, C. J. N. Stronck, W. A. Traag, and J. van der Meer, “Non-ortho and mono-ortho substituted chlorobiphenyls and chlorinated dibenzo-p-dioxins and dibenzofurans in marine and freshwater fish and shellfish from The Netherlands”, *Chemosphere*, vol. 26, no. 10, pp. 1823–1842, 1993.

[104] Y. Wang, Y. Wen, J. J. Li, *et al.*, “Investigation on the relationship between bioconcentration factor and distribution coefficient based on class-based compounds: The factors that affect bioconcentration”, *Environmental Toxicology and Pharmacology*, vol. 38, no. 2, pp. 388–396, 2014.

[105] L. B. Kier and L. H. Hall, “Derivation and significance of valence molecular connectivity”, en, *J. Pharm. Sci.*, vol. 70, no. 6, pp. 583–589, 1981.

[106] E. A. Meyer, R. K. Castellano, and F. Diederich, “Interactions with Aromatic Rings in Chemical and Biological Recognition”, en, *Angewandte Chemie International Edition*, vol. 42, no. 11, pp. 1210–1250, 2003.

[107] S. Agatonovic-Kustrin, R. Beresford, and A. P. M. Yusof, “Theoretically-derived molecular descriptors important in human intestinal absorption”, *Journal of Pharmaceutical and Biomedical Analysis*, vol. 25, no. 2, pp. 227–237, 2001.

[108] S. D. Dimitrov, O. G. Mekenyan, and J. D. Walker, “Non-linear modeling of bioconcentration using partition coefficients for narcotic chemicals”, *SAR and QSAR in Environmental Research*, vol. 13, no. 1, pp. 177–184, 2002.

[109] D. Bonchev and N. Trinajstič, “Overall Molecular Descriptors. 3. Overall Zagreb Indices”, *SAR and QSAR in Environmental Research*, vol. 12, no. 1-2, pp. 213–236, 2001.

[110] R. E. Carhart, D. H. Smith, and R. Venkataraghavan, “Atom pairs as molecular features in structure-activity studies: Definition and applications”, *J. Chem. Inf. Comput. Sci.*, vol. 25, no. 2, pp. 64–73, 1985.

[111] V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, and R. K. Robins, “Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics”, *J. Chem. Inf. Comput. Sci.*, vol. 29, no. 3, pp. 163–172, 1989.

[112] S. Dimitrov, N. Dimitrova, T. Parkerton, *et al.*, “Base-line model for identifying the bioaccumulation potential of chemicals”, *SAR and QSAR in Environmental Research*, vol. 16, no. 6, pp. 531–554, 2005.

[113] S. Dimitrov, N. Dimitrova, J. Walker, G. Veith, and O. Mekenyan, “Predicting bioconcentration factors of highly hydrophobic chemicals. Effects of molecular size”, *Pure and Applied Chemistry*, vol. 74, no. 10, pp. 1823–1830, 2002.

[114] A. Opperhuizen, E. W. v. d. Volde, F. A. P. C. Gobas, *et al.*, “Relationship between bioconcentration in fish and steric factors of hydrophobic chemicals”, *Chemosphere*, vol. 14, no. 11–12, pp. 1871–1896, 1985.

[115] J. A. Arnot, M. I. Arnot, D. Mackay, *et al.*, "Molecular size cutoff criteria for screening bioaccumulation potential: Fact or fiction?", en, *Integr Environ Assess Manag*, vol. 6, no. 2, pp. 210–224, 2010.

[116] S. M. Schrap and A. Opperhuizen, "Relationship between bioavailability and hydrophobicity: Reduction of the uptake of organic chemicals by fish due to the sorption on particles", en, *Environmental Toxicology and Chemistry*, vol. 9, no. 6, pp. 715–724, 1990.

[117] J. A. Arnot, W. Meylan, J. Tunkel, *et al.*, "A quantitative structure-activity relationship for predicting metabolic biotransformation rates for organic chemicals in fish", en, *Environmental Toxicology and Chemistry*, vol. 28, no. 6, pp. 1168–1177, 2009.

[118] E. J. Matthews and J. F. Contrera, "A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASEQSAR-ES software", *Regulatory Toxicology and Pharmacology*, vol. 28, no. 3, pp. 242–264, 1998.

[119] G. Klopman and S. K. Chakravarti, "Screening of high production volume chemicals for estrogen receptor binding activity (II) by the MultiCASE expert system", *Chemosphere*, vol. 51, no. 6, pp. 461–468, 2003.

[120] E. J. Matthews, N. L. Kruhlak, R. D. Benz, *et al.*, "Identification of structure–activity relationships for adverse effects of pharmaceuticals in humans: Part C: Use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities", *Regulatory toxicology and pharmacology*, vol. 54, no. 1, pp. 43–65, 2009.

[121] J. S. Jaworska, M Comber, C Auer, and C. J. Van Leeuwen, "Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints.", *Environ Health Perspect*, vol. 111, no. 10, pp. 1358–1360, 2003.

[122] W. M. Meylan and P. H. Howard, “Atom/fragment contribution method for estimating octanol- water partition coefficients”, *Journal of pharmaceutical sciences*, vol. 84, no. 1, pp. 83–92, 1995.

[123] A. K. Ghose and G. M. Crippen, “Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. partition coefficients as a measure of hydrophobicity”, 1986.

[124] Istituto di Ricerche Farmacologiche Mario Negri, *VEGA non-interactive client*, version 1.1.0, 2015.

[125] C. I. Cappelli, E. Benfenati, and J. Cester, “Evaluation of QSAR models for predicting the partition coefficient (log P) of chemicals under the REACH regulation”, *Environmental Research*, vol. 143, Part A, pp. 26–32, 2015.

[126] F. Grisoni, V. Consonni, S. Nembri, and R. Todeschini, “How to weight Hasse matrices and reduce incomparabilities”, *Chemometrics and Intelligent Laboratory Systems*, vol. 147, pp. 95–104, 2015.

[127] H. Hasse and J. Martinet, *Uber die Klassenzahl abelscher Zahlkörper*. Citeseer, 1985, vol. 1.

[128] J. A. Arnot and C. L. Quinn, “Development and Evaluation of a Database of Dietary Bioaccumulation Test Data for Organic Chemicals in Fish”, *Environ. Sci. Technol.*, vol. 49, no. 8, pp. 4783–4796, 2015.

[129] O. for Economic Co-operation and D. (OECD), *OECD Guidelines for the Testing of Chemicals: Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure*. Paris, 2012.

[130] Chemical Abstract Service, *Scifinder, www.scifinder.cas.org*, 2015.

[131] T. Kohonen, “Self-organization and associative memory”, 1988.

[132] B. J. Konwick, A. W. Garrison, J. K. Avants, and A. T. Fisk, “Bioaccumulation and biotransformation of chiral triazole fungicides in rainbow trout (oncorhynchus mykiss)”, *Aquatic toxicology*, vol. 80, no. 4, pp. 372–381, 2006.

[133] A. Tropsha, P. Gramatica, and V. K. Gombar, “The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models”, *QSAR & Combinatorial Science*, vol. 22, no. 1, pp. 69–77, 2003.

[134] V. Consonni, D. Ballabio, and R. Todeschini, “Evaluation of model predictive ability by external validation techniques”, en, *Journal of Chemometrics*, vol. 24, no. 3-4, pp. 194–201, 2010.

[135] V. Vapnik, S. E. Golowich, and A. Smola, “Support vector method for function approximation, regression estimation, and signal processing”, in *Advances in Neural Information Processing Systems 9*, Citeseer, 1996.

[136] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2009.

[137] H. Wold, “Partial least squares”, *Encyclopedia of statistical sciences*, 1985.

[138] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso”, *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996, ArticleType: research-article / Full publication date: 1996 / Copyright © 1996 Royal Statistical Society.

[139] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression”, *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.

[140] R. Todeschini, D. Ballabio, F. Grisoni, and V. Consonni, “A new concept of higher-order similarity and the role of the distance/similarity measures in local classification methods.”, *Chemometrics and Intelligent Laboratory Systems (to be submitted)*, 2015.

[141] R. Todeschini, D. Ballabio, M. Cassotti, and V. Consonni, “N3 and BNN: Two new similarity based classification methods in comparison with other classifiers”, *Journal of chemical information and modeling*, vol. 55, no. 11, pp. 2365–2374, 2015.

[142] R. Leardi, R. Boggia, and M. Terrile, “Genetic algorithms as a strategy for feature selection”, en, *Journal of Chemometrics*, vol. 6, no. 5, pp. 267–281, 1992.

[143] I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[144] I. Jolliffe, “Principal Component Analysis”, en, in *Wiley StatsRef: STATISTICS Reference Online*, John Wiley & Sons, Ltd, 2014.

[145] R. Todeschini and V. Consonni, “New local vertex invariants and molecular descriptors based on functions of the vertex degrees”, *MATCH, Communications in Mathematical and in Computer Chemistry*, vol. 64, pp. 359–372, 2010.

[146] L. B. Kier and L. H. Hall, “An Electrotopological-State Index for Atoms in Molecules”, en, *Pharm Res*, vol. 7, no. 8, pp. 801–807, 1990.

[147] V. Consonni and R. Todeschini, “New spectral indices for molecule description”, *MATCH, Communications in Mathematical and in Computer Chemistry*, vol. 60, no. 1, pp. 3–14, 2008.