UNIVERSITY OF MILANO – BICOCCA

DEPARTMENT OF EARTH AND ENVIRONMENTAL SCIENCES

Doctoral Degree Course in Chemical Sciences

Cycle XXVII

Ph.D. Thesis

QSAR study of aquatic toxicity by chemometrics methods in the framework of REACH regulation

Matteo Cassotti

Tutor: Prof. Roberto Todeschini

Co-Tutor: Dr. Viviana Consonni

                 Dr. Davide Ballabio

Academic year: 2013-2014

Acknowledgements

Above all, I would like to thank my supervisor Prof. Roberto Todeschini for giving me the opportunity to undertake this Ph.D. project, being always open to teach me chemometrics and involving me in other interesting activities.

Special thanks to Dr. Davide Ballabio who supported and revised my work, taught me how to use MATLAB and asked me to help with teaching.

I am grateful to Dr. Viviana Consonni and Dr. Andrea Mauri, especially for the guidance and help about molecular descriptors and chemoinformatics aspects.

I acknowledge Dr. Igor Tetko who allowed me to initiate this work within the ECO project, which was a positive international experience.

I wish to thank Dr. Eva Bay Wedebye and Dr. Nikolai Georgiev Nikolov for the good discussions and help in preparing the data.

I am thankful to Prof. Rasmus Bro for suggesting new chemometrics methods and ways (free of charge) to learn Danish.

I would like to thank all the other people that I met in the lab in these years, who gave their contribution on a work and social level: Kamel, Faizan, Kai, Pantelis, Valentina, Ioana, Stefan, Alberto, Francesca, Eva, Svava, Rikke, Sine, Marianne, Monika.

In the end, I am thankful to Tine for the scientific discussions, emotional support, networking and for believing in me.

To those who believe in me

more than I do

Contents

Abbreviations

Abbreviation	Meaning
AD	Applicability Domain
(A)NN	(Artificial) Neural Network
ASM	All Subset Models
CMR	Carcinogenic, Mutagenic or toxic to Reproduction
e.g.	From Latin "exempli gratia", stands for "for instance/for example"
GAs	Genetic Algorithms
GPR	Gaussian Process Regression
i.e.	From Latin "id est", stands for "that is"
kNN	K-Nearest Neighbours
LC$_{50}$	Lethal Concentration causing death in 50% test organisms
LogP	n-octanol-water partition coefficient
LV	Latent Variable
MLR	Multiple Linear Regression
MoA	Mode of Action
OLS	Ordinary Least Squares
PBT	Persistent, Bioaccumulative and Toxic
PCA	Principal Component Analysis
PLS	Partial Least Squares
PRESS	Predictive Error Sum of Squares
QSAR	Quantitative Structure-Activity Relationship
RMSEC	Root Mean Square Error in Calculation
RMSECV	Root Mean Square Error in Cross-Validation
RMSEP	Root Mean Square Error in Prediction
RSR	Reshaped Sequential Replacement
RSS	Residual Sum of Squares
SMILES	Simplified Molecular Input Line Entry System
SR	Sequential Replacement
SVHC	Substance of Very High Concern
SVR	Support Vector Regression
TSS	Total Sum of Squares
vPvB	very Persistent and very Bioaccumulative

Preface

Rationale and scopes

The rationale of the work carried out in this project stems from the deep impact that the introduction of the European regulation on the registration, evaluation, authorization and restriction of chemicals (REACH) had on several components of the social matrix, industry, regulatory bodies and the scientific community first. This work intends to build bridges between these components in order to assist the implementation of REACH on one side, and contribute to the reduction of animal tests through the application of *in silico* methodologies on the other side.

The final scope is the development of quantitative structure-activity relationship (QSAR) models that can be applied in the framework of REACH to fill data gaps of substances lacking experimental data. To this end, models were developed aiming to comply to the full extent with the requirements of REACH in order to reduce potential limitations to their application in the regulatory framework. In particular, this study focuses on the analysis of the acute toxicity of chemicals towards two species of aquatic organisms, namely *Daphnia magna* (water flea) and *Pimephales promelas* (fathead minnow), being this information required for most substances subject to REACH (Annexes VII and VIII).

This work was initiated and partly funded during a short-term fellowship within the Marie Curie initial training network - environmental chemoinformatics project (ECO-ITN, grant agreement no. 238701). The ECO project was achieved through a joint action involving seven institutions in five EU member Countries and aimed at training a new generation of researchers in the field of environmental sciences.

Thesis outline

This thesis is organised in six main chapters summarised as follows.

The Introduction chapter depicts the state of the art of the three main fields that flow into this work, i.e. toxicology, REACH regulation and QSAR analysis. Basic concepts of toxicology are given prior to a more detailed treatment of aquatic toxicology that encompasses some types of experimental tests and the information that can be derived. The European REACH regulation is introduced with a viewpoint on the rationale behind its development and the scenarios opened from its introduction, with particular focus on the implications for QSAR analysis. The main features of current QSAR investigations are given in a nutshell before describing its role in the regulatory context. In the end, the state of the art of QSAR analysis for the prediction of the toxicity towards *Daphnia magna* and *Pimephales promelas* is outlined.

The Data chapter begins with an overview of experimental data and related issues. Afterwards, the sources of data and the treatment procedures adopted in order to define the datasets used to derive the QSAR models are described.

The Methods chapter serves to detail the approaches employed for the development of QSAR models and their evaluation. First, the molecular descriptors and fingerprints used to describe the molecular structures are introduced. The algorithms employed to select the relevant molecular descriptors are then outlined, including an algorithm developed in this project. The mathematical methods used to calibrate the regression models are then described together with the validation techniques and the statistical parameters used to evaluate the goodness-of-fit and predictive power of the models. Eventually, the approaches for the assessment of the applicability domain of the models are discussed.

The Results chapters report and discuss the main findings of this work, separately for *Daphnia magna* and *Pimephales promelas*. The models developed with different approaches are compared and discussed in order to give insight into the advantages and drawbacks of each of them. The models chosen as most appropriate are analysed for their scientific validity and discussed in the light of literature models and the requirements of REACH.

The last chapter draws conclusions from the previous discussion and enunciates the future perspectives for this work.

The relevant research articles published in peer-reviewed scientific journals are reported in the appendices.

CHAPTER 1
Introduction

'All things are poisons, for there is nothing without poisonous qualities. It is only the dose which makes a thing poison.'

Paracelsus (born Philippus Aureolus Theophrastus Bombastus von Hohenheim), Responsio ad quasdam accusationes & calumnias suorum aemulorum et obtrectatorum. Defensio III. Descriptionis & designationis nouorum Receptorum, 1538.

1.1 Aquatic toxicity

In its broad meaning, toxicity refers to the ability of a substance to cause damages to a living organism. Toxicity comprises very seldom a single molecular event; rather, it encompasses a series of events that begins with the exposure of an organism and ends, in case via metabolic processes, with the interaction of the toxicant with target macromolecules. The xenobiotic-macromolecule interaction gives rise to the expression of a toxic endpoint. This sequence of events can also be mitigated by excretion/elimination or repair processes.

Since toxicity is a quantitative concept, every substance has the potential to become a lethal toxicant above certain doses/concentrations, while it results completely harmless below lower levels. In between these two limits, there is a range of potential effects that span from low chronic toxicity to instant lethality. The importance of the dose is epitomised by essential metals (e.g. copper, iron, zinc, etc.), whose lack in food diet can cause pathologies, but high doses of which can generate serious toxic effects as well.

The toxic effect of a substance depends on both its concentration and the duration of the exposure. In order for the exposure to occur, the substance needs to be available for absorption. In fact, a substance that is toxic to a specific organism is harmless if it is not absorbed. The availability for absorption depends on several

factors, including the chemical-physical state of the toxicant, the physiological, morphological and pathological features of the organism and the environmental conditions.

A chemical substance that enters the environment can undergo biotic and abiotic degradation processes that determine its chemical-physical nature and therefore its persistence, reactivity with environmental constituents, partitioning in the environmental compartments and biomagnification phenomena. Since individuals are not isolated, and populations are connected with one another, damages to members of a certain species can spread, threatening the survival of other parts of the ecosystem [Newsome et al., (1996)]. A scheme of the potential repercussions on the ecosystem due to exposure to chemicals is depicted in Figure 1.1. After the exposure, either the cell is able to repair the damages or, when the critical level is surpassed, a biochemical effect is triggered, which spreads at level of organism. The pattern applies to every tier and can eventually lead to the destruction of an ecosystem.
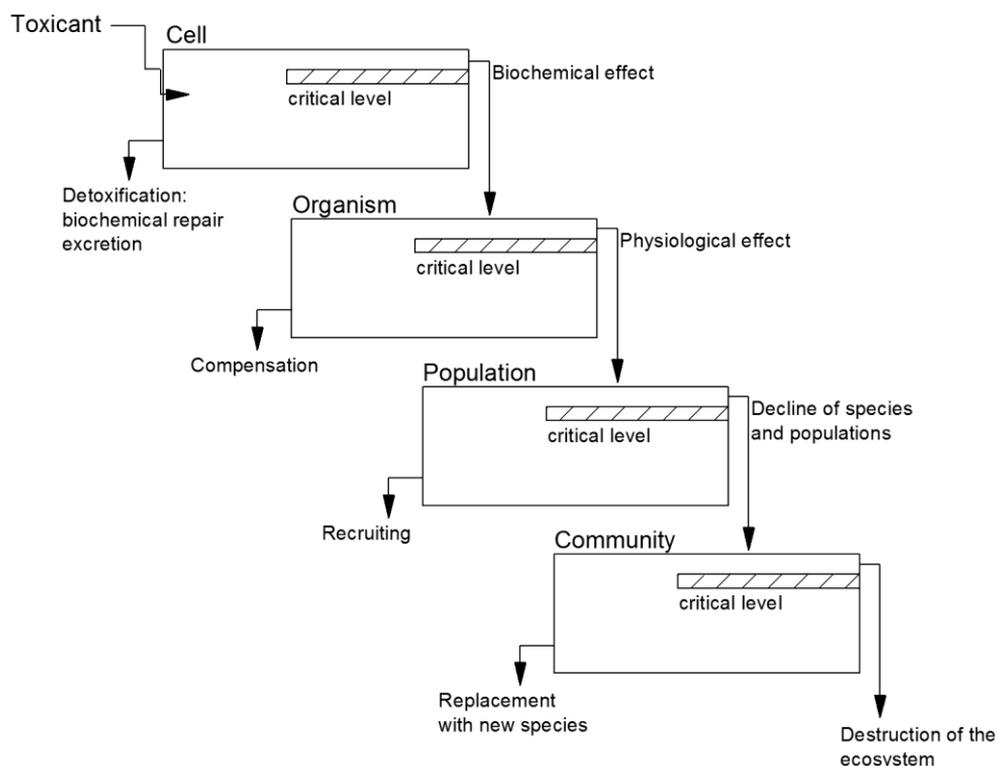
Figure 1.1. Scheme of the potential consequences on the ecosystem of exposure to chemicals.

The aquatic ecosystem is of particular concern because many chemicals released into the environment, either from direct discharge into water or from terrestrial runoff and atmospheric deposition, eventually partition in water [Pritchard, (1993)]. Furthermore, there is a number of characteristics that make aquatic environments more sensitive to contaminants, such as the low level of oxygen, which makes persistence a major problem, the restricted habitat of aquatic organisms and the high permeability of their skin, gills and eggs (for fish and amphibians) [Klaassen, (2001)].

Aquatic toxicity tests serve to measure the effects of chemical exposure on a variety of endpoints, including survival, reproduction, and physiologic and biochemical responses. The duration allows distinguishing between chronic tests, in which organisms are exposed for long periods to low concentrations of chemicals, and acute tests, designed with short exposure times and relatively high concentrations. Different designs are possible, namely flow-through, static renewal and static systems, where test water is continually, periodically and not renewed, respectively. Aside laboratory, field studies can also be conducted. In manipulative field studies, the level of contamination is under control of the experimenter and these tests are carried out in artificial habitats comprising more than one species of test organism (microcosm or mesocosm). In observational field studies, instead, the level of contamination is not under control of the experimenter [Klaassen, (2001)]. Different parameters can be calculated from aquatic toxicity tests. Often the results of acute tests are expressed in terms of effective or lethal concentrations (EC_x or LC_x, respectively), which represent the concentration at which a certain percentage, x, of the maximum effect occurs. Typically, a percentage equal to fifty (EC_{50} or LC_{50}) is used. From chronic tests, parameters such as the no observable effect concentration (NOEC) or lowest observed effect concentration (LOEC) can be derived. These two indices correspond to the highest concentration at which no statistically significant effect was observed and the lowest concentration at which a statistically significant effect was detected compared to the control, respectively. Different test conditions can have a great impact on the measured values and sometimes differences of orders of magnitude are encountered in the toxicity parameters for the same chemical (paragraph 2.2).

The information derived from different types of tests allowed to discover that chemicals can exert toxicity via a variety of different mechanisms, known as modes of toxic action (MoAs). According to this understanding, contaminants can be

divided into two main categories based on the way they interact with the host organism. Some toxicants can disturb the normal cellular functions without interacting with specific target macromolecules. This type of toxicity is known as narcosis and is characterised by lethargy, unconsciousness and, eventually, death, without signs of other specific symptoms [Veith and Broderius, (1990)]. Narcosis, regarded as being the baseline or minimum toxicity, derives essentially from the presence of toxicants within the cell or cellular membrane and, thus, is driven by the ability of the molecule to partition in the organism. Consequently, the relative toxic potency is a function of the lipophilicity. Examples of narcotic chemicals are chlorinated hydrocarbons, linear ethers, aliphatic alcohols (not propargylic and allilic), aliphatic secondary and tertiary amines [Newsome et al., (1996); Verhaar et al., (1992)]. Other contaminants (or their metabolites) can interact directly with critical biological macromolecules. These interactions/reactions give an additional contribution to the narcotic toxicity. The resulting toxic effect deviates from that expected based on the sole narcosis models and, therefore, these compounds are said to exert an excess toxicity. Examples of reactive chemicals are alkylating agents, epoxides, nitrogen and sulphur mustards [Verhaar et al., (1992)].

Several studies further investigated the reactivity of chemicals towards aquatic organisms with the aim to further separate the two aforementioned categories (narcotic and non-narcotic chemicals) into classes of chemicals that share the same (or at least similar) modes of toxic action (MoAs). To this end, data from a variety of tests, such as joint toxicity studies [Broderius and Kahl, (1985); Broderius et al., (1995)], behavioural assessments [Drummond and Russom, (1990); Drummond et al., (1986)], dose-response curves and fish acute toxicity syndrome (FATS) investigations [McKim et al., (1987a); McKim et al., (1987b); Bradbury et al., (1989); Bradbury et al., (1991)], were combined. For example, some narcotics were noticed to be slightly more toxic than baseline narcosis models estimate. Some authors suggested to discriminate between two types of narcosis, often referred to as type I and type II. Type II narcotics are said to be "polar narcotics" because they feature a polar group, which has often been related to the higher toxic potency [Veith and Broderius, (1990); Broderius et al., (1995)]. Nevertheless, Vaes et al., (1998) consider the difference between non-polar and polar narcotics to be an artefact generated by the misuse of the *n*-octanol – water partition coefficient (LogP) as parameter. They showed that if the membrane (L-α-dimyristoyl phosphatidyl-choline [DMPC]) – water partition coefficient (LogK$_{DMPC}$) is used in place of LogP, no distinction

between type I and type II narcotics is observed. Regarding chemicals that exert excess toxicity, different sub-classifications were proposed. Verhaar et al., (1992) considered two broad categories, namely reactive chemicals and specifically acting chemicals, the second class comprising compounds whose specific biological target is known. Russom et al., (1997) established a database were 461 chemicals were associated with one of eight modes of action for their toxicity to the *Pimephales promelas* (fathead minnow). The eight recognised MoAs were Narcosis I (non-polar narcosis), II (polar narcosis) and III (ester narcosis), acetylcholinesterase (AChE) inhibitors, central nervous system (CNS) seizure agents, oxidative phosphorylation uncouplers, respiratory inhibitors and electrophiles/pro-electrophiles.

1.2 REACH regulation

The Regulation (EC) N° 1907/2006 of the European Parliament and of the Council concerning the registration, evaluation, authorisation and restriction of chemicals (REACH) entered into force in 2007. REACH regulation was born to address previous issues, including:

- Lack of information about the properties of several chemicals.

- Only few thousands substances addressed by previous legislation.

- Inadequate risk control.

- Poor information regarding risk assessment procedures between EU member Countries.

REACH was designed to pursue also other general objectives including mapping of chemicals circulating over Europe, gain in depth knowledge about their effects on human health and the environment, definition of substance-hazard-use correlations (concept of identified use), replacement of hazardous substances (CMR, PBT, vPvB) with safer alternatives and provision EU member Countries with a common and simplified supranational legislative framework. This system was implemented through four main actions, namely:

- Registration of substances imported or manufactured in quantities larger than one tonne/year.

- Evaluation of substances in terms of safety.

- Authorisation for substances of very high concern (SVHC).

1.2 REACH regulation

- Restrictions to use.

Contextually, the European Chemicals Agency (ECHA), based in Helsinki (Finland), was established as the regulatory authority appointed to drive the implementation of REACH.

The domain of applicability of REACH is very broad since the registration of chemicals under REACH is required for:

- All substances imported or manufactured in quantities greater than one tonne/year.

- Monomers in polymers if present in percentages equal to or greater than 2% weight by weight and if the total quantity of monomer is greater than one tonne/year.

- Substances in articles if the total amount is greater than one tonne/year and their release is intended under standard conditions of use.

Nevertheless, some substances lie outside the domain of REACH or are not subject to registration as stated in Title I, article 2.

Probably the most groundbreaking modification introduced by REACH is the inversion of the burden of proof from regulators to industry, by imposing the concept of "no data, no market". Producers are required to know the composition of their products and to prove that the included substances do not pose risks to human health and the environment. To this end, registrants must submit to ECHA a technical dossier, and in case a chemical safety report (CSR), which includes information about the physical-chemical, toxicological, ecotoxicological and environmental fate properties of the substance, as specified in Annexes VII to X of REACH. In addition, registrants must provide the information required in Annex XIII for the classification as persistent, bioaccumulative or toxic (PBT) or very persistent and very bioaccumulative (vPvB).

An important mission that regulators pursued in the "making of" REACH is the will to avoid unnecessary testing and reduce animal tests. This is apparent in several parts of the framework of REACH. Indeed, a number of tools were implemented with this objective, including the push to share data, the creation of substance information exchange forums (SIEF, Title III) and the promotion of alternative testing strategies. Regarding the latter point, the text of REACH reports:

> *"The Commission, Member States, industry and other stakeholders should continue to contribute to the*

promotion of alternative test methods on an international and national level including computer supported methodologies, in vitro methodologies, as appropriate, those based on toxicogenomics, and other relevant methodologies."

The scenarios envisioned by REACH for alternative test methods generated interest in the scientific community. As expected, much effort was put into the development and optimisation of methodologies and protocols that could meet the applicability requirements demanded by regulators. In the field of computer-based methodologies, this is proved also by the large number of European funded projects aimed at developing mathematical models and designing informatics systems that can be used by registrants (CAESAR [Benfenati, (2010)] and OECD QSAR Toolbox [*The OECD QSAR Toolbox*, (2013)] just to name two). It is in this context that also the present study was initiated.

The relevance given to alternative testing strategies for the generation of new data to use in registration dossiers is further emphasised in Title II, article 13, where particular stress is given to the beneficial aspects in terms of animal welfare. It is in this paragraph that quantitative structure-activity relationships (QSARs) are explicitly mentioned.

1.3 QSAR: background and role in the regulatory context

The assumption that the structure of chemicals is related to their activities and properties constitutes the basis of a scientific field whose first steps can be traced back to the end of the XIX century [Meyer, (1899); Overton, (1901)]. This branch of science aims at identifying and rationalizing the functional relationships between chemical structure and observed properties. The final goal is to enhance the understanding of the chemical and biological phenomena under analysis and to allow the prediction of the behaviour of new systems without the need for experimental measurements.

The study of structure-activity relationships (SARs) and their quantification (quantitative structure-activity relationships, QSARs) owes much of its development to the research carried out by Corwin Hansch, Spencer Free, James Wilson and co-workers from the 1960s. Hansch equation related the potency of a biological effect with lipophilic, electronic and steric properties [Hansch et al., (1962)], whereas in the

1.3 QSAR: background and role in the regulatory context

Free-Wilson analysis [Free and Wilson, (1964)] the biological activity was directly related with structural features (substituents and their position).

Since the 1960s, QSAR analysis focused more and more on the development of theoretical variables that are not derived from experiments, the so called molecular descriptors [Todeschini and Consonni, (2009)]. The development of a solely theoretical and computerised description of the molecular structure and its properties (i.e. theoretical molecular descriptors) was partly made possible by progresses in informatics and graph theory [Bondy and Murty, (2008)], as well as increase in the power of computers. This development lead to the definition of a vast number of descriptors and was accompanied by a change in the scientific paradigm. In fact, instead of choosing *a priori* which descriptors to use, current QSAR analysis is often retrospective in that several descriptors are calculated and then the best subset is searched for. Here a new problem is introduced, i.e. the generation and evaluation of several different combinations of descriptors in a reasonable time. To face this issue, scientists needed to borrow and adapt techniques from other scientific fields, such as optimisation methods. Modern QSAR analysis takes advantage of a number of advanced mathematical and statistical methods and is strictly connected to other scientific fields like multivariate analysis, chemometrics and chemoinformatics. Figure 1.2 presents a simplified scheme of the main steps of a typical QSAR analysis.

As outlined in paragraph 1.2, REACH regulation promotes the use of QSAR models. In principle, the new data generated can be employed in different contexts, such as prioritisation of chemicals, design of experiments, mechanistic understanding and data gap filling. Particularly related to the last point, QSAR-generated data can be used as one-to-one replacement of experimental measurements or in a *weight of evidence* approach, depending on the confidence that is associated with the prediction. Since the overall goal of REACH is the enhancement of human health and environmental protection, regulators will accept predictions from QSARs on their own (i.e. as one-to-one replacement of measured data) only under conditions that guarantee (with a certain confidence) that the results are relevant, reliable and adequate. These conditions are enunciated in Annex XI of REACH and detailed in the most comprehensive guidance currently available for the application of QSARs within REACH, i.e. Chapter R.6 of the guidance on information requirements and chemical safety assessment [ECHA, (2008)]. Results of QSARs may be used instead of testing when:

- The model is scientifically valid.

- The model is applicable to the chemical of interest.

- The prediction (result) is relevant for the regulatory purpose.

- Appropriate documentation on the method and result is given.
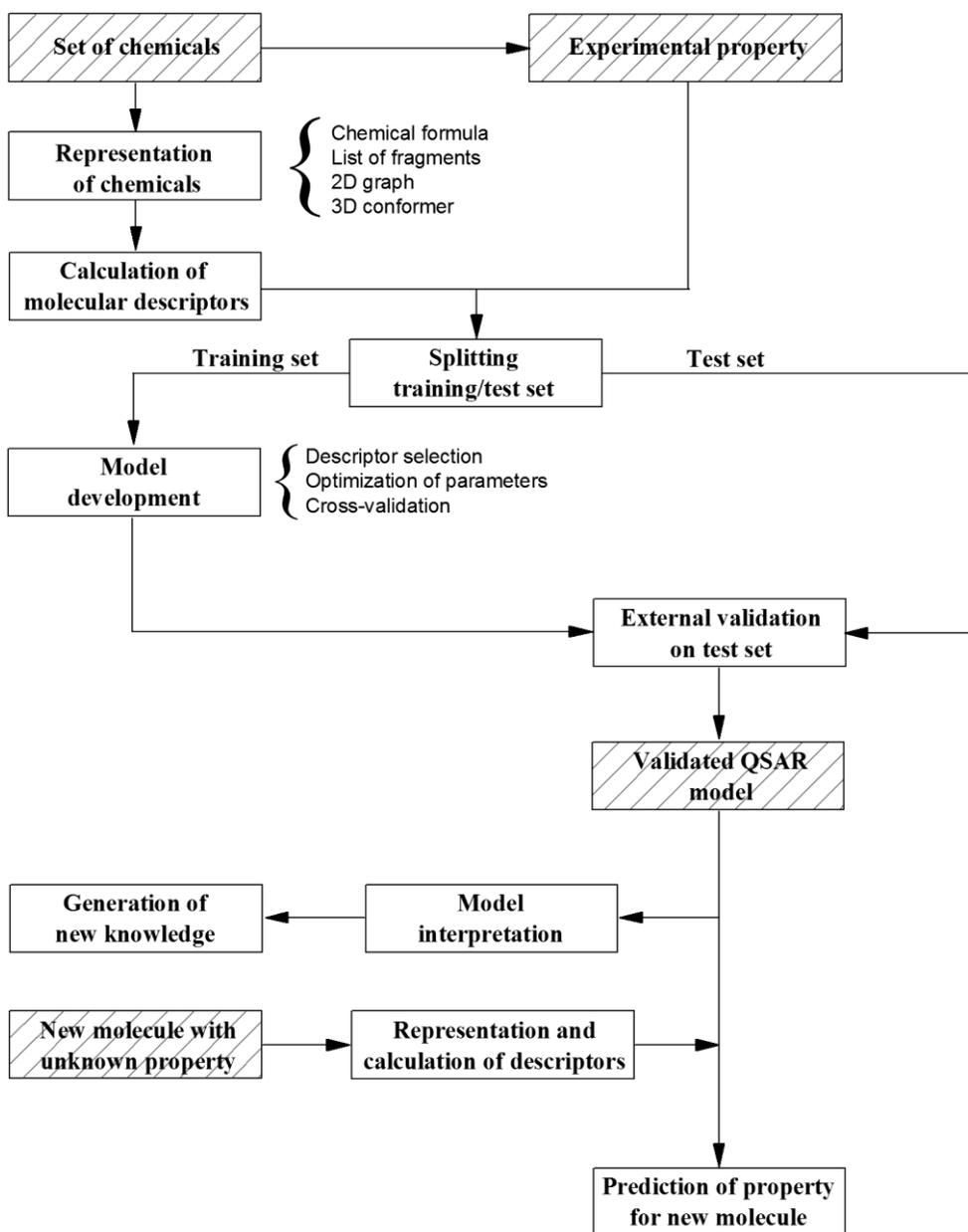
Figure 1.2. Sketch of the steps of a QSAR analysis. White boxes represent 'actions', dash-filled boxes represent 'objects' (set of data and model algorithm).

1.3 QSAR: background and role in the regulatory context

The first condition requires the QSAR model to have a proved validity from a scientific point of view. The validity for regulatory purposes stems from the fulfilment of five principles for validation that were defined by the Organization for Economic Co-operation and Development (OECD) [OECD, (2007)], which will be discussed below in this paragraph.

According to the second condition, the validity of a model is a necessary but not sufficient condition for its safe application: the chemical of interest must also fall inside the domain of applicability of the model, i.e. the chemical space where the model is assumed to provide accurate (reliable) predictions. From the fulfilment of the first two conditions stems the reliability of the prediction.

The third condition adds to the first two the issue regarding the regulatory relevance. It states that not only the chemical of interest should fall inside the applicability domain of a valid QSAR model, but the prediction should also be appropriate for the regulatory purpose. The fulfilment of this condition (on top of the first two) renders a QSAR-generated datum adequate.

Eventually, the last condition requires the information to demonstrate the adequacy of a QSAR prediction be communicated in a clear, complete and appropriate manner. To this end, two protocols were designed: the QSAR model reporting format (QMRF) and the QSAR prediction reporting format (QPRF).

The relationships between the first three conditions are depicted in Figure 1.3. It is shown that a prediction deriving from a valid QSAR model (i.e. meeting the five OECD principles) for which the molecule falls inside the applicability domain is a reliable result, but in order to have an adequate result the model should also be relevant for the particular regulatory purpose.

As aforementioned, five principles were enunciated by the OECD in order to define the scientific validity of QSAR models. The five principles state that

> *"To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:*
>
> *1) a defined endpoint;*
>
> *2) an unambiguous algorithm;*
>
> *3) a defined domain of applicability;*
>
> *4) appropriate measures of goodness-of–fit, robustness and predictivity;*
>
> *5) a mechanistic interpretation, if possible."*

The first two principles serve to enhance the confidence in the use (or acceptance) of a QSAR model. Principle one demands the endpoint for which the model gives predictions be well-defined, since different protocols or experimental conditions could be used to determine a certain endpoint. Principle two addresses the reproducibility of the results by demanding transparency in the model algorithm. The third principle tackles the issue related to the implicit limitations of a QSAR model that derive from the finite set of chemicals (the training set) used to develop the model itself. In other words, it is necessary to define a chemical domain based on the training molecules where the model is assumed to provide accurate (reliable) predictions. Principle four states that the model needs to have been subject to appropriate validation procedures in order to estimate not only its ability to fit the data in the training set, but also its accuracy in predicting properties for new suitable chemicals (test set). Eventually, principle five promotes the interpretation of model descriptors in relation to the property being addressed, since this can give additional confidence in the use (and acceptance) of model predictions.
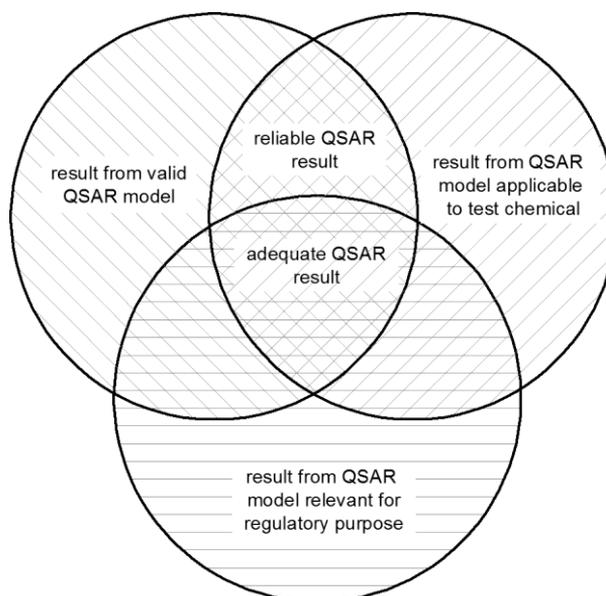
Figure 1.3. Inter-relationships among the first three conditions of REACH Annex XI for the regulatory use of QSAR models. Scheme adapted from Chapter R.6 of the guidance on information requirements and chemical safety assessment [ECHA, (2008)].

1.4 State of the art of QSAR in aquatic toxicity

Several QSAR studies were conducted on the aquatic toxicity of chemicals. Early studies generally targeted homogeneous sets of compounds, often a single chemical class or compounds presumed to share a common mode of action. Later, with the development of new methods, the progresses in informatics systems and the availability of new experimental data, research focused also on the analysis of large heterogeneous datasets, where several different chemical classes, implying different modes of action, were contemporarily represented.

A general scheme that encompasses aquatic toxicity in general, i.e. it is not related to a specific species, was defined by Verhaar et al., (1992). The scheme assigns chemicals to one of four identified toxicity categories, which are associated to estimated ranges of effective concentrations. Hence, the scheme can be used for preliminary screening or prioritisation for testing. The identified classes are outlined below.

1. **Inert chemicals** are not reactive and give narcosis-type toxicity. As aforementioned, the toxic potency of these chemicals depends only on their lipophilicity and is the minimum or baseline toxicity.

2. **Less inert chemicals** are also not reactive but their toxicity is slightly higher than that of inert chemicals. These chemicals are said to be "polar narcotics" because they feature a polar group, which has often been related to the higher toxic potency.

3. **Reactive chemicals** show an excess toxicity compared to the baseline set by narcosis. This broad definition actually comprises diverse classes of chemicals that can react with biological structures following different modes of action.

4. **Specifically acting chemicals** is a category that includes diverse chemical classes that are known to react with specific biological targets

QSAR models developed specifically for *Daphnia magna* and *Pimephales promelas* are discussed in the following two paragraphs. The discussion does not encompass all the models that were developed. Particular focus is given to results in regression obtained using large heterogeneous datasets for sake of comparison with the results obtained in this study. Commercial models are not discussed because information regarding dataset and algorithm is often not provided. This causes also

resistance from the regulators' side to the acceptance of the predictions obtained via these systems.

1.4.1 QSAR models for *Daphnia magna*

Acute toxicity towards *Daphnia magna* is commonly estimated by a quantitative parameter, usually the lethal or effective concentration for 50% organisms for a test duration of 24 or 48 hours (LC_{50} or EC_{50} 24 or 48 hours). In case of effective concentrations, immobilisation is the observed effect. Models aiming to classify chemicals for their mode of action [von der Ohe et al., (2005); Raevskii et al., (2008); Grigor'ev et al., (2014)] or their toxicity level [Roy and Das, (2013)] were also developed.

Regression analysis was employed to study both homogeneous sets of chemicals, defined as single chemical class or single mode of action, and heterogeneous datasets. Examples of some early and recent studies on single chemical classes are given in Table 1.1. Overall, good correlations were obtained based mainly on linear approaches, such as multiple linear regression (MLR) and partial least squares (PLS) regression. The reader can refer to the following manuscripts for models developed for single modes of action: Hermens et al., (1984); Urrestarazu Ramos et al., (1998); Zhao et al., (1998); and for single chemical classes: Vighi and Calamari, (1985); Devillers and Chambon, (1986); Deneer et al., (1989); Vighi et al., (1991); Tosato et al., (1993); Newsome et al., (1993); Todeschini et al., (1996); He and Wang, (1996); Chen et al., (1996); Wong et al., (1997); Zhu et al., (1999); Marchini et al., (1999); Dai and Wang, (2000); Cronin et al., (2000); Wei et al., (2001); Liu et al., (2003); Morrall et al., (2003); Davies et al., (2004); Padmanabhan et al., (2006); Boeije et al., (2006); Hodges et al., (2006); Zvinavashe et al., (2009); Song et al., (2011); Ismail Hossain et al., (2011); Furuhama et al., (2012); Roberts et al., (2013); Cassani et al., (2013a); Cassani et al., (2013b); Roy et al., (2014). A computer module implementing models for over 110 chemical classes was also developed by the U.S. Environmental Protection Agency as a part of the EPI Suite package [*EPI Suite*, (2012)]: the ECOSAR program [Mayo-Bean, (2012)].

The discussion will now focus on QSAR models developed from large heterogeneous datasets coherently with the intents of this study. Regression models calibrated on large heterogeneous datasets take advantage of a single model applicable to many chemical classes, in contrast to models defined for individual chemical classes. The latter ones, in fact, present potential drawbacks when applied

to multifunctional molecules. In these cases, in fact, the problem regards the choice of the proper model or the procedure to combine predictions from models specific for each different moiety in the molecule of interest.

When dealing with the increased structural diversity in heterogeneous datasets, both global and local approaches were used. Details regarding dataset, modelling method and performance of QSAR models developed from heterogeneous datasets are given in Table 1.2. Global models were mainly based on linear approaches, such as multiple linear regression (MLR), and partial least squares (PLS) regression. Non-linear methods, such as probabilistic neural networks (PNN), and the multi-linear splines were also used. Local approaches were used in order to overcome the differences between modes of action. The basic assumption was that chemicals with similar structure should act via a common mode of action. Local models were implemented both by grouping chemicals into clusters, for which *ad hoc* regression models were calibrated [Martin et al., (2012)], and by employing read-across approaches. Martin et al., (2012) developed two models where the prediction is derived from one or more clusters of similar molecules (SC+MLR and HC+MLR in Table 1.2). Read-across was based on the *k*-nearest neighbours (*k*NN) method with the introduction of a similarity-based assessment of the applicability domain [Martin et al., (2012); Kühne et al., (2013)]. Kühne et al., (2013), in particular, combined a global LogP-based regression to calculate the baseline toxicity with *k*NN to estimate the toxicity enhancement over the baseline. This model was integrated in a decision-tree that was able to provide either a quantitative or a qualitative estimate of the toxicity.

Two reasons can be envisioned for the higher statistics obtained on homogeneous datasets compared to those obtained on heterogeneous ones. On one hand, the homogeneity of structures often translated into similar mechanisms of toxicity and limited range of toxicity values, which enabled a "local" analysis to be carried out; on the other hand, the experimental measurements were often carried out in a single laboratory following common procedures, thus generating consistent data. In fact, the reproducibility of independent measurements is a major issue, since differences of several orders of magnitude were sometimes detected among data for the same chemical deriving from different laboratories (paragraph 2.2). Emblematic are in this regard the results obtained by Tao et al., (2002) who developed a fragment-based model from 217 chemicals. Their model reached R^2 and Q^2_{cv} values of 0.97 and a considerable contribution to the excellent statistics can be reasonably ascribed

to their data selection approach: only molecules who showed an experimental variability smaller than one order of magnitude between the minimum and maximum values were, in fact, retained in the dataset. If this is the case, the experimental variability can be said to have an extreme influence on the performance of QSAR models.

Table 1.1. Characteristics of literature models for acute toxicity towards Daphnia magna based on chemical classes. In case of multiple models, the range of the statistics is reported between square brackets. Hyphens are used for not definable information.

Reference	Chemical class	Endpoint[a]	No. mod.[b]	Method[c]	n train[d]	p[e]	R²
				Homogeneous datasets			
Vighi and Calamari, (1985)	organotin	EC50 24 h	14	MLR	[3-15]	[1-3]	[0.26-0.99]
Devillers and Chambon, (1986)	chlorophenols	IC50 24 h	15	MLR	17	[1-3]	[0.39-0.89]
Deneer et al., (1989)	nitroaromatics	IC50 48 h	3	MLR	[15-22]	[1-2]	[0.36-0.56]
Vighi et al., (1991)	organo-phosphorus	EC50 24 h	1	QMLR	22	6	0.90
Todeschini et al., (1996)	amines	EC50 24 h	1	MLR	8	4	1.00
	chlorobenzenes	EC50 24 h	1	MLR	6	3	1.00
	organotin	EC50 24 h	1	MLR	15	6	0.99
	organo-phosphorus	EC50 24 h	1	MLR	20	5	0.92
Zvinavashe et al., (2009)	organo-thiophosphate	EC50 24 h	3	MLR	10+5[f]	[1-2]	[0.80-0.82]+ [0.61-0.71][g]
Ismail Hossain et al., (2011)	ionic liquids	EC50 24 h	1	MLR	64	36	0.97
Cassani et al., (2013a)	triazoles and benzotriazoles	EC50 48 h	2	MLR	90	[5-6]	[0.79-0.79]
			2	PLS	90	[243-245]	[0.59-0.80]
			1	ASNN	90	132	0.70
			1	*consensus*	90	-	0.82
Cassani et al., (2013b)	triazoles and benzotriazoles	EC50 48 h	2	MLR	97[h]	[5-5]	[0.73-0.77]+ [0.68-0.83][g]

[a] endpoint as reported in the literature; h = hours; [b] number of developed models; [c] MLR = Multiple Linear Regression; QMLR = Multiple Linear Regression with quadratic terms; PLS = Partial Least Squares; ASNN = Associative Neural Networks; [d] number of compounds in training set; [e] number of model descriptors; [f] training set + test set; [g] fitting + external validation; [h] split into training-test sets (70%-30%) twice.

Table 1.2. Characteristics of literature models for acute toxicity towards Daphnia magna developed from heterogeneous datasets. In case of multiple models, the range of the statistics is reported between square brackets. Hyphens are used for lacking or not definable information.

Reference	Endpoint[a]	No. mod.[b]	Method[c]	Heterogeneous datasets					
				n train[d]	n test[e]	p[f]	R^2	Q^2_{cv}	Q^2_{ext}
Devillers et al., (1987)	IC_{50} 24 h	4	MLR	57	44	[1-4]	[0.70-0.89]	-	0.70
Todeschini et al., (1996)	EC_{50} 24 h	5	MLR	49	-	[3-7]	[0.68-0.82]	[0.64-0.71][g]	-
Faucon et al., (2001)	EC_{50} 48 h	1	MLR	61	35	2	0.54	0.49	0.57
Kaiser and Niculescu, (2001)	LC_{50} 24 h LC_{50} 48 h	4	PNN	700	76	57	[0.87-0.88]	-	[0.76-0.76]
Tao et al., (2002)	EC_{50} 48 h	1	MLR	217	-	122	0.97	0.97[h]	-
Toropov and Benfenati, (2006)	EC_{50} 96 h	1	MLR	220	42	1	0.78	-	0.74
Moosus and Maran, (2011)	LC_{50} 48 h	1	MLR	118	117	4	0.74	0.74[i]	0.56
Toropova et al., (2012a)	LC_{50} 48 h	1	MLR	114+108[j]	75	1	0.71	0.72[k]	0.78
Toropova et al., (2012b)	LC_{50} 48 h	4	MLR	[107-152]+ [57-115][j]	[75-89]	1	[0.65-0.73]	[0.80-0.88][k]	[0.75-0.80]
Katritzky et al., (2009)	EC_{50} 48 h	4	MLR	[86-130]	[0-44]	[5-5]	[0.70-0.78]	[0.64-0.75]	[0.54-0.74]
Kar and Roy, (2010)	LC_{50} 48 h	6	MLR	222	75	[5-12]	[0.66-0.74]	[0.64-0.70][l]	[0.59-0.72]
		3	MLR-Spline	222	75	[5-7]	[0.70-0.71]	[0.67-0.69][l]	[0.63-0.67]
		6	PLS	222	75	[5-8]	[0.63-0.70]	[0.61-0.68][l]	[0.61-0.74]
		3	PLS-Spline	222	75	[4-9]	[0.65-0.66]	[0.61-0.64][l]	[0.61-0.63]
		2	LR+kNN	1365	-	-	[0.75-0.75]	[0.75-0.75][l]	-
Kühne et al., (2013)	LC_{50} 48 h	1	LR+kNN (0.75)	1365 (972)[m]	-	-	0.82	0.82[l]	-
		1	LR+kNN (0.87)	1365 (524)[m]	-	-	0.86	0.86[l]	-
		2	LR+kNN (tree)	1365 ([757-759])[m]	-	-	[0.82-0.85]	[0.82-0.84][l]	-
Martin et al., (2012)	LC_{50} 48 h	1	HC+MLR	283	70	$\leq n_k/5$[n]	-	-	0.70
		1	SC+MLR	283	70	-	-	-	0.56
		2	MLR	283	70	-	-	-	[0.67-0.70]
		1	kNN	283	70	-	-	-	0.73
				283	70	-	-	-	0.74
		1	*Consensus*	283	333(333)[m,o]	-	-	-	0.46[o]

Table 1.2. Continued

Reference	Endpoint[a]	No. mod.[b]	Method[c]	n train[d]	n test[e]	p[f]	R²	Q²cv	Q²ext
VEGA, EPA[p]	LC$_{50}$ 48 h	1	MLR	269	68(15)[m] 374(135)[m,o]	17	0.71	-	0.75 0.54[o]
VEGA, DEMETRA[q]	LC$_{50}$ 48 h	1	neuro-fuzzy	220(105)[m]	43(18)[m] 220(31)[m,o] 135[r]	16	0.93	-	0.97 0.14[o] 0.63[r]

[a] endpoint as reported in the literature; [b] number of developed models; [c] MLR = Multiple Linear Regression; PNN = Probabilistic Neural Networks; PLS = Partial Least Squares; LR+kNN = Linear regression coupled with k-Nearest Neighbours; LR+kNN (0.75) = Linear regression coupled with k-Nearest Neighbours with similarity threshold = 0.75; LR+kNN (0.87) = Linear regression coupled with k-Nearest Neighbours with similarity threshold = 0.87; LR+kNN (tree) = decision tree based on Linear regression coupled with k-Nearest Neighbours with both similarity thresholds; HC+MLR = Hierarchical Clustering coupled with Multiple Linear Regression; SC+MLR = Single Clustering coupled with Multiple Linear Regression; kNN = k-Nearest Neighbours; neuro-fuzzy = hybrid system that combines one PLS and two NN models; [d] number of compounds in training set; [e] number of compounds in the test set; [f] number of model descriptors; [g] leave-20%-out; [h] leave-20 molecules-out repeated 30 times; [i] iterated leave-30%-out; [j] training set + validation set; [k] Q² on the validation set; [l] leave-one-out; [m] number of compounds inside the Applicability Domain; [n] maximum allowed number of descriptors in each cluster model (n_k = number of compounds in the cluster); [o] validation reported by Golbamaki et al., (2014); [p] re-implementation in VEGA of the MLR model of Martin et al., (2012); [q] re-implementation in VEGA of the DEMETRA model (specific for pesticides); [r] validation reported by Porcelli et al., (2008)

1.4.2 QSAR models for *Pimephales promelas*

An invaluable work that allowed much of the QSAR analysis undertaken on the toxicity towards the *Pimepahles promelas* was the creation of a consistent database of experimental data for about six hundred chemicals with assigned modes of action (MoAs), namely the MED-Duluth fathead minnow database [Russom et al., (1997)].

As the case of *D. magna*, some studies aimed at predicting the mode of action of chemicals [Russom et al., (1997); Michielan et al., (2010); Casalegno and Sello, (2013); Nendza et al., (2014); Ren, (2002); Ren and Schultz, (2002); Lozano et al., (2010)]. The rationale is that reliable QSARs can only be derived from sets of chemicals acting via the same mechanism. Hence, the possibility to assign chemicals to the correct cluster (or class) is the primal objective, prior to developing MoA-based QSARs (unless the MoA is experimentally determined). Based on these considerations, regression models to estimate the acute toxicity were calibrated for single chemical classes or single MoAs. Examples of early QSAR studies based on chemical classes or modes of action are given in Table 1.3. QSAR models developed from such homogeneous datasets were based on linear regression employing a small number of descriptors. Regression statistics were in general high in fitting, presumably because chemicals shared a common mode of action. The reader can refer to the following manuscripts for MoA-based QSARS: Veith et al., (1983); Veith and Broderius, (1987); Nendza and Russom, (1991); Karabunarliev et al., (1996a); Bearden and Schultz, (1997); Zhao et al., (1998); Gunatilleka and Poole, (1999); Öberg, (2004); Papa et al., (2005); Yuan et al., (2007); Qin et al., (2010); Lozano et al., (2010); and chemical class based ones: Basak and Magnuson, (1983); Basak et al., (1984); Hall et al., (1984); Nendza and Russom, (1991); Newsome et al., (1991); Karabunarliev et al., (1996b); Mount et al., (1997); Wong et al., (1997); Parkerton and Konkel, (2000); Freidig and Hermens, (2000); Cui et al., (2008).

However, the classification of chemicals to known MoAs is not an easy task and is made even more complicated by the difficulties in the determination of the MoA itself. Consequently, many investigations were also addressed to quantitatively model large heterogeneous datasets as a whole. The parameter being modelled was typically the lethal concentration for 50% test fish for a test duration of 96 hours (LC$_{50}$ 96 hours). The discussion will focus on this case.

Modelling the toxicity of large heterogeneous datasets presented additional challenges due to the increased structural diversity, which implied the concomitant

presence of different modes of action (MoAs). The majority of literature models were still based on global strategies but often taking advantage of non-linear methods. Regarding linear methods, multiple linear regression (MLR) was still the preferred choice for many investigations, whereas partial least squares (PLS) regression was seldom used. A multi-linear method, i.e. spline, was also employed. Neural networks (NN) were often used to model such heterogeneous datasets. Interestingly, only one QSAR model based on NN was developed for the toxicity towards *D. magna*. Another non-linear method that lately received much attention in other scientific fields, namely support vector regression (SVR), was used to study the toxicity towards the *P. promelas*. Additionally, some models explicitly or implicitly included also the information regarding the mode of action by implementing a preliminary classification step that assigned chemicals to more homogeneous categories (not necessarily corresponding to known modes of action), for which local regressions were calibrated. In et al., (2012) divided chemicals into narcotics and reactive chemicals and developed local linear regression models. Gini et al., (2004) clustered chemicals into nine groups using self-organizing neural networks and then calibrated regression models for each cluster using feedforward neural networks. The hierarchical clustering model in T.E.S.T. software averaged the predictions obtained from several linear models developed for a number of clusters, whereas the single cluster model (called FDA method) used a linear equation fitted on a single cluster of 30-75 similar chemicals [Martin et al., (2012)]. Colombo et al., (2008) employed

Table 1.3. Characteristics of literature models for LC$_{50}$ 96 hours towards Pimephales promelas based on chemical classes or modes of action. In case of multiple models, the range of the statistics is reported between square brackets.

Homogeneous datasets					
Reference	**Chemical Class / MoA**	**No. mod.**[a]	***n* train**[b]	***p***[c]	**R^2**
Basak and Magnuson, (1983)	alcohols	2	10	1	[0.98-0.99]
Basak et al., (1984)	esters	25	15	[1-3]	[0.023-0.96]
Veith and Broderius, (1987)	polar narcotics	1	39	1	0.90
Newsome et al., (1991)	amines	30	[7-41]	[1-2]	[0.03-0.98]
Nendza and Russom, (1991)	non-polar narcotics	1	147	1	0.85
	polar aliphatic	1	33	1	0.96
	polar aromatic	1	118	1	0.83

[a] number of developed models; [b] number of compounds in training set; [c] number of model descriptors.

Table 1.4. Characteristics of literature models for LC_{50} 96 hours towards Pimephales promelas developed from large heterogeneous datasets. In case of multiple models, the range of the statistics is reported between square brackets. Hyphens are used for lacking or not definable information.

| Reference | No. mod.[a] | Method[b] | Heterogeneous datasets | | | | | |
			n train[c]	n test[d]	p[e]	R^2	Q^2_{cv}	Q^2_{ext}
Netzeva et al., (2005)	62	MLR	[560-568]	-	[1-4]	[0.61-0.70]	[0.61-0.69][f]	-
	2	PLS	562	-	13	[0.72-0.73]	[0.71-0.72]	-
Pavan et al., (2006)	1	MLR	408	57	4	0.80	0.80[g]	0.72
Roy and Das, (2012)	6	MLR	344	115	[6-10]	[0.76-0.79]	[0.75-0.76][f]	[0.75-0.79]
	3	MLR-Spline	344	115	[5-7]	[0.76-0.78]	[0.75-0.77][f]	[0.78-0.78]
	1	MLR	445	110	5	0.71	-	0.55
In et al., (2012)	1	ANN	334+111[h]	110	5	0.80	0.62[i]	0.62
	2	RP-MLR	445	110	[5-5][j]	[0.75-0.76]	-	[0.60-0.63]
	8	Consensus	445	110	-	[0.78-0.80]	-	[0.63,0.67]
Gini et al., (2004)	1	NN+NN	454	114	156	-	-	0.76
Wang et al., (2010)	1	SVR	457	114	8	0.83	-	0.80
Schüürmann et al., (2011)	2	LR+kNN	692	-	-	[0.73-0.73]	[0.72-0.72][f]	-
	2	LR+kNN (0.8)	692 (419)[k]	-	-	[0.78-0.78]	[0.76-0.78][f]	-
	2	LR+kNN (0.9)	692 (230)[k]	-	-	[0.87-0.87]	[0.87-0.87][f]	-
Mazzatorta et al., (2003a)	11	Fuzzy NN	392	170	9	[0.20-0.70]	-	[0.00-0.50]
Mazzatorta et al., (2003b)	1	CPNN	275	274	150	0.97	-	0.59
	1	HC+MLR	659	164	$\leq n_k/5$[l]	-	-	0.71
	1	SC+MLR	659	164	-	-	-	0.63
Martin et al., (2012)	2	MLR	659	164	-	-	-	[0.69-0.70]
	1	kNN	659	164	-	-	-	0.67
	1	Consensus	659	164	-	-	-	0.73
Devillers, (2005)	2	MLR+NN	484	85	[10-23]	[0.82-0.75][m]	-	[0.64-0.66][m]
	3	MLR	607[n]	-	147	[0.83-0.87]	[0.46-0.54][i]	-
Casalegno et al., (2005)	3	PLS	607[n]	-	147	[0.81-0.83]	[0.59-0.67][i]	-
	3	NN	607[n]	-	147	[0.89-0.92]	[0.62-0.70][i]	-
	3	PMM	607[n]	-	147	[0.79-0.82]	[0.48-0.60][i]	-
Colombo et al., (2008)	1	Tree + MLR	560	-	[3-5][j]	[0.83-0.99][o]	[0.81-0.98][f, o]	-
Eldred et al., (1999)	1	MLR	287	88[p]	8	0.83[m]	0.90[m]	0.75[m]
	2	NN	287	88[p]	[8-8]	[0.81-0.71][m]	[0.88-0.77][i,m]	[0.84-0.74][m]

Table 1.4. Continued

Reference	No. mod.[a]	Method[b]	n train[c]	n test[d]	p[e]	R²	Q²cv	Q²ext
Hewitt et al., (2007)	1	MLR	484	121	2	0.71	0.70	0.61
	2	*Consensus*	484	121	-	NR	[0.71-0.71]	[0.59-0.60]
Klopman et al., (2000)	1	MLR+ L-MLR	675	-	[2-8]j	0.88	0.95q	-
Lozano et al., (2010)	10	MLR	557	-	[4-17]	[0.62-0.73]	[0.50-0.70]r	-
	1	*Consensus*	557	201+ 144s	-	0.71	-	0.60+0.58t
Maran et al., (2007)	1	MLR	373	188	6	0.73	0.72f	0.66
	1	BPNN	373	188	6	0.78	-	0.73
Nendza and Russom, (1991)	1	MLR	532	-	2	0.61	-	-
Niculescu et al., (2004)	2	PNN	800	86	76	[0.89-0.99]	-	[0.78-0.52]
Papa et al., (2005)	2	MLR	249	200	[5-6]	[0.79-0.81]	[0.78-0.80]g	[0.71-0.72]
Toropova et al., (2012c)	3	MLR	[246-271]+ [144-164]h	[148-158]	[1-1]	[0.67-0.68]	[0.79-0.85]i	[0.77-0.79]
VEGA, EPA[u]	1	MLR	652	164 (39)k	21	0.69	-	0.69

[a] number of developed models; [b] MLR = Multiple Linear Regression; PLS = Partial Least Squares; ANN = Artificial Neural Network; RP-MLR = Recursive Partitioning coupled with MLR for each class; NN+NN = Clustering by means of self-organised Neural Network coupled with local regression by means of feedforward Neural Networks; SVR = Support Vector Regression; LR+kNN = Linear regression coupled with k-Nearest Neighbours; LR+kNN (0.8) = Linear regression coupled with k-Nearest Neighbours with similarity threshold = 0.8; LR+kNN (0.9) = Linear regression coupled with k-Nearest Neighbours with similarity threshold = 0.9; Fuzzy NN = fuzzy Neural Network; CPNN = CounterPropagation Neural Network; HC+MLR = Hierarchical Clustering coupled with Multiple Linear Regression; SC+MLR = Single Clustering coupled with Multiple Linear Regression; kNN = k-Nearest Neighbours; MLR+NN = Multiple Linear Regression for baseline coupled with Neural Networks to model the residuals; PMM = Powell's Minimization Method; Tree + MLR = decision tree to partition chemicals into 9 clusters coupled with MLR for each cluster; MLR+L-MLR = Multiple Linear Regression for baseline coupled with local MLR for clusters of molecules sharing a common toxicophore; BPNN = BackPropagation Neural Network; PNN = Probabilistic Neural Networks; [c] number of compounds in training set; [d] number of compounds in the test set; [e] number of model descriptors; [f] leave-one-out cross-validation; [g] bootstrap with 5000 iterations; [h] training set + validation set; [j] Number of descriptors in each MLR; [k] number of compounds inside the Applicability Domain; [l] maximum allowed number of descriptors in each cluster model (n_k = number of compounds in the cluster); [m] Root Mean Square Residuals; [n] split in training-validation sets (2:1) 3 times; [o] range of statistics for the 9 cluster models (overall statistics not reported); [p] divided into validation and test sets; [q] leave-10%-out cross-validation, repeated 3 times; [r] 3-fold cross-validation; [s] number of molecules in each of the two test sets; [t] results on each of the two test sets; [u] re-implementation in VEGA of the MLR model of Martin et al., (2012).

a scheme to cluster chemicals into nine groups and developed MLR models for each cluster. Klopman *et al.*, (2000) used LogP to calculate the baseline toxicity and calibrated local MLR models for clusters of chemicals sharing a common toxicophore fragment. Local models were also developed by using the read-across approach, consisting in the determination, for each molecule, of a local neighbourhood, which is used to estimate the property. Read-across was implemented by means of the *k*-nearest neighbours (*k*NN) method and used in combination with a similarity-based assessment of the applicability domain [Martin et al., (2012); Schüürmann et al., (2011)]. Schüürmann et al., (2011) designed a model that combined a LogP-based linear equation to estimate the baseline toxicity and read-across to assess the toxicity enhancement over the baseline.

As it could be expected, the statistics of these models were lower compared to those of models developed for specific chemical classes or modes of action (Table 1.3). However, satisfactory performance was obtained for both internal and external validation. Table 1.4 gives details about the aforementioned models regarding both the dataset, the modelling method and the performance.

1.5 Is there need for new models?

As one can see in the previous paragraphs, several QSAR models were developed to predict the acute toxicity of chemicals towards *Daphnia magna* and *Pimephales promelas* and it would therefore be reasonable to ask whether (and why) there is need for new models. The answer lies in the applicability of existing models for regulatory purposes within REACH, which largely depends on the fulfilment of the five OECD principles.

By analysing the level of compliance of existing models with the OECD principles, it can be said that the endpoint was usually well defined. Regarding principle two, the algorithm was sometimes not fully described in all the mathematical details. Moreover, it is reasonable to assume that the simpler and the more transparent the algorithm, the more confidence the user and the regulators will have in its application. Therefore, the use of complex methods, such as neural networks and support vector regression, should be limited to cases where they outperform simpler methods. The assessment of the applicability domain is perhaps the weakest point in several cases, since only few QSAR models implemented an approach to address this point. Regarding the validation of QSAR models, even

though there is no general agreement on which specific procedure to adopt, OECD principle four recommends the use of internal validation to estimate the goodness-of-fit and robustness, and external validation to assess the predictivity. It is also intuitive that higher confidence be given to models able to withstand harsh validation conditions, as these should provide a realistic estimate of the performance of the model. In this regard, some models where internally validated by means of mild conditions, such as leave-one-out cross-validation, and/or were not tested on an external set of chemicals. Even though the last OECD principle, provision of a mechanistic interpretation, is not a strict requirement, some investigations did not attempt to provide an interpretation of the observed correlation between molecular descriptors and toxicity. It is obvious that some of the aforementioned models were not specifically developed to comply with the OECD principles, being the focus on the development or application of new modelling algorithms or approaches. Consequently, these models could face difficulties if seeking regulatory application.

As hinted in the preface, the present work intends to develop QSAR models that can be applied in the framework of REACH. Consequently, the entire model development was carried out with the aim to comply with the requirements of the five OECD principles to the full extent in order to reduce potential limitations in real-life use. Treatment of molecular structures and measured toxicity values was undertaken in order to define consistent datasets, as much as allowed by the need for large heterogeneous datasets. Simple modelling algorithms were preferred in order to assure confidence in the equations. The assessment of the applicability domain was investigated by means of different approaches. The models were thoroughly validated by means of both internal and external procedures in order to assure proper estimate of stability and predictivity. Eventually, an interpretation of model descriptors in relation to current knowledge of aquatic toxicity was carried out

CHAPTER 2

Data

'Garbage In, Garbage Out'

Syndicated newspaper article of the U.S. Internal
Revenue Service, 1 April 1963.

*'On two occasions I have been asked, "Pray, Mr.
Babbage, if you put into the machine wrong
figures, will the right answers come out?" ... I am
not able rightly to apprehend the kind of confusion
of ideas that could provoke such a question.'*

Charles Babbage, Passages from the Life of a
Philosopher, 1864.

2.1 Acute lethal toxicity tests

Toxicity tests are usually performed in standardised conditions in order to reduce the variability. Different organizations have developed protocols and guidelines for acute toxicity testing, such as the Organization for Economic Co-operation and Development (OECD), the American Society for Testing and Materials (ASTM), the International Organization for Standardization (ISO), Environment Canada and the U.S. Environmental Protection Agency (U.S. EPA) [Rand, (1995)]. Nevertheless, the variability intra- and especially inter-laboratories cannot be eliminated and affects the results.

Aquatic acute toxicity tests are performed by exposing test organisms to the toxicant and observing their behaviour for a certain duration, usually at predefined time-points. Organisms are exposed to different concentrations of the chemical with the aim to obtain a concentration-response curve, from which the endpoints (e.g. LC_{50}, EC_{50}, NOEC, LOEC, etc.) can be calculated by means of statistical analysis. Figure 2.1 illustrates an example of a concentration-response curve. This study focuses on lethality tests, in which the observed effect is mortality and the corresponding calculated parameter is the concentration that kills 50% of test

organisms (LC$_{50}$) after 48 hours for *Daphnia magna* and 96 hours for *Pimephales promelas*.

Guidelines for acute toxicity tests with daphnids are the Canadian guideline described in the report EPS 1/RM/11 [Environment Canada, (1996)] for lethality testing and the OECD TG 202 [OECD, (2004)], the ASTM E729 [ASTM, (2007)], the ISO 6341:2012 [ISO, (2012)] and the U.S. EPA OPPTS 850.1010 [US EPA, (1996a)] for immobilisation tests. Guidelines for conducting acute lethality tests with fish are the OECD TG 203 [OECD, (1992)], the ASTM E729 [ASTM, (2007)], the U.S. EPA OPPTS 850.1075 [US EPA, (1996b)], the ISO 7346:1996 [ISO, (2010)] and the Canadian guideline described in the report EPS 1/RM/22 [Environment Canada, (2011)].
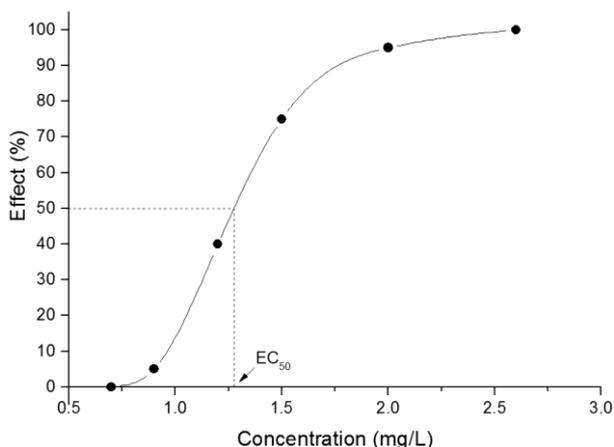
Figure 2.1. Example of concentration-response curve and corresponding EC$_{50}$ value.

2.2 Data quality

The quality of the experimental data used to develop QSAR models has an essential role. Good quality data are a promising starting point. Unfortunately, the reproducibility of experimental measurements intra- and inter-laboratories is a serious problem for acute aquatic toxicity endpoints. The often encountered large variability among measured LC$_{50}$ values depends on several factors, such as test conditions (e.g. water pH and temperature), test design (flow-through, static renewal and static) and organism [Golbamaki et al., (2014)].

In regard of acute toxicity testing with daphnids, it was shown that significantly different data can be obtained using genetically different clones of *Daphnia magna*,

even under the same experimental conditions [Picado et al., (2007)]. This caused different laboratories to produce data that were clone-dependent. Additionally, the criteria for death seem not univocal. Obviously, there is agreement on the lack of movement, also after gentle agitation of the vessel, but some studies and guidelines mention also the absence of heartbeat as additional criterion [Environment Canada, (1996); Katritzky et al., (2009); Yang et al., (2014)]. Since the observation of the heartbeat requires appropriate equipment (e.g. a dissecting microscope) and given that with some narcotic chemicals the heartbeat of *Daphniae* can slow down to 1-2 beats/min, it is not always straightforward to distinguish between dead and immobile daphnids [Environment Canada, (1996)]. Due to this reason, sometimes it is preferred to record the immobilisation, where a daphnid is considered immobile if it does not swim, also after gentle prodding, with the exclusion of minor movements of the appendages [Rand, (1995)]. Therefore, there is some overlap between these two endpoints, because dead daphnids are counted in immobilisation tests and some immobile *Daphniae* might be erroneously considered dead. It is intuitive that this 'confusion' can be an additional source of variability in the data.

An estimation of the variability of the data can be found in a number of studies: Picado *et al.*, (2007) found differences of one order of magnitude for effluents only due to the genetic variability among populations of daphnids. Golbamaki et al., (2014) reported that 23% of their data (480 chemicals) had a variability within 10-fold range and 4.6% up to 100-fold. Barron et al., (2012) noticed extreme variability for carbaryl (2992-fold range) and benzene (62-fold). Regarding the data analysed in this study, the highest variability was observed for pentachlorophenol, whose LC_{50} values ranged from 3.01 to 6.85 [–Log(mol/L)]. Ranges for the ten chemicals associated with the largest variability are reported in Table 2.1. The pooled standard deviation over the entire dataset calculated from the molecules having multiple experimental values was equal to 0.368 [–Log(mol/L)].

For several fish species differences in the response to contaminants exposure were detected as well [Maes et al., (2005); Wedekind et al., (2007); Coe et al., (2009)]. The variations in responses to toxicants detected inter-laboratories were ascribed to the genetic differences among laboratory populations. In fact, laboratory populations generally showed reduced genetic variability compared to wildlife. Ankley and Villeneuve, (2006) discussed the use of *Pimephales promelas* in aquatic toxicology and concluded that higher quality data would be obtained by a stricter control of the diet and genetic composition of stocks of fish used for testing. Lethality

tests with fish do not seem affected by the death/immobilisation issue mentioned for daphnids. Death of fish is commonly described as the cessation of all visible signs of movement, especially absence of respiratory movements.

Regarding estimates of the experimental variability, Barron et al., (2012) reported differences within a range up to 7.5-fold for Endrin. In this study, the largest variability was that of 4-(dimethylamino)-3-methylphenyl N-methylcarbamate (Aminocarb), whose LC_{50} values ranged from 4.39 to 6.44 [-Log(mol/L)]. Ranges for the ten chemicals associated with the largest variability are reported in Table 2.2. The pooled standard deviation over the entire dataset calculated from the molecules having multiple experimental values was equal to 0.234 [–Log(mol/L)]. The variability within fish seems lower compared to that of daphnids. Some factors may contribute to this fact, such as no confusion among observed endpoints and natural differences in species sensitivities. Indeed, Raimondo et al., (2008) analysed species sensitivity distributions (SSDs) for 68 chemicals and 291 species and noticed that crustaceans were the most sensitive taxa.

Based on these considerations, it would seem reasonable (and advisable) to use experimental data produced using the same test conditions, design and strain of organism, since this would contribute to the elimination, or at least the reduction, of sources of variation other than the molecular structure. The problem here lies in the number of data having such characteristics that contrasts with the intent of this study, which is to develop QSAR models from large heterogeneous datasets. Hence, a dual problem had to be faced: on one side, the need for consistency demanded a selection of the data; on the other side, the desire to expand the dataset implied the employment of all the available data. The latter aspect was prioritised, aware that the experimental variability would have affected the performance of the models. Nevertheless, actions were devised in order to detect and remove at least erroneous experimental data and ambiguous molecular structures.

The strategy adopted for the preparation of the toxicity data towards *Daphnia magna* and *Pimephales promelas* presented one substantial difference related to the treatment of salts and mixtures. Therefore, the data preparation is described separately for the two organisms in paragraphs 2.3 and 2.4.

Table 2.1. Ten chemicals with the largest ranges of experimental LC$_{50}$ values towards Daphnia magna.

CAS-RN	Name	Max. LC$_{50}$		Min. LC$_{50}$	
		-Log(mol/L)	mg/L	-Log(mol/L)	mg/L
124-18-5	Decane	6.70	0.028	3.90	18.00
85-68-7	Butyl benzyl phthalate	5.24	1.80	3.53	92.00
28249-77-6	Thiobencarb	6.43	0.097	4.67	5.51
63-25-2	Carbaryl	7.51	6.3E-3	4.57	5.40
52-68-6	Trichlorfon	9.09	2.1E-4	6.31	0.13
122-14-5	Fenitrothion	8.58	7.3E-4	6.68	0.058
87-86-5	Pentachlorophenol	6.85	0.038	3.01	260.00
91465-08-6	Cyhalothrin	9.06	3.9E-4	7.22	0.027
95-76-1	3,4-dichloroaniline	6.21	0.10	4.10	13.00
206-44-0	Fluoranthene	6.28	0.11	2.80	320.00

Table 2.2. Ten chemicals with the largest ranges of experimental LC$_{50}$ values towards Pimephales promelas.

CAS-RN	Name	Max. LC$_{50}$		Min. LC$_{50}$	
		-Log(mol/L)	mg/L	-Log(mol/L)	mg/L
1071-83-6	Glyphosate	4.87	2.30	3.24	97.00
111-42-2	Diethanolamine	1.89	1.37E3	0.35	4.71E4
122-34-9	Simazine	4.61	5.00	2.60	510.00
13071-79-9	Terbufos	7.34	0.013	5.87	0.39
1740-19-8	Dehydroabietic acid	5.30	1.50	3.82	45.50
2032-59-9	Aminocarb	6.44	0.075	4.39	8.50
51630-58-1/ 66230-04-4[a]	Fenvalerate/ Esfenvalerate	9.32	2.0E-4	7.89	5.4E-3
83-79-4	Rotenone	7.93	4.6E-3	6.44	0.14
86-50-0	Azinphos-methyl	6.93	0.037	4.99	3.26
87-86-5/ 131-52-2[b]	Pentachlorophenol / Pentachlorophenol Na salt	7.16	0.020	5.40	1.07

[a] data for two isomers merged; [b] data of sodium salt merged together with discrete molecule.

2.3 *Daphnia magna* dataset

2.3.1 Data for model development

The endpoint considered for the development of QSAR models was the concentration causing death in 50% test organisms (LC$_{50}$) after a test duration of 48 hours. The experimental values used in this study were retrieved from three databases, namely ECOTOX [US EPA, (ECOTOX)], EAT5 [ECETOC, (2003)] and OASIS, and available scientific publications [Bernot et al., (2009); Randall et al., (1979); Sanderson and Thomsen, (2009); Jemec et al., (2007); Zou and Fingerman, (1997); Costanzo et al., (2007); Staples and Davis, (2002); Martins et al., (2007); von der Ohe et al., (2005); Williams et al., (2011); Nørgaard and Cedergreen, (2010); di

2.3 Daphnia magna dataset

Delupis et al., (1992); Ferrari et al., (2004); Foit et al., (2012); Ochoa-Acuña et al., (2009); Horn et al., (2004); Kyriakopoulou et al., (2009)]. The OASIS database was downloaded from the OECD QSAR Toolbox [*The OECD QSAR Toolbox*, (2010)]. In the EAT5 database, LC_{50} data were reported as EC_{50} (effective concentration) with lethality as the observed effect.

As aforementioned, priority was given to the definition of a large heterogeneous dataset, at the expense of the homogeneity of the experimental measurements. Therefore, no restriction was applied to the test conditions, design or laboratory. It is noteworthy that this information was not available for all the data. Nevertheless, data were screened with respect to both the molecular structures and the experimental toxicity values in order to disregard erroneous and ambiguous records. The treatment of the data was based on the following actions:

- Removal of records from the ECOTOX database indicating ranges or thresholds of experimental values.
- Check of correspondence between CAS registry numbers and chemical names by means of queries to the ChemSpider database [Royal Society of Chemistry] and Chemical Identifier Resolver (CIR) at NCI/NIH [NCI/CADD Group].
- Manual check of all records that showed mismatches with the additional support of the Sigma-Aldrich [Sigma-Aldrich Co.] and PubChem [Bolton et al., (2008)] databases. All the mismatches not resolved were deleted.
- Removal of disconnected structures (salts and mixtures).
- Removal of inorganic compounds.
- Deletion of toxicity data expressed as "%", "% v/v".
- Identification of duplicates of the same measurement. Only one value was retained.
- Deletion of information regarding stereochemistry.
- Calculation of the median toxicity for molecules having multiple experimental values.
- Derivation of an alert for inconsistent experimental data to be used for molecules having multiple toxicity values. The pooled standard deviation over the entire dataset was used to derive the alert.
- Check in the original publications of the experimental values for all the molecules associated with standard deviations larger than the alert. If the

original study was not accessible or not found, the corresponding value was removed.

Further details regarding these steps are reported in the scientific publication in Appendix I, which describes the model calibrated on this dataset. The final dataset comprised 546 organic molecules (Table 2.3) and can be downloaded at:

http://michem.disat.unimib.it/chm/download/toxicity.htm.

This dataset was used for the development of QSAR models with LC_{50} values expressed as negative logarithm of molarity (-Log(mol/L)). For sake of clarity, this dataset will be hereinafter referred to as the MICHEM dataset.

2.3.2 Additional data for model validation and extension

In addition to the data presented in the previous paragraph (MICHEM dataset), two new sets of data were later gathered and used to validate the model calibrated on the MICHEM dataset. In particular, data for 1360 compounds were obtained from the model implemented in the ChemProp software [Kühne et al., (2013); *ChemProp*, (2013)] and for 388 chemicals from the QSAR group at the Technical University of Denmark [Niemelä et al., (2010)]. The latter dataset is also available in the OECD QSAR Toolbox [*The OECD QSAR Toolbox*, (2013)]. These two sets of data will be referred to as ChemProp dataset and DTU dataset, respectively.

In order to use these data to further validate the model calibrated on MICHEM dataset, it was first necessary to investigate the overlaps among the three sets of data, namely ChemProp, DTU and MICHEM. Twelve chemicals from the DTU set were removed because their identity was not clearly defined. The remaining 376 compounds in the DTU set were all external to MICHEM dataset, since MICHEM and DTU sets are not overlapping, as shown in Figure 2.2. The procedure used to check the overlaps and define the validation subsets was as follows:

- Merge DTU, ChemProp and MICHEM sets.
- Check for the presence of records with the same CAS registry number and/or structure from different source sets.
- Remove records that had duplicates in terms of either CAS-RN or structure, thus indicating mismatches between CAS-RNs and structures in different source sets.
- Calculate average LC_{50} value for molecules present in both DTU and ChemProp sets.

- Define three validation subsets:

    o 'External to MICHEM' comprises 1009 molecules from DTU and ChemProp not present in the MICHEM set.

    o 'External to ChemProp' comprises 228 molecules from MICHEM and DTU not present in the ChemProp set.

    o 'External to both MICHEM and ChemProp' includes 128 molecules from DTU absent in both ChemProp and MICHEM sets.

The distribution of molecules in DTU, ChemProp and MICHEM sets and the number of chemicals in each validation subset are reported in Figure 2.2.

The three validation sets were used to further validate the model calibrated on MICHEM dataset (described in paragraph 4.5 and in the scientific publication in Appendix I) and directly compare its performance with that of the ChemProp model, as summarised in paragraph 4.6 and detailed in the scientific publication in Appendix II. Additionally, an extended dataset comprising 1555 chemicals was defined by merging MICHEM and 'External to MICHEM' sets (Table 2.3). This dataset was used to extend the previous model and develop novel models as summarised in paragraph 4.6 and detailed in the scientific publication in Appendix II.
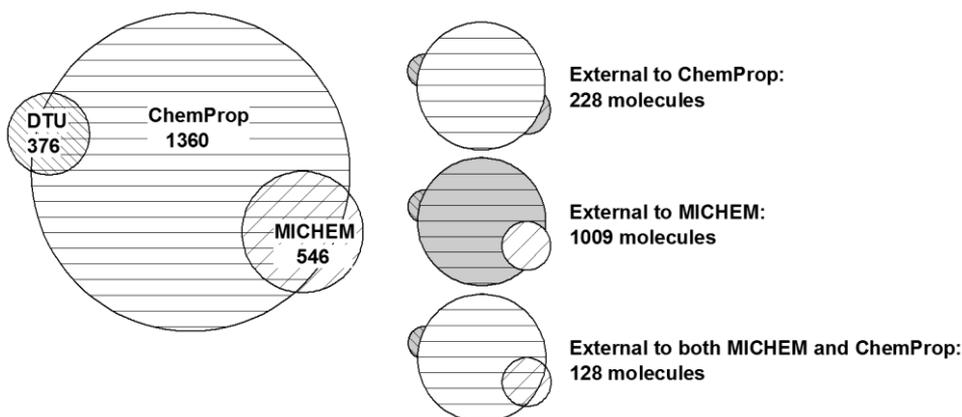

Figure 2.2. Illustration of the retrieved datasets and definition of the three validation subsets (grey areas).

2.4 *Pimephales promelas* dataset

Toxicity data expressing the concentration that kills 50% test fish (LC_{50}) over a test duration of 96 hours were used as the endpoint for the development of QSAR models. Unlike the case of *Daphnia magna*, a large database of consistent data measured in the same laboratory for more than 600 chemicals was available: the MED-Duluth fathead minnow database [Russom et al., (1997)]. However, since the priority was to define a dataset that was as large as possible, data from other sources were also gathered. The experimental data used in this study were retrieved from three databases, namely ECOTOX [US EPA, (ECOTOX)], EAT5 [ECETOC, (2003)] and OASIS. The OASIS database was downloaded from the OECD QSAR Toolbox [*The OECD QSAR Toolbox*, (2010)]. In the EAT5 database, LC_{50} was reported as EC_{50} with mortality as observed effect. Similarly to the *Daphnia magna* dataset, no filter was applied on test conditions and design. The preparation of the data was carried out in a similar manner with a substantial difference concerning the treatment of salts and mixtures. The dataset preparation was as follows:

- Deletion of LC_{50} data reported as ranges or as greater/smaller than a certain value.
- Check that CAS registry numbers and chemical names corresponded to the same structure by means of queries to the ChemSpider database [Royal Society of Chemistry] and the Chemical Identifier Resolver (CIR) at NCI/NIH [NCI/CADD Group].
- Manual inspection of all records that had mismatches between the results obtained from the chemical name and the CAS-RN by the additional support of the PubChem [Bolton et al., (2008)] and the Sigma-Aldrich [Sigma-Aldrich Co.] databases.
- Deletion of all the records for which the mismatch was not resolved.
- Deletion of toxicity data expressed as "%", "% v/v" and "AI ng/L".
- Identification of duplicates of the same measurement. Only one value was retained.
- Removal of inorganic compounds.
- Deletion of information regarding stereochemistry.
- Inspection of disconnected structures (salts and mixtures) by means of a dissociation algorithm for the identification of dissociable chemical species.

2.4 Pimephales promelas dataset

- Screening of the potential dissociation products for non-toxic species: if more than one chemical species was considered the source of toxicity, the record was removed; otherwise, the only one dissociation product assumed the sole source of the measured toxicity was retained in the dataset.
- Neutralisation of the retained dissociation products (with the exception of quaternary ammonium ions).
- Final validation of the structures by comparison with the structures in the OpenTox database.
- Calculation of the median toxicity for molecules with multiple experimental values.
- Definition of an alert for inconsistent data from the pooled standard deviation over the entire dataset.
- For all molecules with standard deviation larger than the alert, check of the experimental values in the original scientific publication. If the original study was not available or not found, the corresponding value was deleted.

Additional details about the data treatment steps can be found in the manuscript in Appendix III. The software used for the data preparation and the OpenTox database were made available by the QSAR group at the Technical University of Denmark.

The final dataset was made up of 908 organic chemicals (Table 2.3) and can be downloaded from:

http://michem.disat.unimib.it/chm/download/toxicityfish.htm.

This dataset was used for the development of QSAR models with LC_{50} values expressed as negative logarithm of molarity (-Log(mol/L)). A few models were also developed on the sole MED-Duluth fathead minnow database (566 compounds) [Russom et al., (1997)] in order to check if, and to what extent, the higher heterogeneity of the 908 chemicals dataset affected the performance of the developed QSAR models. These models were then externally validated on the subset of compounds not present in the MED-Duluth database (349 compounds).

Table 2.3. Characteristics of the analysed datasets.

Dataset	Endpoint	Duration	No. of compounds
Daphnia magna MICHEM	LC_{50}	48 hours	546
Daphnia magna Extended MICHEM	LC_{50}	48 hours	1555
Pimephales promelas	LC_{50}	96 hours	908
MED-Duluth database	LC_{50}	96 hours	566

CHAPTER 3

Methods

'Keep it simple, stupid'

Design principle noted by the U.S. Navy, 1960.

'It is pointless to do with more what can be done with fewer'

William of Ockham, *Summa Totius Logicae*, 1323.

Notation

The following mathematical notation is used throughout this thesis:

- Scalars are indicated by italic lower-case characters (e.g. x_{ij}).
- Vectors are represented by bold-type lower-case characters (e.g. $\mathbf{x}$).
- Matrices (e.g. two-dimensional arrays) are indicated by bold-type upper-case letters (e.g. $\mathbf{X}$). The size of a matrix is (n x p), where n is the number of rows and p the number of columns. If not otherwise specified, chemicals are arranged in rows and molecular descriptors in columns. The value of the j-th descriptor for the i-th molecule is indicated as x_{ij}.
- Scalar products (inner products) between two vectors are indicated as $\mathbf{x^t y}$, where $\mathbf{x^t}$ is the transposed (a row vector) of vector $\mathbf{x}$ and $\mathbf{y}$ is a column vector.

3.1 Description of molecular structure

3.1.1 Molecular format: SMILES notation

The structure of the chemicals under study was represented by means of the simplified molecular input line entry system (SMILES). SMILES is a molecular format initiated by David Weininger at the United States Environmental Protection Agency (US EPA) and completed at Pomona College. SMILES make use of a line

notation, i.e. a typographical notation system using printable characters. The information regarding elements, connectivity, bond orders and configuration can be encoded in the SMILES string. What SMILES notation does not specify is the spatial disposition of the atoms, which makes it a two-dimensional molecular representation. The SMILES theory is given here concisely:

- Atoms are specified by the chemical symbol between square brackets. Brackets can be omitted for common elements of organic compounds (B, C, N, O, P, S, F, Cl, Br, and I). In this case, the number of bonded hydrogen atoms for standard valence states is assumed implicitly.

- Bonds are assumed to exist between adjacent atoms in the SMILES string. Single bonds are either implicit or represented by hyphens ('-'); double bonds are coded by the equal symbol ('='); triple bonds are represented by hash marks ('#').

- Atom labelling with numbers is used to specify a bond between non-adjacent atoms in the SMILES string: this notation is used to indicate a ring closure for cyclic structures.

- Branches are enclosed in brackets.

- Aromatic structures are specified in the Kekulé-like form with conjugated double bonds, provided the availability of algorithms able to determine aromaticity. Aromaticity can be explicitly specified using lower case atomic symbols for rings constituted of C, N, O, S, P, As, Se.

- R/S configuration around a tetrahedral centre is specified by the use of a single or double at-sign ('@' or '@@'), which indicates that the substituents are listed in the SMILES string anticlockwise or clockwise, respectively.

- E/Z configuration is coded by a pair of slash ('/') and/or backslash ('\') symbols. For the E-configuration, two slash or two backslash symbols are used; for the Z-configuration, a pair of slash-backslash is used.

- Disconnected structures are specified by individual SMILES strings separated by a full stop ('.').

An example of the generation of SMILES strings for toluene and 1,3-butadiene is given in Figure 3.1. It is noteworthy that SMILES strings are not a unique identifier. This means that the structure of a chemical compound can be specified by different SMILES strings. Therefore, the direct comparison of SMILES strings to check if two structures are identical is not feasible. Before performing such comparison, it is

necessary to generate canonical SMILES. Canonical SMILES are obtained via canonicalisation algorithms able to generate always the same SMILES string for a particular structure. Different algorithms for the generation of canonical SMILES can output different results. Therefore, it is necessary to compare canonical SMILES obtained from the same algorithm or software package: under these conditions, SMILES can be considered a unique identifier of the molecular structure. The full theory of SMILES is described by Daylight Inc. [Daylight, SMILES].
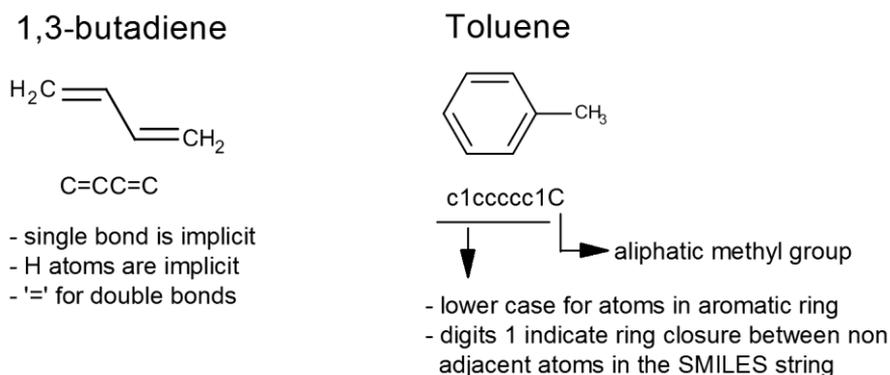

Figure 3.1. Generation of SMILES strings for 1,3-butadiene and toluene.

3.1.2 Molecular descriptors

Molecular descriptors are the independent variables used in QSAR investigations to describe the molecular structure and its properties. One definition of molecular descriptor was given by Todeschini and Consonni, (2000):

> *"The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment"*

This definition tells that the molecular descriptor is, ultimately, a number that encodes a fraction of the enormous information that is enclosed in the chemical structure. A first main distinction must be made between experimental measurements and theoretical molecular descriptors, which derive from a symbolic representation of the molecule. Theoretical molecular descriptors became in the last decades by far the most commonly employed variables in QSAR investigations, hence, hereinafter, they will simply be referred to as molecular descriptors or descriptors for sake of simplicity. Different symbolic representations of the chemical structure were defined

and, consequently, different types of information can be extracted through the application of algorithms. For instance, the chemical formula only includes information about the constituent atoms; a 2D molecular graph can additionally provide information about connectivity and bond orders; a 3D conformer specifies also the spatial disposition of the atoms. Hence, the information content of a molecular descriptor depends on both the type of chemical representation from which it is calculated, and the algorithm defined for its calculation.

Molecular descriptors can be classified according to the type of molecular representation they are calculated from.

Molecular descriptors obtained from the chemical formula can be considered zero-dimensional (0D) because no information regarding bonds is accounted for. Examples are the number of atoms of a certain element, molecular weight and, in general, all constitutional descriptors and atomic properties functions. Atomic properties are often used as weights to characterise atoms of the molecule.

Presence/absence or count of different functional groups or substructures can be defined as one-dimensional (1D) descriptors because the connectivity between atoms is expressed within each substructure.

From a molecular graph it is possible to compute descriptors that take into account the connectivity between atoms, such as topological and connectivity indices, pairs of atoms at a certain topological distance and others. Hence, this molecular representation and the descriptors calculated from it are considered two-dimensional (2D).

The geometrical (3D) representation of a molecule allows the derivation of descriptors that encode information about the spatial distribution of the atoms. These descriptors are commonly called 3D or geometrical descriptors. Since a geometrical representation implies the knowledge of the relative positions of the atoms in the three-dimensional space, i.e. the atomic coordinates (x, y, z), geometrical descriptors, usually, provide a larger amount of information than 2D descriptors. The additional specification of the electronic structure enables the calculation of quantum-chemical descriptors. Despite their high information content, geometrical descriptors present some drawbacks too. First, they require molecular geometry to be defined (typically, the minimum energy conformation is used). Second, some descriptors used in grid-based QSAR analyses need alignment rules, necessary to obtain comparable results. Additionally, the geometries obtained by means of different methods or software packages may differ. Consequently, a QSAR model based on 3D descriptors should

either implement also the geometry optimisation algorithm used for the preparation of the dataset, or require users to follow the same steps.

Four properties were recognised as basic requirements that molecular descriptors must possess [Todeschini and Consonni, (2000); Consonni and Todeschini, tutorial]:

- Invariance to atomic labelling and numeration.
- Invariance to rotation and translation of the molecule.
- An unambiguous definition, computable by means of algorithms.
- Values in a suitable numeric interval.

Additional characteristics that good molecular descriptors should possess were discussed by Randić [Randić, (1996)].

Currently, thousands of molecular descriptors were defined and can be calculated by means of dedicated software. In this study, different types of molecular descriptors were investigated for the calibration of QSAR models. Given the issues related to the geometrical description of the molecular structure (time required for the optimisation, consistency of the optimisation results, conformer used for modelling), only zero-, one- and two-dimensional descriptors were used.

3.1.3 Binary fingerprints

In the last decades, a particular description of the molecular structure was given much interest, especially for problems related to similarity searching, which is the identification of similar structures in large libraries. This description, referred to as binary fingerprints, consists in binary vectors encoding the presence (1) or absence (0) of specific fragments or substructures. Different algorithms for the calculation of binary fingerprints were defined and implemented in software packages [Yap, (2010)]. The main difference is related to the definition of the fragments, which can be either based on a pre-existing library, or derived from the analysed dataset through the generation of all the fragments meeting some criteria. Especially in the latter case, the identification of all the fragments from large heterogeneous datasets can lead to the generation of several thousands of fragments. The corresponding binary vectors are generally sparse, i.e. they will mainly be constituted by null entries. The mathematical treatment of such vectors is not trivial. Another issue is related to the potential identification of new fragments if new molecules to which the model is applied are considered. A way to handle this situation is the re-calculation of the

fingerprints for all the compounds. In order to overcome these issues, particular attention was paid to the use of hashing algorithms. These algorithms allow compressing the information in a binary vector of predefined length (the length is defined as number of bits). The advantages connected with the definition of shorter vectors are an easier mathematical treatment and the possibility to independently calculate the fingerprints for new sets of compounds. Obviously, the compression of the information is accompanied with a loss of information. In particular, the correspondence between each bit and a specific fragment is lost, i.e. more fragments will degenerate on the same bit. The implications are a confusion between different fragments (which can be chemically different) and a more difficult interpretation of the results from a chemical viewpoint.

In this study, two different types of hashed binary fingerprints were used:

- Extended connectivity fingerprints: they are obtained through the generation of atom-centred fragments in an iterative procedure [Morgan, (1965); Rogers and Hahn, (2010)]. Different properties can be used (also in combination) to discriminate between fragments, such as atom type, charge, aromaticity, attached hydrogens, connectivity and bond order. The procedure to generate fragments starts by considering only atoms (*radius* equal to 0); the second iteration considers also bonded atoms (*radius* equal to 1), the third iteration considers bonded atoms and atoms at topological distance equal to 2 (*radius* equal to 2). The algorithm proceeds in this manner until a pre-defined maximum *radius*.
- Path fingerprints: the same properties for discrimination among fragments can be applied to path fingerprints as well. The main difference with respect to extended connectivity fingerprints relies in the fact that linear fragments are taken into account [Daylight, Fingerprints].

A stepwise example of the generation of extended connectivity fingerprints is given in the scientific publication in Appendix II.

3.2 Selection of molecular descriptors

A situation encountered in many scientific fields is the presence of a large number of variables that are investigated by means of multivariate techniques. QSAR investigations are particularly affected by this problem because several thousands of descriptors can be easily and quickly calculated by means of software packages. As

outlined in paragraph 1.3, the approach adopted by scientists often consists in the generation of a large pool of molecular descriptors, followed by the application of techniques able to identify an optimal subset for the problem under study.

The reason why it is advisable to select a subset of descriptors instead of using all the available ones is dual. On one hand, it is sensible to assume that not all the calculated molecular descriptors are relevant for the analysed problem. Therefore, the inclusion of molecular descriptors unrelated to the property under study would be chemically inappropriate and difficult to justify. On the other hand, the presence of irrelevant descriptors is not inconsequential because it can negatively affect the quality of the derived mathematical model. For example, some descriptors may model the experimental noise, which typically one would not like to account for. A principle often invoked is the principle of parsimony (known as Occam's razor), which indicates to opt for the simplest solution. In the framework of QSAR analysis, it is translated as an indication to keep models as simple as possible, which implies, also, a low number of molecular descriptors [Cronin et al., (2004); Dearden et al., (2009); Yi and Zhang, (2012)].

In this study, a preliminary filter was applied in order to discard useless descriptors according to the following two criteria:

- Presence of missing values.
- Constant or near-constant values for all chemicals in the dataset

The techniques used afterwards to select molecular descriptors can be distinguished in unsupervised and supervised methods. Unsupervised and supervised methods used in this study are presented in paragraph 3.2.1 and paragraph 3.2.2, respectively.

3.2.1 Unsupervised variable reduction

Unsupervised methods, sometimes referred to as variable reduction methods, only consider relationships between descriptors. The property being modelled, hereinafter referred to as response vector, is never used, so it can be said that there is no 'supervision' of the response. These techniques are often used as preliminary filter to remove descriptors according to some criteria.

Unsupervised variable reduction was carried out by analysing the correlation between pairs of descriptors. The presence of correlated descriptors in regression models is, in fact, dissuaded. On one side, multicollinear variables negatively affect

the model, especially when linear approaches are used; on the other side, the concomitant presence of descriptors that encode the same information is inappropriate from a chemical viewpoint. The check on correlation was based on the comparison of the absolute coefficient of correlation between pairs of descriptors with a fixed threshold. If the absolute correlation coefficient was greater than the threshold, then the descriptor possessing the largest average correlation with all the other descriptors was discarded. A threshold equal to 0.95 was used.

3.2.2 Supervised variable selection

Supervised methods, often referred to as variable selection methods, are used to identify optimal subsets of descriptors for modelling the property under study. To this end, the response vector is used with several different combinations of molecular descriptors to derive mathematical models in a trial and error procedure. The quality of each model is measured by means of the fitness function, which is the quantity to be optimised. The fitness function can coincide with one statistical parameter (e.g. the coefficient of determination or the root mean square error) or be a function of several parameters.

One approach to variable selection can be based on the generation of all the possible combinations of the available descriptors. Although this strategy, sometimes called all subset models (ASM), would guarantee the best subset of descriptors be identified, it is often impracticable due to the enormous calculation time needed. Indeed, the total number of combinations (t_{ASM}), i.e. models, of p descriptors is given by:

$$t_{ASM} = 2^p - 1 \qquad (1)$$

Typically, one is interested in simple models that comprise few molecular descriptors (V). The number of combinations of p descriptors in subsets of size from 1 to V is:

$$t = \sum_{k=1}^{V} \left(\frac{p!}{k!(p-k)!} \right) \leq 2^p - 1 \qquad (2)$$

However, also in this case, the number of combinations can be very large, making the calculation not feasible. Figure 3.2 shows the increase in the number of combinations, i.e. models, (in logarithmic scale) as a function of the number of
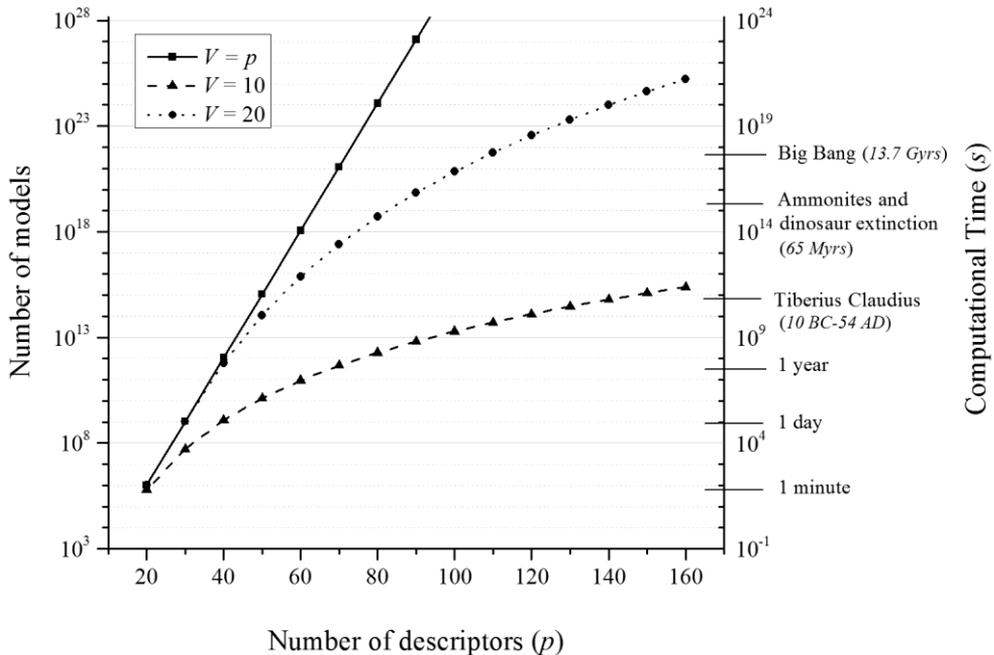
Figure 3.2. Number of models versus number of descriptors, p, for an all subset models method with V=10, V=20 and V=p, assuming a computational speed of 10,000 models per second. Y-axes are reported in logarithmic scale.

descriptors *p* for three cases with *V* equal to 10, 20 and *p*, respectively. On the secondary vertical axis, an estimation of the time needed for calculation is reported, assuming a computational speed of 10,000 models per second (reasonable estimate for current laptops). For example, the generation of all the models comprising from one to ten descriptors, selected from a pool of 140 descriptors, would take so long that we should have started the calculation during the empire of Tiberius Claudius (41-54 AD) to have it completed by now.

Several methods were defined in order to overcome these issues, such as stepwise selection [Efroymson, (1960); Hastie et al., (2009)], genetic algorithms [Holland, (1992); Leardi, (2003)], ant colony optimisation [Goodarzi et al., (2009)], particle swarm optimisation [Shen et al., (2004)] and sequential replacement [Miller, (1984); Miller, (2002)]. Additionally, regression methods able to carry out a concomitant selection of the variables were developed, such as LASSO and elastic net [Efron et al., (2004); Tibshirani, (1996)].

In this study, variable selection was carried out by means of genetic algorithms (GAs) and an adaptation of the sequential replacement method of Miller [Miller, (1984)] developed during this project thesis, called reshaped sequential replacement

(RSR). These two methods are described in paragraph 3.2.2.1 and 3.2.2.2, respectively.

3.2.2.1 Genetic algorithms

Genetic algorithms (GAs) are a nature-inspired optimisation method that implements the concept of 'survival of the fittest' expressed in Darwin's theory of the evolution. In fact, GAs are based on a population of agents, called *chromosomes*, that compete with each other for survival in the population [Kim and Cho, (2006)]. Being an optimisation method, GAs aim at finding a condition of maximum or minimum of a defined response (or fitness function) for a certain number of independent variables, which the response itself depends on. A peculiar application of genetic algorithms consists in searching for an optimal subset of molecular descriptors, given a set that comprises a high number of independent descriptors.

In GAs application to variable selection, each chromosome is a binary vector of p bits, where p is the total number of descriptors; hence, there is a univocal correspondence between bits and descriptors. Each chromosome is associated to a model comprising some specific descriptors: a bit set to 1 means that the corresponding descriptor is included in the model and vice versa for null bits. A statistical parameter, the fitness function, is used to express the quality of each chromosome and is the variable to be optimised. Two operations are used to generate new chromosomes, namely crossover and mutation:

- Crossover implies the selection of two parent chromosomes from the population (the selection can be random or biased towards the best chromosomes) and the generation of offspring. Children chromosomes have the shared genetic pool of the parents: bits being 0 or 1 in both parents remain 0 or 1 in the offspring; bits mismatching between parents can be set to 0 or 1 according to a crossover probability.
- Mutation is the operation by which bits of a chromosome can be inverted generating mutants. The likelihood for mutation to occur is expressed by a mutation probability that is typically smaller than the crossover probability. Mutation is used to limit the chance that the population becomes stuck in local optima. However, the mutation probability should not be too high to avoid the population to drift away from the likely optimal region.

The size of the population is kept constant during the run. Consequently, the worst chromosomes are discarded while better ones enter the population. The analogy with Darwin's theory is trivial.

A typical implementation of genetic algorithms is made up of the following steps:

1. Creation of a random initial population of chromosomes.
2. Evaluation of the fitness function of initial chromosomes.
3. Generation of new chromosomes by crossover and mutation.
4. Evaluation of the fitness function of children and mutants.
5. Update of the population.
6. Iteration of steps 2-5 until a stop criterion, e.g. maximum number of iterations or stability of the population over a certain number of iterations.

When applied to datasets consisting in a large number of molecular descriptors or low number of compounds, GAs are affected by the risk of overfitting, i.e. the definition of models that lack of generality and/or model noise, and of convergence to local optima. In order to overcome these issues, the strategy suggested by Leardi and González was used [Leardi and González, (1998)]. According to this strategy, instead of carrying out one single run for several iterations, a number of short independent runs are carried out starting with initial random populations. The frequency of selection of molecular descriptors is recorded over the independent runs. The final stage is a stepwise selection that, starting from a one-descriptor model (the most frequent one), adds further descriptors, based on their frequency of selection over the runs.

Since in QSAR analysis many descriptors can be highly correlated, they can alternatively be included in the models, leading to pairs (or clusters) of correlated descriptors with equal or similar frequencies of selection. The strategy of Leardi and González would add them together in the final stepwise selection. In order to overcome this issue and better explore the optimal region, a part from carrying out the final stepwise selection based only on the frequency of selection, final models were also developed from an all subset strategy based on the generation of all the combinations of a small pool of descriptors.

3.2.2.2 Reshaped sequential replacement

The reshaped sequential replacement (RSR) method was developed during this study as a modern adaptation of the sequential replacement (SR) method proposed by Miller [Miller, (1984)]. The core of the algorithm, i.e. the replacement procedure, was kept faithful to the original definition, but new features were added in order to overcome some drawbacks that the original method suffered from.

The RSR algorithm, like GAs, is based on a population of agents (corresponding to models) whose quality is measured by a fitness function. In the terminology of RSR, agents are called *seeds* but they are identical entities to chromosomes: binary vectors of *p* bits, *p* being the total number of molecular descriptors. In analogy with GAs, the identification of optimal subsets of descriptors is obtained through a trial and error procedure that implies the generation of new seeds (models) and the storage of the information through the update of the population. The generation of new seeds is accomplished via the replacement procedure defined by Miller, which is exemplified in Figure 3.3 for a single seed. In this example, typographic characters are used instead of binary vectors for sake of clarity. An initial seed comprising descriptors *O, P, T, I, C, U, Y* is randomly generated. The first descriptor, *O*, is replaced with all the remaining ones one at a time, keeping the other descriptors in the model fixed. Thus, *O* is replaced with *A* to give model *APTICUY*, then with *B* to give *BTICUY*, and so on. Descriptors already present in the model (*PTICUY*) are obviously not considered in the replacement. The fitness function of all the models obtained from the replacement procedure is evaluated and only the best, say *LPTICUY*, is retained. In analogous manner, the replacement procedure is applied to the other descriptors in the initial model. The outcome is a population of seven improved models. The best one, say *OPTIMUY*, is retained and transferred to the second iteration, where the replacement procedure is applied in the same way. Note that in the second iteration no replacement of descriptor *U* gives improvements. The procedure is iterated as long as replacements provide improved models, i.e. until convergence of the population. The final model, *OPTIMUS*, is obtained by the replacement of descriptor *Y* in model *OPTIMUY* with descriptor *S*.

An important difference with respect to chromosomes used in GAs is that the number of bits set to 1, i.e. the number of descriptors included in the model, is kept constant, whereas in chromosomes this can change as a consequence of crossover and mutation.
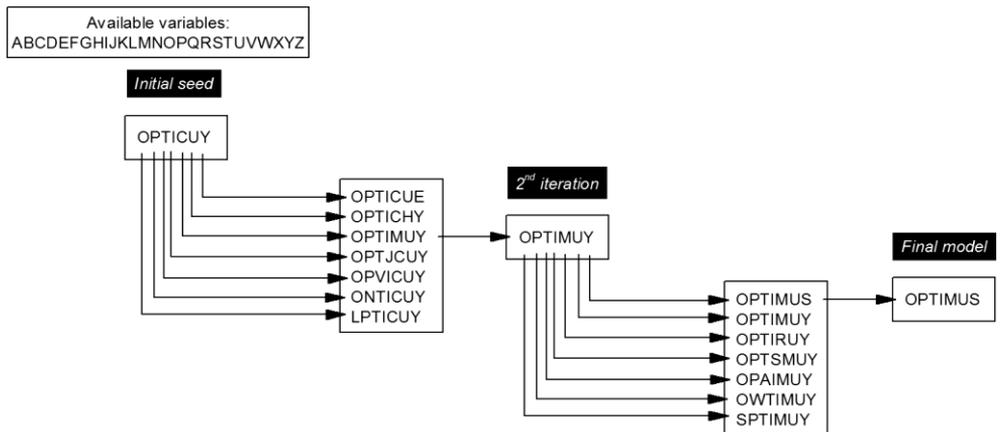
Figure 3.3. Replacement procedure of the sequential replacement method.

The original sequential replacement method generates a large number of combinations of the descriptors and any combination providing an improvement in the fitness function is retained, regardless of the magnitude. Hence, the method is subject to the risk of overfitting. Additional drawbacks are connected with the potentially long computational time required to reach convergence and the fact that the fitness function in the original formulation was not related to the predictive power, but only to the ability to fit the training data.

The reshaped sequential replacement (RSR) intends to overcome these and other issues described in the literature [Todeschini et al., (2004)] through the implementation of additional features. The implemented new functionalities are briefly described below. The full theory is explained in the scientific publication in Appendix IV.

- **Fitness function**: instead of the residual sum of squares (*RSS*), which is only related to the fitting ability of the model, the coefficient of determination in cross-validation (Q^2_{cv}) is used.
- **Tabu list**: the tabu list is a preliminary coarse-grained filter applied prior to the replacement procedure. Descriptors expected to be poor for modelling according to a certain criterion (e.g. their correlation with the modelled property) are stored in the tabu list and excluded from the replacement phase. They can be recovered at a later stage. The tabu list is intended to speed up the modelling time by reducing the number of available descriptors.
- **Roulette wheel**: the roulette wheel is an algorithm biased towards high quality solutions used in genetic algorithms for the selection of parent chromosomes.

In the RSR algorithm, it is used to generate the initial population of seeds by forcing the inclusion of descriptors supposed to be relevant for modelling according to a certain criterion (e.g. their correlation with the modelled property). The roulette wheel aims at generating an initial population close to the optimal region.

- **QUIK rule**: the Q^2 under influence of K (QUIK) rule is a statistical test used to detect models suffering from multicollinearity between descriptors [Todeschini et al., (1999)]. The test is based on the K correlation index [Todeschini, (1997)], which describes the overall correlation among a set of descriptors. The test is carried during the replacement procedure.

These four features directly affect the replacement phase leading to potentially different final populations of models. Additional functions and tests were implemented to analyse the final population of models:

- **Y-scrambling**: the y-scrambling is a well-known permutation test used to detect the presence of chance correlation between descriptors and response [Lindgren et al., (1996)].

- **R-function based rules**: two rules defined in the literature to detect models suffering from redundancy of good descriptors and presence of noisy descriptors were implemented [Todeschini et al., (2004)]. These tests are based on two functions, R^p and R^N, that compare the correlation coefficient of the model and of each individual model descriptor with the response.

- **Nested model test**: this test is intended to identify nested models. A model F is considered nested of another model G, if G contains all the descriptors of F and the predictive power of G is only slightly higher than that of F.

- **Model distance and model correlation**: a measure of correlation and distance between the final models is provided based on the *CMC* and *CMD* indices [Todeschini et al., (2009)]. These two indices help to understand whether models based on different sets of descriptors are actually different in their nature or not.

A workflow of the algorithm is provided in Figure 3.4. The new functionalities are highlighted in grey boxes. Applications of the RSR algorithm to QSAR datasets are provided in the scientific publication in Appendix V. The RSR toolbox for MATLAB can be downloaded at:

http://michem.disat.unimib.it/chm/download/rsrinfo.htm.
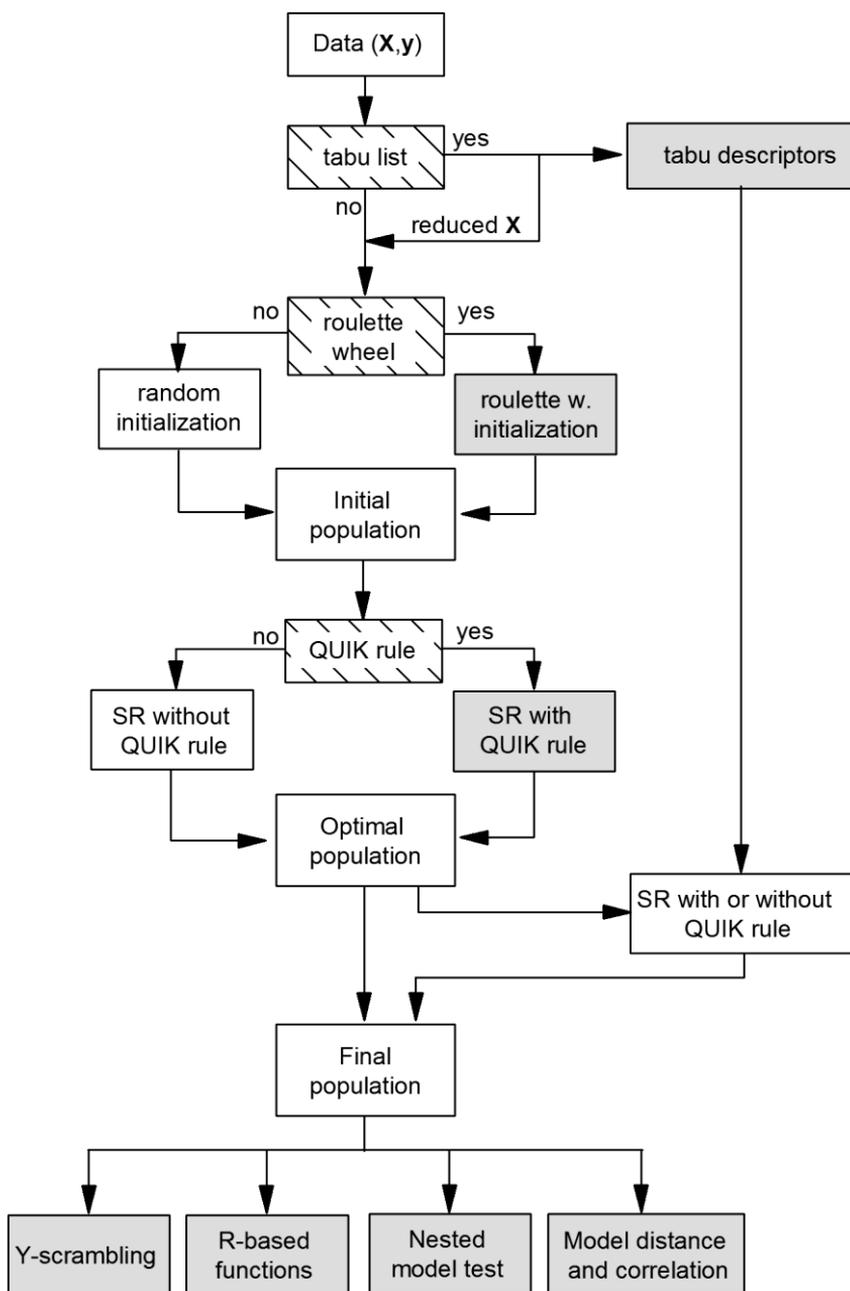
Figure 3.4. Workflow of the RSR algorithm. New functionalities are highlighted by grey boxes.

3.3 Regression methods

A variety of regression methods, possessing different characteristics, was used in order to establish mathematical relationships between the analysed properties and molecular descriptors.

3.3.1 Multiple linear regression

Multiple linear regression (MLR) is a mathematical approach to establish linear relationships between a set of descriptors, the independent variables $\mathbf{x}_j$, and a quantitative response, the dependent variable $\mathbf{y}$ [Hastie et al., (2009)]. The relationship takes the form of:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3}$$

where $\mathbf{y}$ is the response vector, $\mathbf{X}$ is the model matrix, $\boldsymbol{\beta}$ is the vector of true regression coefficients and $\mathbf{e}$ is the vector of the errors. The corresponding mathematical model is expressed as:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} \tag{4}$$

where $\mathbf{b}$ is the vector of the estimates of the true coefficients $\boldsymbol{\beta}$ and $\hat{\mathbf{y}}$ is the vector of calculated responses. Alternatively, the model can be written in non-matrix terms as:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} \tag{5}$$

where $\hat{y}_i$ is the calculated response of the *i-th* molecule, b_0 is the intercept, b_j are the coefficients of the p model descriptors and x_{ij} is the value of the *j-th* descriptor for the *i-th* molecule.

Different methods can be used to estimate the coefficients of the model. In this study, ordinary least squares (OLS) and partial least squares (PLS) methods were used.

3.3.1.1 Ordinary least squares regression

The ordinary least squares (OLS) method provides an estimate of the coefficients of a linear model by minimizing the residual sum of squares (*RSS*) between the calculated and the experimental response vectors, defined as:

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{6}$$

where y_i is the experimental response of the *i-th* molecule, $\hat{y}_i$ is the corresponding value calculated by the model and the summation runs over the *n* molecules.

The coefficients of the model, b_j, are estimated as:

$$\mathbf{b} = \left(\mathbf{X^tX}\right)^{-1}\mathbf{X^ty} \tag{7}$$

where $\mathbf{X}$ is the matrix of model descriptors and $\mathbf{y}$ is the response vector.

From the regression coefficients, it is possible to derive the standardised regression coefficients (equation 8):

$$b_j^{'} = b_j \cdot \frac{s_j}{s_y} \tag{8}$$

where s_j and s_y are the standard deviations of the *j-th* descriptor and of the response, respectively. Standardised coefficients are independent from the descriptors scale and, therefore, tell about the importance of each individual descriptor in calculating the modelled response.

The estimation of the coefficients is negatively affected by multicollinearity between model descriptors due to problems with the inversion of the matrix $\mathbf{X^tX}$. Therefore, the correlation among model descriptors should be kept low. Furthermore, the estimates of the coefficients are inaccurate also when the number of descriptors is greater than the number of molecules.

3.3.1.2 Partial least squares regression

Partial least squares (PLS) is a linear regression method that has some similarities with principal component analysis (paragraph 3.8) because it operates a projection of the molecules in a new space defined by linear combinations of the original descriptors, called latent variables (LVs). PLS is suitable when dealing with a large number of descriptors, when the ratio molecules/descriptors is lower than one and in presence of correlated descriptors (multicollinearity). Therefore, PLS is suitable in situations when OLS is not appropriate.

Unlike PCA, PLS uses the $\mathbf{y}$ vector to find the new directions (the latent variables). Therefore, PLS deals with a dual problem where the aim is to find directions of maximum variance in $\mathbf{X}$ (like PCA) and of maximum correlation with $\mathbf{y}$ at the same time [Hastie et al., (2009); Andersson, (2009)]. The projection in the new space is carried out as:

3.3 Regression methods

$$\mathbf{T} = \mathbf{XV} \tag{9}$$

where $\mathbf{T}$ is the score matrix, i.e. the coordinates of the molecules in the new space, $\mathbf{X}$ is the original data matrix and $\mathbf{V}$ is the matrix of the loadings. The model is then calibrated in the space of the LVs as:

$$\mathbf{y} = \mathbf{Tq'} + \mathbf{e} \tag{10}$$

with $\mathbf{q}$ being the vector of PLS regression coefficients. The replacement of $\mathbf{T}$ in equation 10 with its expression from equation 9 leads to:

$$\mathbf{y} = \mathbf{XVq'} + \mathbf{e} \tag{11}$$

The problem is solved for $\mathbf{q}$ and then the coefficients of the original descriptors in the model of equation 4 can be obtained as:

$$\mathbf{b} = \mathbf{Vq'} \tag{12}$$

A number of algorithms exist to solve equation 11 and they return results that are similar but not identical.

PLS allows to reduce the dimensionality of the data by selecting a number of LVs smaller than the number of original variables. The selection of the optimal number of LVs is typically carried-out in cross-validation.

3.3.2 *K*-nearest neighbours

The *k*-nearest neighbours (*k*NN) is a method initially defined to deal with discrete or qualitative responses (classification) [Kowalski and Bender, (1972)]. Nevertheless, it is suitable and easily adapted to the regression case. The algorithm implies the calculation of the pairwise distance of each molecule from the other compounds in the training set according to a predefined metric. The *k* closest molecules, the nearest neighbours, are identified and used to calculate the response as mean or weighted mean. The main differences with OLS and PLS regard the fact that *k*NN does not provide a functional model and it acts only at a local level. It is evident how *k*NN embodies the congenericity principle by assuming that molecules close in the descriptors space, i.e. considered structurally similar, possess enough similar experimental responses to enable a local prediction.

In this study, the response of a compound was calculated as weighted mean of the experimental toxicities of the *k* nearest neighbours according to equation 13:

$$\hat{y}_r = \sum\nolimits_{t=1}^{k} y_t . w_t \tag{13}$$

where y_t and w_t are the experimental response and the weight of the t–th neighbour (taken from the training set), respectively, and the sum runs over the k neighbours. The weights, w_t, represent the similarity so that more similar neighbour molecules (i.e. closer) are associated with higher weights and vice versa. In this way, closer molecules have a larger contribution in the calculation of the property value, thus enforcing the congenericity principle. The value of k is typically selected in cross-validation.

It is noteworthy that the scaling of the descriptors, as well as the distance measure, largely affect the results because different molecules can be identified as nearest neighbours of a certain compound. The choice of the proper distance and scaling is not always trivial and it can be needed to try different pairs and select the one giving the best results.

3.3.2.1 Distance measures

Several different distance measures were used in order to evaluate the similarity between chemicals when using the kNN method. For real data, the Euclidean, Mahalanobis, Lance-Williams, Jaccard-Tanimoto, and Soergel metrics were used [Todeschini et al., (2015)]. Let $\mathbf{x_s}$ and $\mathbf{x_t}$ be two vectors that contain the values of p descriptors for molecules s and t, respectively. The Euclidean distance, d_{st}, between the two molecules, is calculated as:

$$d_{st} = \sqrt{\left(\mathbf{x_s} - \mathbf{x_t}\right)^t \left(\mathbf{x_s} - \mathbf{x_t}\right)} = \sqrt{\sum_{j=1}^{p} \left(x_{sj} - x_{tj}\right)^2} \tag{14}$$

where x_{sj} is the value of the j-th descriptor for molecule s, x_{tj} is the corresponding value for molecule t and the summation runs over the p descriptors.

The Mahalanobis distance takes into account also the covariance in the data. The distance between molecules s and t is defined as:

$$d_{st} = \sqrt{\left(\mathbf{x}_s - \mathbf{x}_t\right)^t \mathbf{S^{-1}} \left(\mathbf{x}_s - \mathbf{x}_t\right)} \tag{15}$$

where $\mathbf{S^{-1}}$ is the inverse of the covariance matrix of all the data.

The Lance-Williams distance between molecules s and t is defined as:

$$d_{st} = \frac{|\mathbf{x_s} \text{ - } \mathbf{x_t}|^t \mathbf{1}}{(|\mathbf{x_s}| + |\mathbf{x_t}|)^t . \mathbf{1}} = \frac{\sum_{j=1}^{p} |x_{sj} - x_{tj}|}{\sum_{j=1}^{p} \left(|x_{sj}| + |x_{tj}|\right)} \quad 0 \le d_{st} \le 1 \tag{16}$$

3.3 Regression methods

where x_{sj} is the value of the *j-th* descriptor for molecule s, x_{tj} is the corresponding value for molecule t and the summation runs over the p descriptors. **1** is a unit vector of length p.

The Jaccard-Tanimoto distance between molecules s and t, was derived from the corresponding coefficient as [Todeschini et al., (2015)]:

$$d_{st} = \sqrt{1 - \frac{\sum_{j=1}^{p} x_{sj} \cdot x_{tj}}{\sum_{j=1}^{p} x_{sj}^2 + \sum_{j=1}^{p} x_{tj}^2 - \sum_{j=1}^{p} x_{sj} \cdot x_{tj}}} =$$
$$= \sqrt{\frac{d_{st,euclidean}^2}{\sum_{j=1}^{p} x_{sj}^2 + \sum_{j=1}^{p} x_{tj}^2 - \sum_{j=1}^{p} x_{sj} \cdot x_{tj}}} \quad 0 \le d_{st} \le 1 \tag{17}$$

where $d_{st,euclidean}^2$ is the squared Euclidean distance. The Jaccard-Tanimoto distance requires the descriptors to be positively defined. In this condition it is bound in the range [0,1].

The Soergel distance between molecules s and t is defined according to equation (18):

$$d_{st} = \frac{\sum_{j=1}^{p} |x_{sj} - x_{tj}|}{\sum_{j=1}^{p} \max\{x_{sj}, x_{tj}\}} \quad 0 \le d_{st} \le 1 \tag{18}$$

Also the Soergel distance requires the descriptors to have positive values and ranges between [0,1].

The distance between two molecules is a measure of their dissimilarity: the higher the distance, the more dissimilar the molecules. Hence, there is an inverse relationship between distance and similarity measures, where similarity ranges from 0 (completely different structures) to 1 (identical structures). As described in paragraph 3.3.2, *k*NN was used in its weighted formulation, where the weights associated to the k neighbours represent their similarity to the target molecule. The weights were obtained from a transformation of the distance. Equation 19 was used to obtain the similarity from the Euclidean and Mahalanobis distance:

$$s_{st} = \frac{1}{1 + d_{st}} \tag{19}$$

where d_{st} is the distance between molecules s and t. Since the Soergel, Jaccard-Tanimoto and Lance-Williams metrics are bounded in the range [0,1], the corresponding similarity was obtained simply as:

$$s_{st} = 1 - d_{st} \tag{20}$$

where d_{st} is the distance between molecules s and t.

When binary fingerprints are used to describe the molecular structure, $\mathbf{x_s}$ and $\mathbf{x_t}$ are binary vectors that contain only 0 and 1 entries. The distance measures outlined earlier in this paragraph are not appropriate to deal with binary data. Hence, the Jaccard-Tanimoto similarity coefficient for binary data was used to evaluate the similarity [Jaccard, (1908)]. The calculation of the Jaccard-Tanimoto similarity coefficient is based on the analysis of the entries in common and differing in the two binary vectors according to Table 3.1, where a is the frequency of events $s=1$ and $t=1$, b the frequency of events $s=1$ and $t=0$, c the frequency of events $s=0$ and $t=1$, d the frequency of events $s=0$ and $t=0$. Hence, a and d count the frequency of entries in common between the two vectors, whereas b and c the frequency of entries differing in the two vectors.

The Jaccard-Tanimoto similarity coefficient is then calculated as:

$$S_{JT} = \frac{a}{a+b+c} \quad 0 \le S_{JT} \le 1 \tag{21}$$

Table 3.1. Frequency table of the entries in common and differing between two binary vectors.

	$t=1$	$t=0$
$s=1$	a	b
$s=0$	c	d

3.3.3 Support vector regression

Support vector regression (SVR) is the extension to the regression case of the support vector machine (SVM) classification method developed by Vapnik in the sixties [Vapnik and Lerner, (1963); Vapnik and Chervonenkis, (1964)]. SVR can be both a linear and nonlinear regression method that aims at finding a function that is as flat as possible and that has at most ε deviations from the real response values [Smola and Schölkopf, (2004)]. This version, referred to as ε-SVR, implies that a deviation (i.e. a residual between experimental and calculated response values) equal to or smaller than ε is acceptable, whereas larger deviations are penalised by means

3.3 Regression methods

of a loss function: the ε-insensitive loss function (Figure 3.5). The loss contribution, ξ_i, is equal to 0 if the residual is equal to or smaller than ε, otherwise it is proportional to the error:

$$\xi_i = \begin{cases} 0 & if \ |y_i - \hat{y}_i| \le \varepsilon \\ |y_i - \hat{y}_i| - \varepsilon & if \ |y_i - \hat{y}_i| > \varepsilon \end{cases} \tag{22}$$

The mathematical model is expressed in the form:

$$y = \mathbf{w}^t \mathbf{x} + b \tag{23}$$

where $\mathbf{w}$ is the vector of regression coefficient, $\mathbf{x}$ is the data vector and b is the intercept. The solution is found via the minimisation of:

$$\frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{i=1}^{n}\xi_i \tag{24}$$

where $\mathbf{w}^t\mathbf{w}$ is the norm and is related to the requirement of flatness, ξ_i is the loss contribution of the *i-th* molecule and C is a parameter that works as a trade-off between the two targets, namely flatness and prediction accuracy. The solution is defined as:

$$y = \sum_{i=1}^{n}\alpha_i\left(\mathbf{x}_i^t\mathbf{x}\right) + b \tag{25}$$

where α_i are the weights of the support vectors, $\mathbf{x_i}$; $\mathbf{x}$ is a chemical of interest. Therefore the solution provided by SVR is fully specified by a fraction of the training data, i.e. the support vectors. The support vectors are the molecules that lie either on the margin (residual equal to ε) or outside the ε-insensitive tube, whereas all the other training molecules have a null weight α_i and can potentially be removed without affecting the predictions of the model. The solution expressed in equation 25 is valid for linear problems. The nonlinear solution is obtained by a pre-processing step that performs a mapping of the original data from the input space (p descriptors) to a new feature space with more dimensions. The assumption is that exists a higher dimensional space where the problem can be linearly solved. The mapping is obtained via the use of kernel functions, the most common being the polynomial and the Gaussian kernels. The solution can be expressed in analogous way as:

$$y = \sum_{i=1}^{n}\alpha_i K\left(\mathbf{x}_i^t\mathbf{x}\right) + b \tag{26}$$

where K is the kernel function.
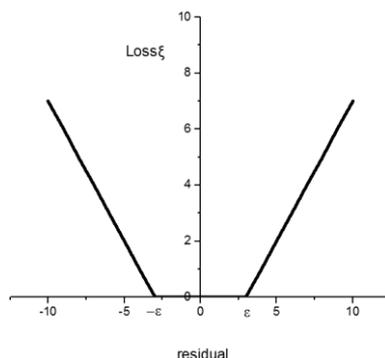
Figure 3.5. ε-insensitive loss function.

A number of loss functions can be used in place of the ε-insensitive function. Additionally, a new algorithm, called *v*-SVR, was developed where the parameter *v* in place of ε has to be defined [Schölkopf et al., (1999); Schölkopf et al., (2000)]. The authors demonstrated that *v* represents a lower bound on the fraction of training data that are support vectors and an upper bound on the fraction of bad predictions (residual greater than ε). The optimal values of the parameters are typically selected in cross-validation.

The data compression property of SVR, achieved by the definition of a number of support vectors that are a fraction of the total number of data, is seriously affected in situations characterised by high-dimensional data and, especially, noisy data. In these cases, the SVR algorithm tends to identify a large number of support vectors [Smola and Schölkopf, (2004)].

3.3.4 Gaussian process regression

Gaussian process regression (GPR) is a nonlinear regression method initially developed by O'Hagan [O'Hagan and Kingman, (1978)]. It is a nonparametric method where the response is modelled as a random variable having a zero mean Gaussian distribution. The predictions are obtained using a covariance function that tells how the model descriptors of training molecules relate to the response being modelled. Several covariance functions can be used, however Chen *et al.* [Chen et al., (2007)] recommend the use of the covariance function in equation 27:

$$C\left(\mathbf{x_i},\mathbf{x_f}\right) = a_0 + a_1\mathbf{x_i^t}\mathbf{x_f} + v_0\exp\left(-\frac{1}{2}\sum_{j=1}^{p}\frac{\left(x_{ij}-x_{fj}\right)^2}{w_j}\right) + \sigma^2\delta_{if} \quad (27)$$

where $\mathbf{x_i}$ and $\mathbf{x_f}$ are the vectors of the descriptors for the *i-th* and *f-th* molecule, respectively; p is the number of molecular descriptors. The set $\mathbf{\Theta} = \{a_0, a_1, v_0, w_1, \dots w_p, \sigma^2\}$ are the hyper-parameters, where a_0, a_1 and v_0 represent the importance of the first three terms of equation 27; $\{w_1, \dots w_p\}$ are 'weights' associated to molecular descriptors and σ^2 is the random noise. The last term is the Kronecker delta. The use of the w_j parameters enables the automatic relevance determination (ARD), which assigns larger weight to molecular descriptors highly related to the response. The covariance function in equation 27 is constituted by four terms associated to a constant offset, linear part, nonlinear part and noise, respectively. Hence, with such covariance function, GPR is able to handle both linear and nonlinear problems.

In GPR, the response is considered to have a joint Gaussian distribution depending on the training data:

$$\mathbf{y} = (y_i, \dots, y_n)^T \sim G(0, \mathbf{C}) \tag{28}$$

with $\mathbf{C}$ being the covariance matrix of training molecules. The assumption is that the response of a new test molecule, y^*, derives from the $(n+1)$-dimensional joint Gaussian distribution, namely:

$$\begin{bmatrix} \mathbf{y} \\ y^* \end{bmatrix} \sim G(0, \mathbf{C_{n+1}}) \tag{29}$$

where $\mathbf{y}$ is the vector of the responses of the training molecules, y^* is the response of the new molecule and $\mathbf{C_{n+1}}$ is the covariance matrix of the training set plus the new molecule. The decomposition of the $(n+1)$-dimensional covariance matrix $\mathbf{C_{n+1}}$ is depicted in Figure 3.6.

Predicting with a Gaussian process corresponds to sampling from this distribution. Hence, the prediction for the new molecule $\boldsymbol{x_*}$ is given in terms of mean and variance using the formulas for conditioning a joint Gaussian distribution:

$$y^* = E(y^*) = \mathbf{c_*^t C_n^{-1} y} \tag{30}$$

$$\sigma^{2*} = c_{**} - \mathbf{c_*^t C_n^{-1} c_*} \tag{31}$$

The optimisation of the hyper-parameters by means of cross-validation is often not feasible due to the large number of hyper-parameters (especially if the ARD is used). Alternatively, the optimisation can be based on Bayesian evidence (BE) starting from initial guesses, such as those suggested by Rasmussen [Rasmussen, (1996)].
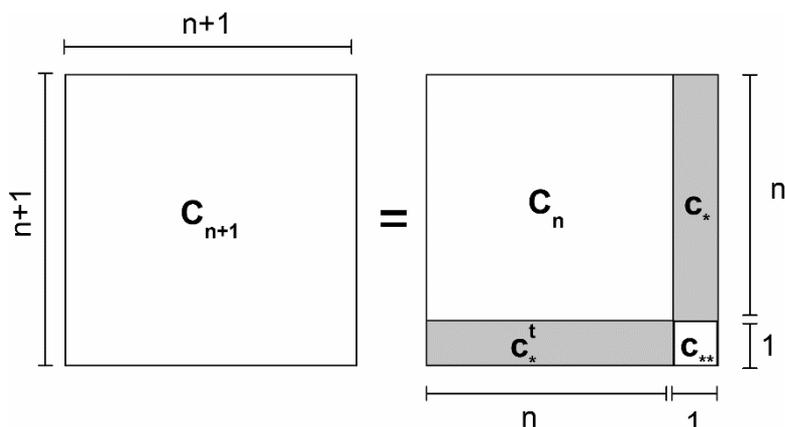
Figure 3.6. Decomposition of the (n+1)-dimensional covariance matrix C_{n+1}. C_n is the covariance matrix of the training set; c_*^t and c_* are the vectors of the covariance between training and test molecules; c_{**} is the auto covariance of the test molecule.

Compared to other machine-learning methods, such as artificial neural networks (ANN), GPR was reported to suffer less from overfitting. On the other side, GPR does not scale well to the number of molecules in the dataset, as the computational cost increases sharply.

3.4 *Consensus* **modelling**

Consensus modelling is an approach based on the combination of the predictions provided by different QSAR models. Based on the consideration that every QSAR model implicitly has some flaws that determine its prediction errors, the assumption of *consensus* modelling is that the weaknesses of one model will be counterbalanced by the strengths of the others and vice versa. The final goal can be the enhancement of prediction accuracy or the broadening of the applicability domain (AD). Previous studies have already shown the beneficial effects of *consensus* modelling for ecotoxicological endpoints [Mansouri et al., (2013); Cheng et al., (2012); Lozano et al., (2010); Stoyanova-Slavova et al., (2014)].

In this study, two strategies to *consensus* modelling were followed, which will be referred to as 'Strict' and 'Loose'. The 'Strict' approach considers only the predictions for chemicals that are inside the applicability domain of all the models and provides a prediction that is the mean of the individual model estimates. Consequently, the applicability domain of the 'Strict' *consensus* model will be narrow because it is the intersection of the applicability domains of the individual

models. The 'Loose' strategy considers the predictions for molecules that are inside the AD of at least one model. The prediction of a compound is the mean of the estimates of the models that the compound falls inside the AD of. If a chemical is inside the AD of only one model, the only available prediction is used. It is evident that the 'Loose' approach provides predictions for a larger percentage of compounds compared to the 'Strict' *consensus* model (i.e. it has a broader AD). In fact, the AD of the 'Loose' *consensus* model is the union of the ADs of the individual models. An illustration of the *consensus* strategies is provided in Figure 3.7 for a case with two models.
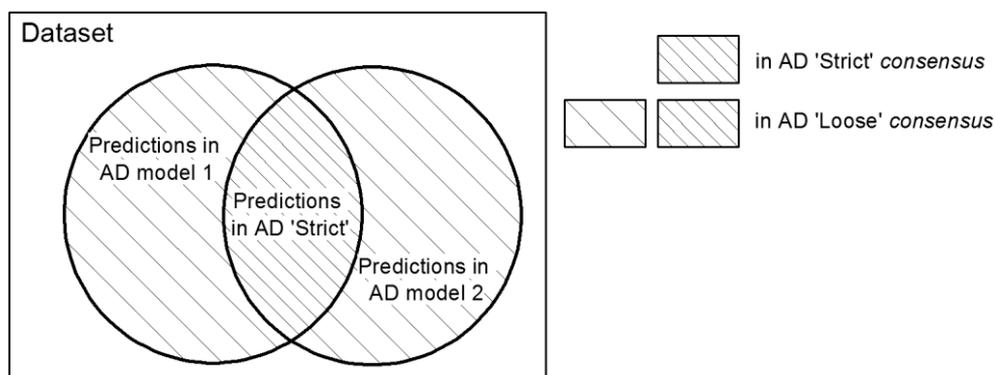
Figure 3.7. Depiction of the applicability domains of the 'Strict' and 'Loose' consensus models for a case with two available QSAR models.

3.5 Applicability domain

The application of any QSAR model to new compounds is implicitly limited by the fact that the model derives from a particular finite set of molecules: the training set. The calibration of a mathematical model is based on the interpolation of the training data and therefore the model can be assumed valid in this space. In other words, there is no guarantee that the model will retain its behaviour when applied outside its training space (extrapolation). It is therefore fundamental to identify the chemical space where a QSAR model is assumed to provide reliable (accurate) predictions, i.e. its applicability domain (AD). It should be recalled that the characterisation of the AD is required by the OECD principle 3 [OECD, (2007)].

The applicability domain can be defined either *a priori* regardless of model descriptors, for instance if the model was developed from a specific chemical class, or it can be assessed *a posteriori* on the basis of the molecular descriptors of the

training set. In the latter case, a variety of approaches was proposed [Sahigara et al., (2012)]. In this study, three main approaches were used, all of them of the *a posteriori* type.

The first approach is based on the leverage, which is the distance of each chemical from the centre of the model space. The leverage matrix (or hat matrix) is calculated from the values of model descriptors as:

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X^t X}\right)^{-1}\mathbf{X^t} \tag{32}$$

where $\mathbf{X}$ is the matrix of model descriptors. The leverage values associated to the compounds are the diagonal elements of this matrix, h_{ii}, and the larger the leverage the more distant a compound from the centre of the model. Hence, molecules with high leverage can be considered different and, in some circumstances, outliers. On one side, training molecules associated with high leverage may be discarded because they can considerably affect the parameters of the model; on the other side, test molecules with high leverage are likely to be associated with unreliable predictions (low accuracy) as a consequence of their difference from the training molecules (extrapolation). The average value of the leverages for molecules in the training set is often used to derive a threshold that works as an upper bound. In this work, five times the mean leverage was considered as an upper threshold. Molecules in the training set associated with a leverage greater than the threshold were regarded as outliers and removed prior to fitting and cross-validating the model. In similar manner, test molecules with leverage values greater than five times the mean value (recalculated on the reduced training set) were considered outliers and not predicted.

The second approach is known as the bounding box. This simple approach takes the minimum and maximum values in the training set for each model descriptor and uses them as lower and upper bounds, respectively. A test molecule is regarded to fall inside the applicability domain of the model if the values of its descriptors are within the range of the training set.

The third approach belongs to the *k*NN similarity-based approaches [Sahigara et al., (2012); Sahigara et al., (2013); Sheridan et al., (2004)]. These methods use a similarity or distance measure to evaluate the resemblance between compounds. Typically, a threshold value is chosen and compared with the similarity (or distance) of each molecule from the others in the training set. If the similarity is greater than the threshold (or if the distance is lower than the threshold), the molecule is considered inside the AD because it has enough similar training chemicals. It is

evident that these similarity-based approaches better characterise the molecules distribution and seem especially useful in combination with similarity-based modelling methods, such as *k*NN. Compared to the previous two approaches, the area inside the AD is not necessarily continuous and can present hollow spaces. In this work, the average distance from the *k* nearest neighbours was compared to a fixed threshold used as upper bound. An average distance larger than the threshold is associated to low similarity and therefore symptom of unreliable (inaccurate) prediction.

An illustration that highlights the differences among these approaches is provided in Figure 3.8 for two simulated datasets. For *k*NN based on the average distance, five nearest neighbours were considered (*k* equal to five). The leverage and bounding box approaches include empty areas in the AD, which is not the case with the *k*NN similarity-based approach.

Further details about the AD approach used in the final models are given in the scientific publications in the appendices I, II and III.

3.6 Model validation

The validation of a mathematical model is a procedure used to test and determine its robustness and predictivity. The robustness refers to the sensitivity of the parameters of the model to changes in the training data. The predictivity is related to the accuracy of the predictions provided by the model for chemicals not used to develop the model itself. According to the guidance on the five OECD principles the robustness should be evaluated by means of internal validation techniques and the predictivity by means of external validation [OECD, (2007)].

Internal validation estimates the model robustness by perturbing the training set. In practice, one or more chemicals are removed from the training set and assigned to a set called *evaluation set*. A partial model is calibrated on the reduced training data and used to predict the response of the chemicals in the evaluation set. Real (experimental) and predicted response values are then used to calculate statistical parameters (see paragraph 3.7). The different approaches for internal validation, also called cross-validation (CV), differ for the way molecules are assigned to the reduced training and evaluation sets. The most common methods are:
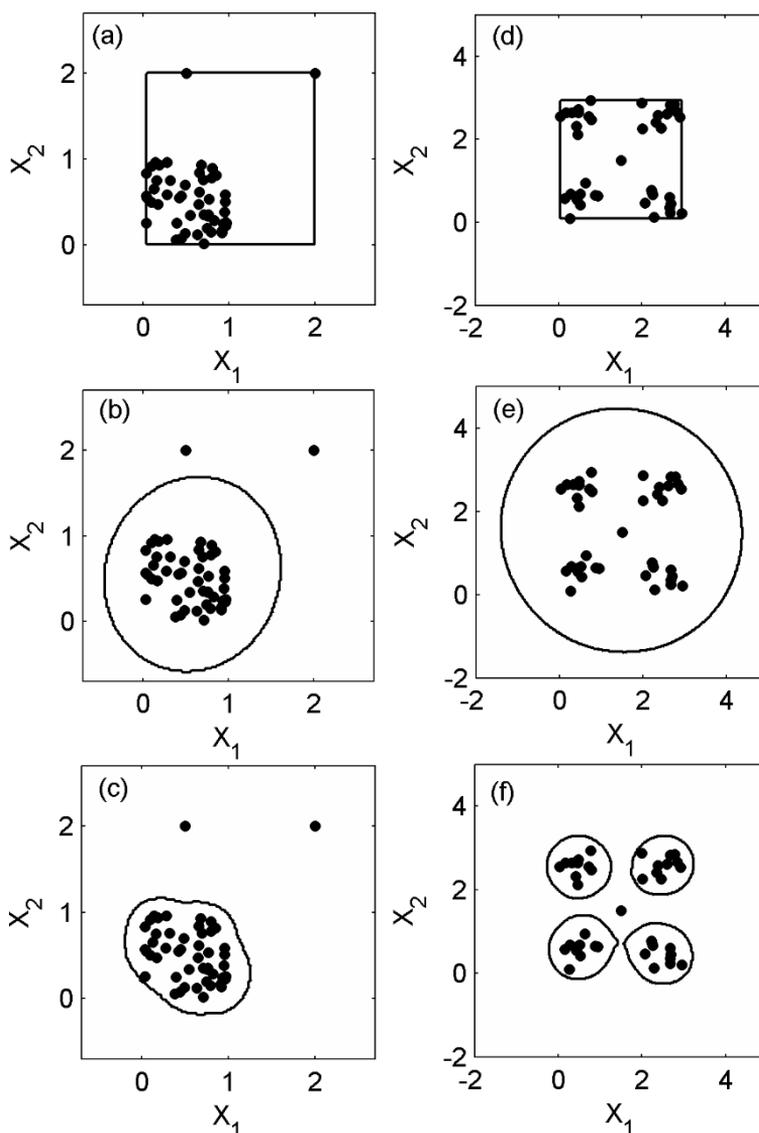
Figure 3.8. Illustration on two simulated datasets of three approaches to evaluate the applicability domain: bounding box (a)-(d); leverage (b)-(e); average distance from five nearest neighbours (c)-(f). Circles represent training compounds; lines delimit the space within the AD.

- **Leave-one-out**: one chemical at a time is assigned to the evaluation set. The procedure is iterated until all chemicals have been excluded once. With this method, n partial models are constructed, n being the number of chemicals in the training set. The approach represents a small perturbation of the dataset and was reported to give optimistic estimates unless the number of chemicals in

the training set is small [Golbraikh and Tropsha, (2002); Esbensen and Geladi, (2010)].

- **Leave-more-out**: this method is in principle analogous to the leave-one-out but more chemicals are assigned to the evaluation set simultaneously. Typically, the training set is divided into G groups that are removed one at a time. Each compound is excluded only once, i.e. it is assigned to only one evaluation group. It is evident that leave-one-out corresponds to leave-more-out with G equal to n.

- **Random splitting**: a certain percentage of molecules is randomly selected and assigned to an evaluation set. Since the results may strongly depend on the particular evaluation-reduced training set pair, the procedure is usually repeated several times. The main difference with the previous two methods is that molecules can be excluded several times or never.

- **Bootstrap**: this method is similar to the random splitting but it is performed with repetitions. This means that each time a training set consisting in n compounds is generated, in which some molecules are present more than once (repeated). The absent molecules constitute the evaluation set. The procedure is repeated several times, usually thousands of times, in order to have reliable estimates.

The external validation consists in predicting the response of compounds never considered for the calibration of the model nor the selection of molecular descriptors, for which the experimental values are known. Since often it is not feasible to measure new data, a test set can be generated before model development takes place. The selection of the compounds to be assigned to the test set can be done in different ways.

- **Response-based selection:** the splitting training-test set is based on the values of the response. Typically, the dataset is sorted based on the response value and molecules uniformly distributed are assigned to the test set. This method is used in order to assure equal coverage of the response domain by training and test sets.

- **Descriptor-based selection:** the splitting training-test set uses some kind of rational sampling method that considers the descriptors values. Examples of such methods are Kennard-Stone algorithm [Kennard and Stone, (1969)], the distance-based optimal design [Marengo and Todeschini, (1992)] and

clustering methods in general. Unlike the response-based, this method aims at assuring equal coverage in the descriptors space. A problem with this selection approach emerges if a subsequent variable selection is carried out: the splitting training-test set is performed on all the descriptors but the final model will comprise only a subset and the mutual relationships between chemicals can change during this dimensional compression.

- **Random selection**: compounds assigned to the test set are selected in a random manner. This is one of the most commonly used methods when a large number of chemicals is available because it is perceived free of 'selection bias' and does not present the problem of the descriptor-based approaches in case of variable selection. Since no information about the coverage both in the descriptors and in the response domain is considered, it can occur that the training and test sets are different. It is therefore extremely important to check if test chemicals fall inside or outside the applicability domain of the model.

The statistical parameters derived from the validation procedures are described in paragraph 3.7.

In this study, internal validation was performed by means of the leave-more-out method with five evaluation groups (five-fold cross-validation) and the random splitting 80%-20% training- evaluation set repeated 1000 times. Test sets comprising approximately 20% of the chemicals in the dataset were always generated by means of random selection. The test sets were never used during variable selection and model calibration: they were only submitted to the final models in order to test their predictivity.

3.7 Statistical parameters for regression diagnostic

The validation procedures described in the previous paragraph allow the calculation of statistical parameters used to quantify the ability of the model to fit the training data, withstand changes in the training data (robustness) and accurately predict the analysed property for molecules not used during the development of the model (predictivity). These parameters rely on the comparison of the real experimental response values with those calculated or predicted from the model. In this thesis, the terms 'calculated' and 'predicted' will be used with reference to fitting and validation (both internal and external), respectively.

3.7 Statistical parameters for regression diagnostic

The coefficient of determination was used to evaluate the performance of the model in fitting, internal and external validation. Equation 33 shows the formula for the calculation of the coefficient of determination in fitting (referred to as R^2) and cross-validation (referred to as Q_{cv}^2).

$$R^2/Q_{cv}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS} \tag{33}$$

where y_i is the experimental response of the *i-th* molecule and $\bar{y}$ is the mean response of the training chemicals. $\hat{y}_i$ is the response calculated when the *i-th* molecule was included in the training set for the fitting case (R^2) or predicted when the *i-th* molecule was part of the evaluation set for cross-validation (Q_{cv}^2). The quantities at the numerator and denominator are the residual sum of squares (*RSS*) and the total sum of squares (*TSS*) and represent the amount of variance not explained by the model (*RSS*) with respect to the total variance (*TSS*). It should be recalled that the *RSS* is the quantity being minimised by the OLS regression method (paragraph 3.3.1.1).

For the calculation of the coefficient of determination in external validation (Q_{ext}^2) the F3 function suggested by Consonni et al., (2009) was used. The formula (equation 34) presents a slight difference with respect to the R^2 and Q_{cv}^2 due to the fact the number of chemicals in the training and test sets is also considered:

$$Q_{ext}^2 = 1 - \frac{\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)/n_{ext}}{\sum_{j=1}^{n_{tr}}(y_j - \bar{y})/n_{tr}} = 1 - \frac{PRESS/n_{ext}}{TSS/n_{tr}} \tag{34}$$

where y_i is the experimental response value for the *i-th* molecule in the test set, $\hat{y}_i$ is its predicted response value, y_j is the experimental response value for the *j-th* molecule in the training set, $\bar{y}$ is the mean response of the training chemicals, n_{ext} and n_{tr} are the number of chemicals in the test and training set, respectively. The sum of squares of the residuals of the test set is referred to as predictive error sum of squares (*PRESS*)

An additional parameter used to quantify the average error of the model is the root mean square error (*RMSE*). The *RMSE* can be calculated from the residual sum of squares obtained in fitting, cross-validation (*RSS*) and external validation (*PRESS*). The corresponding *RMSE* is often referred to as root mean square error in calculation (*RMSEC*) for fitting, root mean square error in cross-validation (*RMSECV*) for internal validation and root mean square error in prediction (*RMSEP*) for external

validation. The formula is the same for the three cases (*RMSEC*, *RMSECV* and *RMSEP*), the difference being the nature of the *RSS* used:

$$RMSEC(V) = \sqrt{\frac{RSS}{n_{tr}}} \quad (35) \qquad RMSEP = \sqrt{\frac{PRESS}{n_{ext}}} \quad (36)$$

where n_{tr} and n_{ext} are the number of chemicals in the training and test set, respectively.

In this study, the coefficient of determination in cross-validation (Q_{cv}^2) was used as fitness function, i.e. as parameter to optimise, during variable selection performed by means of genetic algorithms (GAs) and reshaped sequential replacement (RSR).

3.8 Analysis of data structure: principal component analysis

Multivariate data analysis techniques can be used to investigate the structure of the data. These techniques are useful to highlight the presence of clusters or outliers, analyse relationships between molecules and molecular descriptors, reduce data dimensionality, and for the development of mathematical models.

One of the most commonly employed methods is principal component analysis (PCA). PCA is an unsupervised multivariate analysis technique invented in 1901 by Karl Pearson [Pearson, (1901); Jolliffe, (2005)]. PCA operates a rotation of the data in a new space defined by the so-called principal components. Each principal component is a linear combination of the original descriptors and the coefficients associated to each principal component are called *loadings*. Principal components are orthogonal, so they carry independent information, and are defined in such a way that the first component lies along the direction of maximum variance of the data, the second component lies along the second direction of maximum variance, and so on. Therefore, principal components are ranked according to the amount of information (variance) they explain, which is quantified by the eigenvalues. Typically, the last components explain little information and they are often associated with noise. From this consideration, it emerges that PCA allows separating 'good' information from noise by retaining only few components.

The calculation of principal components implies the diagonalisation of the covariance or the correlation matrix of the data and outputs the loadings and the

eigenvalues matrices. The scores, calculated from the original data matrix and the loadings matrix, are the coordinates of the samples in the new space defined by the principal components.

In this study, principal component analysis (PCA) was used to analyse the data in the space of the molecular descriptors included in the final models. These investigations were carried out to (a) evaluate the relationships between molecules and descriptors, and (b) analyse if patterns emerged with respect to toxicity. This was particularly useful for models based on regression methods that did not provide functional relationships, such as kNN and GPR.

3.9 Software

Workflows designed in KNIME [Berthold et al., (2007)] by the author were used to extract the relevant data from the databases and process them. The canonicalisation of SMILES strings was carried out by means of the KNIME node of OpenBabel [O'Boyle et al., (2011)]. The OASIS Database manager [Nikolov et al., (2006)] was used on the *Pimephales promelas* dataset to retain only organic compounds, apply the dissociation converter and compare the SMILES strings in the dataset and in the OpenTox database. DRAGON [*DRAGON 6,* (2012)] was used to calculate molecular descriptors and apply unsupervised variable reduction. Hashed binary fingerprints were generated by means of in-house software. Variable selection, model calibration and validation were carried out in MATLAB [*MATLAB*, (2012)] by means of in-house routines and toolboxes. LibSVM [Chang and Lin, (2011)] and GPML [Rasmussen and Williams, (2006)] toolboxes were coupled with the genetic algorithms toolbox in MATLAB and used to calibrate models by means of support vector regression (SVR) and Gaussian process regression (GPR), respectively. ChemProp [*ChemProp*, (2013)] was used to retrieve the datasets of the models proposed by Kühne et al., (2013) and Russom et al., (1997) for *Daphnia magna* and *Pimephales promelas*, respectively. The validation of the model for *Daphnia magna* proposed by Kühne et al., (2013) on the validation subsets presented in paragraph 2.3.2 was carried out in ChemProp. Marvin was used for drawing, displaying and characterising chemical structures and substructures [*Marvin*, (2012)].

3.9.1 Leadscope Enterprise™

The *Pimephales promelas* dataset was modelled also by means of Leadscope Enterprise™ software [*Leadscope Enterprise*^{TM}]. Leadscope Enterprise™ uses fragments, scaffolds and molecular properties as descriptors for the calibration of partial least squares (PLS) regression models. The selection of molecular descriptors was carried out by means of the proprietary algorithm: unsupervised variable reduction removes fragments present in less than two or in all compounds; supervised selection is based on t^2 test and analysis of the residuals. The fitness function used by the software to assess the quality of the models is the residual sum of squares (*RSS*) in cross-validation.

The validation of the models in Leadscope Enterprise™ was carried out in a slightly different way. First, models were calibrated using all the molecules in the dataset (908 chemicals) and internally validated by means of random splitting 50%-50% repeated ten times. The best model was also validated by means of the 5 x 2 x 50% off strategy [Ringsted et al., (2009); Jensen et al., (2011)]. To this end, the dataset was randomly split into two sets A and B (50%-50%) but keeping the same coverage in the response domain. A model from each half was built and used to predict the other half. The whole procedure was repeated five times. As a result, ten sub-models were created and the average of their performance was used as a measure of the performance of the final model. With this type of validation, the ten sub-models are constructed following the same procedure as the one followed for the final model, but variable selection is carried out from scratch for every sub-model. This assures that the final model is completely independent from the outcomes of the ten sub-models. It can be said that this strategy validates not only the final model based on a specific set of descriptors, but also the whole variable selection procedure.

The evaluation of the applicability domain (AD) in Leadscope Enterprise™ is based on two criteria. A test molecule is considered outside the AD of the model if it has none of the fragments included in the model or there are no molecules in the training set with a distance score smaller than 0.60. The distance score is calculated combining the distances obtained using all fragments in the training set and only the fragments included in the model, as:

$$d_i = \sqrt{\overline{d}_{i,tf}^2 + \overline{d}_{i,mf}^2} \qquad\qquad (37)$$

where d_i is the overall distance score, $\bar{d}^2_{i,tf}$ is the squared average Tanimoto distance between the *i-th* test molecule and the training molecules using all fragments present in the training set; $\bar{d}^2_{i,mf}$ is the squared average Tanimoto distance between the *i-th* test molecule and the training molecules using only the fragments included in the model. The distance value ranges from zero (identical molecules) to 1.414 (null similarity).

CHAPTER 4

Results on Daphnia magna

'No amount of experimentation can ever prove me right;
a single experiment can prove me wrong'

Albert Einstein (paraphrased)

4.1 Explorative analysis

The preparation of data on acute toxicity towards *Daphnia magna* lead to the definition of a dataset comprising 546 organic molecules (paragraph 2.3.1), which will be referred to as MICHEM dataset. According to current theories on aquatic toxicity, all chemicals exert narcotic effects, which arise from unspecific interactions with biological structures. Several QSAR models showed that narcosis correlates well with parameters describing the hydrophobicity, the *n*-octanol-water partition coefficient (LogP) above all [Veith et al., (1983); Van Leeuwen et al., (1992); Bearden and Schultz, (1997); von der Ohe et al., (2005)]. Some chemicals are more toxic than what expected on the basis of narcosis models because they are able to additionally interact with critical macromolecules [Newsome et al., (1996)]. Theoretical bases for these arguments consider that chemical toxicity derives from a combination of penetration in the biological media and interaction with the site of action [McFarland, (1970); Bearden and Schultz, (1997)]. In order to check this situation in the analysed data, the LogP (calculated) was plotted against the experimental toxicity (Figure 4.1). The baseline toxicity of the model defined in Kühne et al., (2013) was superimposed. In Figure 4.1 the LogP was calculated by means of the KOWWIN software [*KOWWIN*, (2010)] for consistency with the baseline model.

It is apparent from Figure 4.1 that for a subset of chemicals there is a linear relationship between LC_{50} and LogP, which is well fitted by the narcosis model of Kühne et al., (2013). The consideration that all chemicals act at least as narcotics is

4.1 Explorative analysis
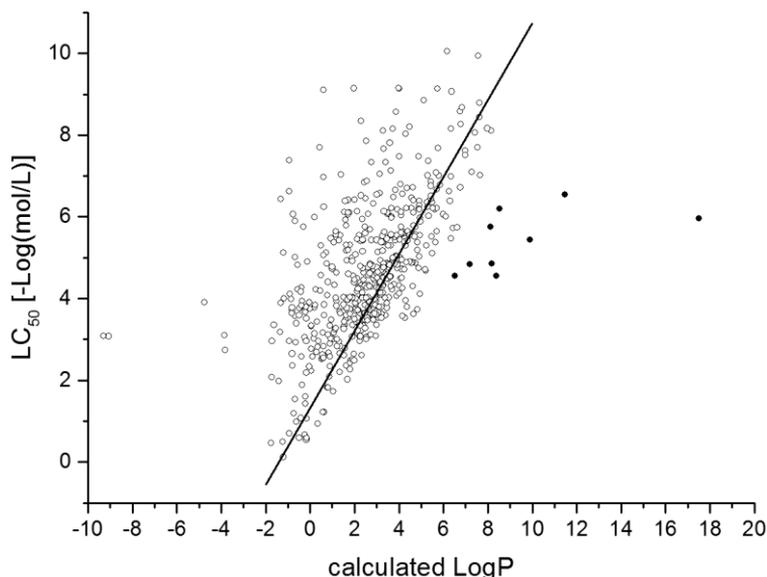
Figure 4.1. Calculated LogP versus LC$_{50}$ 48 hours (-Log(mol/L)) towards Daphnia magna. Solid line: baseline toxicity defined in Kühne et al., (2013). Black circles: nine molecules whose LC$_{50}$ are lower than 2.5 Log units from the baseline.

also exemplified by the fact that almost no compound is less toxic than what expected on the basis of the narcosis model. There are, however, nine compounds whose LC$_{50}$ values are considerably lower (more than 2.5 log units) than what predicted by the baseline (black circles in Figure 4.1). These large deviations from the baseline occur at medium-high LogP values. Three hypotheses were proposed to explain this phenomenon. On one hand, this could be a consequence of the large variability of experimental data, as highlighted in paragraph 2.2. The second hypothesis concerns the accuracy of the calculated LogP values. Hence, the ECHA registration database [*ECHA*], the OECD QSAR Toolbox [*The OECD QSAR Toolbox*, (2013)] and the KOWWIN dataset were analysed searching for experimental LogP values, which were found for five chemicals. Therefore, the LogP was also predicted by means of the *MLOGP* model [Moriguchi et al., (1992); Moriguchi et al., (1994)] implemented in DRAGON in order to have another computational estimate. LogP values are collated in Table 4.1. Their comparison for the five chemicals with available experimental values indicates that KOWWIN tends to overestimate LogP, whereas *MLOGP* tends to underestimate it. The third hypothesis considers the likelihood of inaccurate measurements or LC$_{50}$ determinations. One possibility could be the use of nominal concentrations, which may result in overestimation of lethal concentrations (lower values in the negative logarithmic scale used in Figure 4.1), especially for

chemicals with low water solubility. Therefore, water solubility data were searched for in the OECD QSAR Toolbox and the datasets of WATERNT [*WATERNT*, (2010)] and T.E.S.T. software [Martin et al., (2012)]. When no experimental data was available, predictions were taken by the two software programs (Table 4.1). With the exception of bis(2-ethylhexyl) adipate and the prediction by T.E.S.T. for heptylnonylphtalate, all other water solubility data were greater than the lethal concentrations (LC_{50}) in negative logarithm of molarity, thus corroborating the hypothesis of inaccurate measurements. Based on these considerations, if LogP estimates by KOWWIN are replaced with the experimental LogP and *MLOGP* values, and LC_{50} is replaced by water solubility (when water solubility is higher in negative logarithm of molarity), eight chemicals would considerably approach the baseline. Bis(2-ethylhexyl) adipate would, instead, still lie considerably below. One last possibility considers the chance of inaccurate LC_{50} determination from the dose-response curves.

Given the situation depicted in Figure 4.1, it is clear that a linear model based only on LogP cannot accurately account for the toxicity of all chemicals in the dataset. Hence, it is necessary to either split chemicals into clusters, for instance narcotics and not, and develop local regression models, or include more molecular descriptors and try to develop a unique model in a higher dimensional space where regression can be carried out more accurately. This second route was followed in order to avoid additional steps (e.g. initial clusterisation) that would render the modelling strategy more complex.

Table 4.1. LogP, water solubility and LC_{50} for the nine molecules with toxicity lower than 2.5 logarithmic units from the baseline in Figure 4.1. Water solubility and LC_{50} are reported as – Log(mol/L). Hyphens are used for lacking information.

CAS-RN	Name	LogP KOWWIN	*MLOGP*	Exp. LogP[a]	Water sol.	LC_{50}
35723-83-2	N,N-Diisotridecyl isotridecanamine	17.50	9.15	-	12.00[b]; 8.72[c]	5.96
57157-80-9	Diisotridecylamine	11.47	6.73	-	10.70[b]; 6.85[c]	6.54
5964-62-5	Di(4-(4-hydroxy-2-methyl-5-isopropylphenyl azo)phenyl)sulfone	9.89	4.48	-	11.33[b];6.03[c]	5.44
79-94-7	4,4'-(Methylethylidene)bis (2,6-dibromophenol)	7.20	5.26	5.90	5.69[d]	4.84
103-23-1	Bis(2-ethylhexyl) adipate	8.12	4.74	8.94	5.68[d]	5.75
117-81-7	Bis(2-ethylhexyl) phthalate	8.39	5.43	[7.14,7.94]	6.16[d]	4.55
96829-58-2	Orlistat	8.19	5.44	-	9.51[b];6.41[c]	4.85
1740-19-8	Dehydroabietic acid	6.52	4.55	4.80	4.66 [d]	4.56
19295-81-9	Heptylnonylphtalate	8.54	6.34	8.10	8.93[b];5.88[c]	6.19

[a] experimental LogP; [b] predicted by WATERNT; [c] predicted by T.E.S.T.; [d] experimental.

4.2 Calculation of molecular descriptors and data setup

The SMILES strings of the molecules were imported in DRAGON in order to calculate 0D, 1D and 2D molecular descriptors. Some molecular descriptors were neglected:

- Descriptors from the drug-like block because they were supposed not to be relevant for modelling acute aquatic toxicity.

- Intrinsic state and electrotopological state pseudo-connectivity indices ([Pogliani, (2000); Pogliani, (2004)], Estrada-like indices (*EE_M(w)*), coefficient sum, average coefficient and logarithmic coefficient sum of the last eigenvector (*VE1_M(w), VE2_M(w), VE3_M(w)*), Randic-like, normalised Randic-like and logarithmic Randic-like eigenvector based-index (*VR1_M(w), VR2_M(w), VR3_M(w)*) [Balaban et al., (1991)], Ghose-Crippen-Viswanadhan octanol-water partition coefficient (*ALOGP*) and its squared value (*ALOGP2*) [Ghose and Crippen, (1986); Viswanadhan et al., (1993); Ghose et al., (1998)] and molar refractivity (*AMR*) [Ghose and Crippen, (1987)] because they may not be computable on some structures.

- Baseline toxicity to daphnia, fish and algae because they are trivially defined from the *MLOGP* (ρ = -0.99).

Constant, near-constant and descriptors with missing values were removed in DRAGON, leading to a set of 2187 molecular descriptors, belonging to 18 logical blocks. MICHEM dataset was then randomly split into training and test sets comprising 436 and 110 chemicals, respectively. The training set was used for the selection of relevant molecular descriptors and the calibration of QSAR models, whereas the test set was only submitted to the final models to assess their predictivity.

4.3 Descriptor selection and model calibration

Unsupervised variable reduction based on the pairwise coefficient of correlation was carried out prior to using supervised variable selection methods (GAs and RSR) with OLS and PLS regression. This was done because linear regression models suffer from the presence of correlated descriptors (multicollinearity).

Two different strategies for descriptor selection were followed with genetic algorithms (GAs) and reshaped sequential replacement (RSR) due to the different nature of these methods. Indeed, the available version of GAs attains to the strategy

of Leardi and González, (1998), which is based on the frequency of selection of molecular descriptors over several independent runs; RSR is instead based on a single run carried out until convergence of the population. Previous results showed that RSR tended to converge to the same (or very similar) final population regardless of the initial state. Hence, the following procedures were followed:

- **RSR**: only one run was carried out on all the available molecular descriptors. The tabu list and the roulette wheel were activated with the aim to exclude descriptors uncorrelated to the response and generate an initial population comprising the most promising descriptors.

- **GAs**: several independent runs were carried out on each logical block of molecular descriptors from DRAGON (18 blocks in total). The most frequent descriptors from each block were selected and merged to form a pool of approximately 200 descriptors. GAs were run again on this set. If, at the end of this step, the differences in the frequencies of selection were not large, a reduced subset comprising approximately 100 descriptors with the largest frequencies was extracted and input to GA again. Eventually, all the combinations of a small subset of the most frequently selected descriptors were generated and evaluated by means of an all subset models selection strategy.

So far, only ordinary least squares (OLS) regression and principal component regression (PCR) were implemented in the RSR toolbox. Therefore, RSR was only used to derive OLS models, which were compared with those obtained by means of GAs. On the other side, GAs were used to calibrate models by means of OLS, PLS and kNN regression.

As aforementioned, different approaches for the assessment of the applicability domain were used. The leverage approach was used as detailed in paragraph 3.5 when dealing with both OLS and PLS regression. With kNN as the modelling method, the applicability domain was assessed by comparing the average distance from the k nearest neighbours with a fixed threshold. The optimal value of the distance threshold was chosen in cross-validation for each value of k. In order to avoid the selection of extreme distance thresholds, a constraint was introduced on the maximum percentage of compounds regarded outside the applicability domain. A percentage equal to 40% was chosen as reasonable trade-off between performance and applicability. When using the Soergel distance measure, the descriptors were scaled in the range [0,1] because this metric requires descriptors to be positively defined (paragraph 3.3.2.1).

Table 4.2. Strategies used to derive regression models for Daphnia magna.

Variable reduction	Variable selection	Regression method	AD approach
correlation	GAs	OLS	leverage
correlation	GAs	PLS	leverage
-	GAs	*k*NN	average distance
correlation	RSR	OLS	leverage

It is evident that if a test compound has a value for a descriptor lower than the minimum value in the training set, the scaled value would be negative. In order to avoid this situation, which would not be consistent with the distance requirements, the bounding box approach was used as a preliminary check: test compounds with values of descriptors outside the range of the training set were considered out of AD. A summary of the strategies is given in Table 4.2.

4.4 Summary of results

The best models obtained with OLS, PLS and *k*NN are reported in Table 4.3. Linear models based on OLS and PLS regression had similar statistics in fitting and cross-validation (R^2 between 0.61 and 0.71 and Q^2_{cv} between 0.59 and 0.68), but showed lower predictive ability on the test set (maximum value of Q^2_{ext} equal to 0.54). The percentage of outliers in the training set using the leverage approach was always low (2%-4%), whereas a higher percentage was detected in the test set (up to 10%). OLS models obtained from GAs and RSR shared some descriptors, thus indicating a certain degree of accordance between the outcomes of the two different selection methods. The statistics were comparable but RSR models included less molecular descriptors. This could be an effect of the more thorough search carried out by RSR compared to GAs. The drawback, on the other hand, was that RSR was more sensitive to overfitting. The RSR model with eight descriptors was the first one to fulfil all the evaluation functions (paragraph 3.2.2.2) with default settings. The performance in cross-validation was not high but this model had the most balanced statistics. Interestingly, the RSR model with ten descriptors was the only one not including *MLOGP*.

The results were considerably different with *k*NN regression. Indeed, *k*NN models had considerably higher statistics in fitting, internal and external validation, but also gave larger percentages of compounds outside the applicability domain. This was a consequence of the strict criterion adopted during model calibration, which allowed up to 40% of compounds to be regarded out of AD. In order to check whether

a stricter AD criterion would give beneficial effects also on OLS and PLS, the corresponding models were validated again using three times (instead of five times) the average leverage as threshold. The outcome of this attempt was negative, indeed in most cases the performance in internal and external validation decreased.

KNN models were calibrated by trying three distinct distance measures. The differences among these metrics were highlighted by the fact that *MLOGP* was the only descriptor in common to the three models. The models based on the Euclidean and Mahalanobis distance shared also another descriptor (*GATS1p*). By analysing the results, it emerged that the Euclidean distance, which is often the preferred choice, actually gave the lowest statistics both in fitting, internal and external validation. The model based on the Soergel distance had satisfactory performance in cross-validation (Q_{cv}^2 equal to 0.74) and external validation (Q_{ext}^2 equal to 0.70), but this was achieved at the expense of a large percentage of test molecules regarded out of AD (45%). Hence, the best results were provided by kNN based on the Mahalanobis distance. The most evident characteristic of the Mahalanobis distance is the inclusion of the information on the covariance in the data, which seems beneficial. The kNN model based on the Mahalanobis distance, chosen as the best result, is analysed in the following paragraphs.

Table 4.3. Best models obtained for LC_{50} towards Daphnia magna with different regression methods.

Method[a]	p[b]	Scaling	distance	AD	LVs / k[c]	R^2	Q_{cv}^2	Q_{ext}^2	% out AD train[d]	% out AD test[e]
GA_OLS	12	-	-	$5*\bar{h}$[f]	-	0.65	0.61	0.50	3%	8%
GA_OLS	21	-	-	$5*\bar{h}$[f]	-	0.71	0.66	0.50	2%	6%
RSR_OLS	10	-	-	$5*\bar{h}$[f]	-	0.69	0.68	0.50	4%	4%
RSR_OLS	8	-	-	$5*\bar{h}$[f]	-	0.61	0.59	0.53	3%	7%
GA_PLS	6	auto	-	$5*\bar{h}$[f]	2	0.61	0.60	0.41	3%	5%
GA_PLS	14	auto	-	$5*\bar{h}$[f]	2	0.66	0.65	0.54	3%	10%
GA_kNN	8	auto	Euclidean	1.125[g]	5	0.70	0.69	0.66	36%	29%
GA_kNN	8	auto	Mahalanobis	1.26[g]	3	0.78	0.78	0.72	38%	31%
GA_kNN	9	range	Soergel	0.161[g]	4	0.70	0.74	0.70	35%	45%

[a] GA_OLS: OLS by means of GAs; RSR_OLS: OLS by means of RSR; GA_PLS: PLS by means of GAs; GA_kNN: kNN by means of GAs; [b] number of model descriptors; [c] number of latent variables (LVs) for PLS or number of nearest neighbours (k) for kNN; [d] percentage of training compounds out of AD; [e] percentage of test compounds out of AD; [f] five times the average leverage; [g] threshold on the average distance from the k nearest neighbours.

4.5 Discussion of the kNN model

The kNN model based on the Mahalanobis distance calculated on eight molecular descriptors, hereinafter referred to as the MICHEM model, was selected

as the best result obtained and further analysed. The scientific publication in Appendix I details (a) the preparation of the dataset; (b) the calibration of the model and optimisation of the parameters (number of nearest neighbours, k, and threshold on the average distance from the nearest neighbours); (c) the validation procedures and corresponding statistics; (d) the investigation of the residuals; and (e) the interpretation of model descriptors. Here a summary is given. The values of number of nearest neighbours (k equal to three) and threshold on the average distance (1.26) were optimised in five-fold cross validation. The value of the distance threshold was selected with the aim to maximise the performance, given a constraint on the maximum percentage of molecules out of AD (40%). Molecules in the training set regarded as out of AD were not removed because they could be useful to predict future query compounds. It should be noted that the threshold on the average distance can be changed to tune the strictness of the criterion for the definition of the AD, where low values correspond to a stricter criterion and vice versa. The model had good performance in internal and external validation (Q_{cv}^2 and Q_{ext}^2 equal to 0.78 and 0.72, respectively). The balanced statistics on the training and test sets should indicate lack of overfitting. The eight selected molecular descriptors were topological polar surface area with N, O, S, P polar contributions (*TPSA(Tot)*), surface area of acceptor atoms from P_VSA-like descriptors (*SAacc*), Moriguchi octanol-water partition coefficient (*MLOGP*), reciprocal distance sum Randic-like index (*RDCHI*), number of nitrogen atoms (*nN*), atom-centred fragments of the type R-C(=X)-X / R-C#X / X=C=X (*C-040*) – X being either O, N, S, P, Se, halogens -, number of hydrogen atoms attached to an heteroatom (*H-050*) and Geary autocorrelation of lag one weighted by polarisability (*GATS1p*). These descriptors encoded information about lipophilicity, formation of hydrogen-bonds, polar surface area, polarisability, nucleophilicity and electrophilicity. The following paragraphs report additional investigations not included in the scientific paper in Appendix I.

4.5.1 Analysis of the residuals and neighbourhood behaviour

The analysis of the residuals carried out in the scientific publication in Appendix I justified the introduction of the threshold on the average distance by showing that increasing errors were associated with increasing average distances from the three nearest neighbours. An additional confirmation of this trend is shown in Figure 4.2, which reports the box-whisker plots of the standardised residuals calculated on ten bins of average Mahalanobis distance from the three nearest

neighbours. The range of the residuals, which are approximately centred on zero, tends to increase together with the average Mahalanobis distance. This is an indication that molecules with large average distances with respect to their neighbours are more likely to be associated with large residuals, hence their predictions are less accurate.
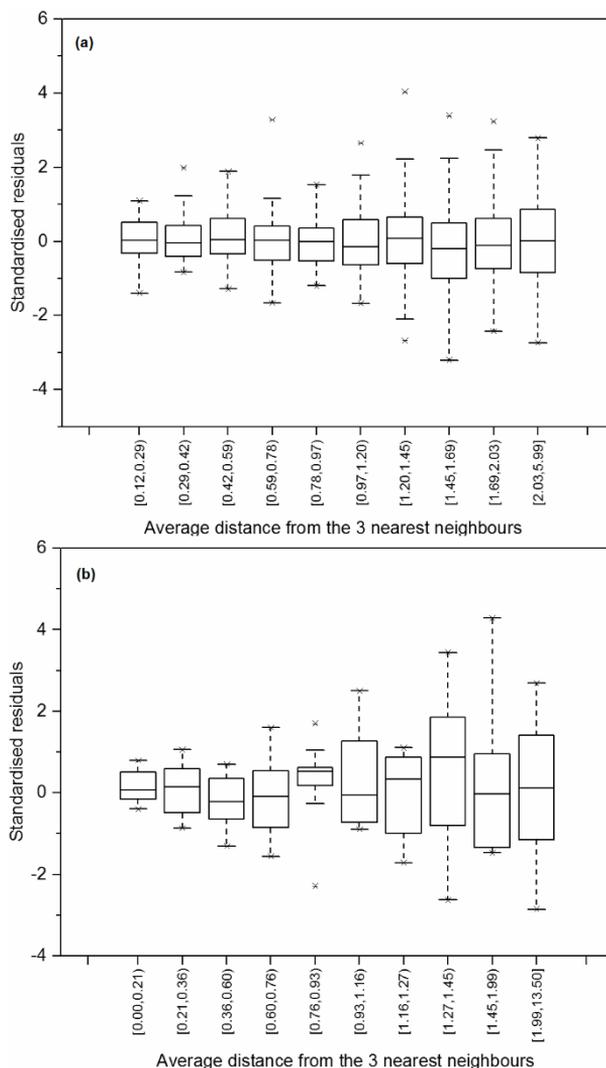

Figure 4.2. Box-whisker plots of the standardised residuals for 10 bins of the average distance from the three nearest neighbours. (a) Training set; (b) test set.

4.5 Discussion of the kNN model

QSAR analysis is based on the congenericity principle, which assumes that molecules that feature similar structures should possess similar activities, and changes in the structure are expressed by changes in the activities. This assumption is even more important when dealing with models based on similarity analysis, such as *k*NN. It is therefore essential to check that this condition is fulfilled by the data under analysis. The Patterson plot can be used to visually check if the data show a neighbourhood behaviour, i.e. if compounds close in the descriptors space possess similar property values [Patterson et al., (1996)]. The Patterson plot reports the distance between pairs of molecules on the x-axis *versus* the absolute difference of their experimental responses on the y-axis. Neighbourhood behaviour implies similar molecules, which are placed on the left-hand side of the plot, to have small differences between their experimental values, and vice versa. The plot can also reveal the presence of activity cliffs if pairs of molecules are placed in the top left corner (similar structures but different experimental values). Hence, the ideal situation for the application of QSAR analysis is when the lower right-hand side triangle is occupied. The Patterson plot for the training set is shown in Figure 4.3, where the distance (x-axis) was calculated as (1-similarity). It can be seen that compounds are distributed in the lower right-hand side triangle and that compounds with large differences between their experimental values of LC_{50} are placed on the right-hand side of the plot (low similarity). This seemed an indication that the dataset was suitable for the application of similarity-based QSAR analysis. However, Sheridan *et al.*, (2004) reported that the Patterson plots of large diverse data sets always show this pattern with substructure descriptors. They suggested the calculation of the Patterson ratio as an estimate of the strength of the neighbourhood behaviour in the dataset. The Patterson ratio is defined as:

$$P_r = \frac{\dfrac{\sum_{s=1}^{n-1}\sum_{t=s+1}^{n}|y_s - y_t|}{ntp}}{\dfrac{\sum_{s=1}^{n-1}\sum_{t=s+1}^{n}|y_s - y_t|\delta_d}{nsp}} \qquad \delta_d := \begin{cases} 1 \, if \, d_{st} \leq 0.3 \\ 0 \, if \, d_{st} > 0.3 \end{cases} \qquad (38)$$

where y_s and y_t are the response values of the *s-th* and the *t-th* molecule, respectively; δ_d is the Kronecker delta which is equal to 1 if the distance (d_{st}) between molecules *s* and *t* is smaller than 0.3 and 0 otherwise; *ntp* is the total number of pairs in the training set and *nsp* the number of pairs with distance smaller than or equal to 0.3. If the ratio is equal to one there is no neighbourhood behaviour, while the larger

the ratio, the stronger the neighbourhood behaviour. The Patterson ratio for the dataset based on the eight molecular descriptors of the *k*NN model is 2.59, thus confirming a neighbourhood behaviour. However, it should be noticed that the high similarity region of the Patterson plot (left-hand side) is quite sparsely populated.

In order to further analyse the relationships between chemical and toxicological similarities, for each molecule included in the test set, the average absolute difference of its experimental toxicity with respect to the three nearest neighbours and the average Mahalanobis distance from the three nearest neighbours were evaluated (Figure 4.4). The scatterplot confirms the considerations drawn from the Patterson plot on the training set: the lower the similarity (or likewise, the larger the distance), the larger the difference in the toxicity activity. Thus, it can be concluded that structural dissimilarity of compounds (molecular descriptors domain) corresponds to dissimilarity in toxicological activities (response domain). This can further justify the introduction of a threshold on the average distance from the three nearest neighbours, so that predictions based on dissimilar molecules are hindered because likely to be inaccurate.
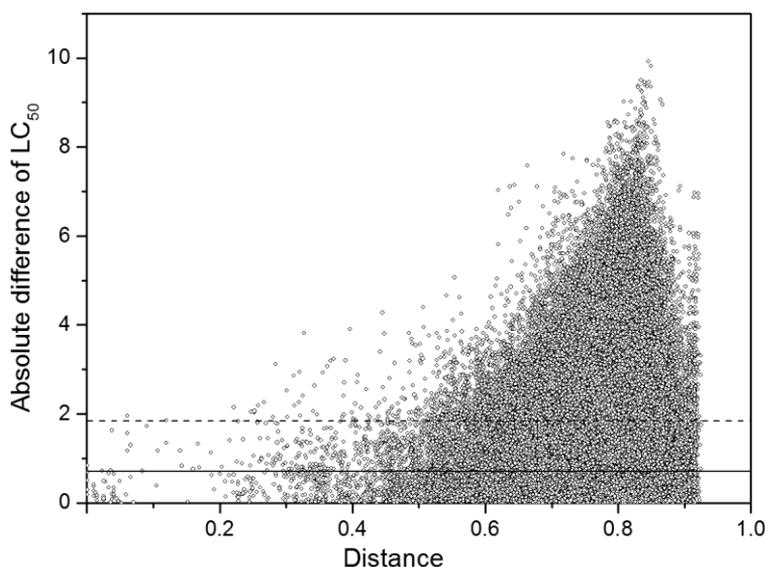

Figure 4.3. Patterson plot of the training set. X-axis: pairwise distance (1-similarity); y-axis: absolute difference between toxicity values. Dashed line: mean absolute difference of the toxicity on all pairs in the training set; solid line: mean absolute difference of the toxicity on pairs of molecules with distance smaller than or equal to 0.3.
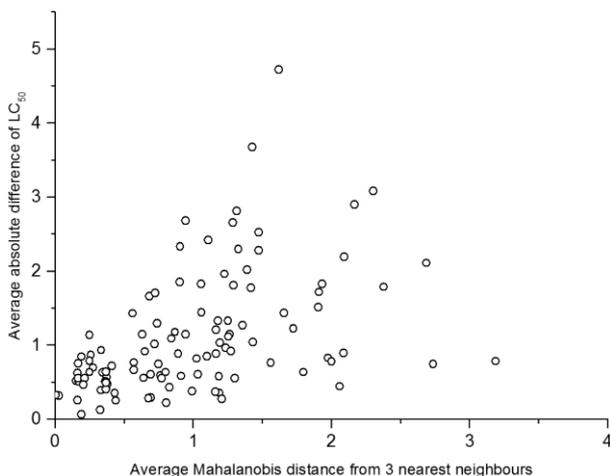
Figure 4.4. Average distance from the three nearest neighbours versus average absolute difference of toxicity from the three nearest neighbours for the test set. One molecule had an average Mahalanobis distance equal to 13.5; in order to make the plot more readable, the x-axis was cut at the value of 4.

4.5.2 Correlation between model descriptors and toxicity

As afore-mentioned, the proposed *k*NN model comprised eight molecular descriptors, whose interpretation was provided in the scientific publication in Appendix I. In order to analyse the relationships between molecular descriptors and toxicity, a principal component analysis (PCA) was carried out. The score and loading plots of PC1 and PC2 for the training set are reported in Figure 4.5. Molecules were coloured based on the toxicity values, the toxicity increases from black to white. The loading plot (Figure 4.5b) shows that the two descriptors accounting for the polar surface area (*TPSA(Tot)* and *SAacc*) are correlated. It has been mentioned in the scientific publication in Appendix I that there is indeed a partial overlap in the information they provide, but also that *TPSA(Tot)* is hypothesised to be involved mostly for molecules containing phosphorous and sulphur, while *SAacc* accounting specifically for the formation of hydrogen-bonds. It can also be noticed that descriptors encoding information about heteroatoms (*nN*, *C-040*, *H-050*, *TPSA(Tot)* and *SAacc*) are not (or partially) correlated with descriptors that instead do not include this type of information (*RDCHI* and partly *MLOGP*). A general trend in increasing the toxicity moving downwards along the second principal component is highlighted (coefficient of correlation between PC2 and toxicity equal to -0.59). The descriptors that have larger loadings on the second component and seem therefore useful for the interpretation of this trend are *MLOGP*, *RDCHI* and *GATS1p*.
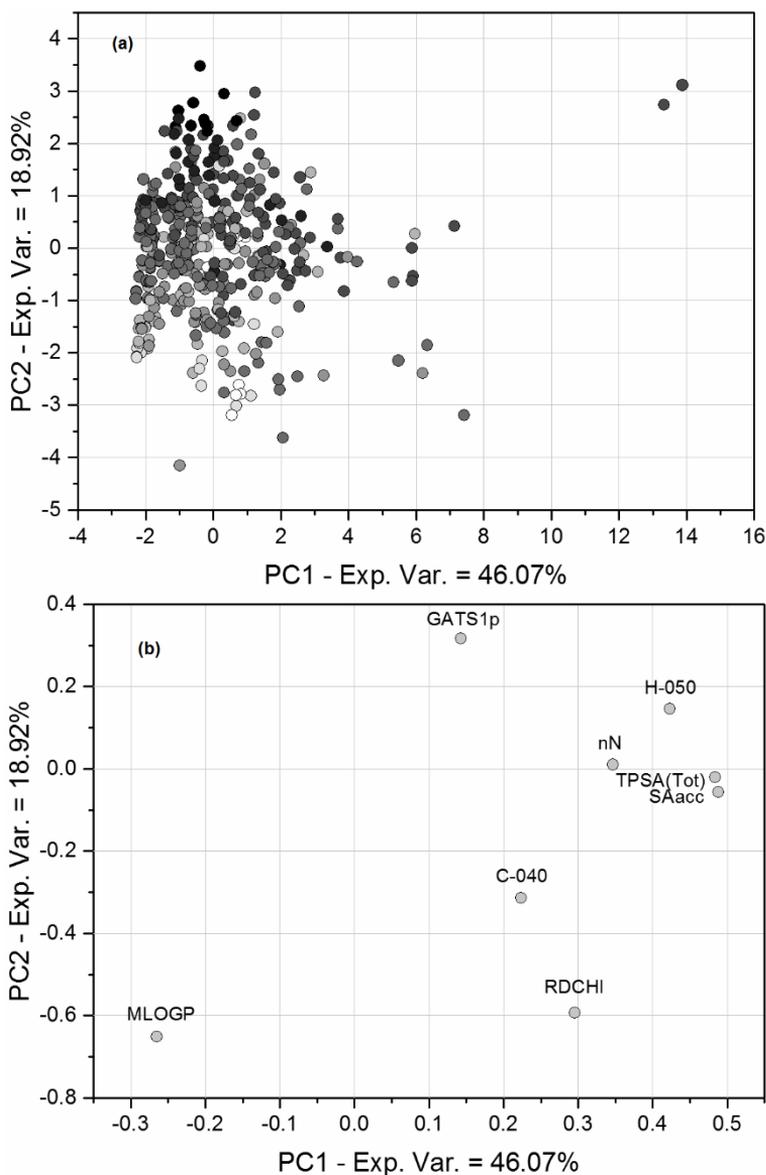
Figure 4.5. Score (a) and loading (b) plot of the training set. In the score plot, compounds are coloured based on the toxicity, which increases from black to white.

In order to confirm the latter consideration, scatterplots of pairs of model descriptors were made, where molecules were coloured based on their toxicity. The pairs of descriptors that showed the highest correlations with respect to toxicity are shown in Figure 4.6. The scatterplots essentially confirm the considerations drawn from the PCA, i.e. that the descriptors with larger correlations with the toxicity are *MLOGP*, *RDCHI* and *GATS1p*.

4.5 Discussion of the kNN model
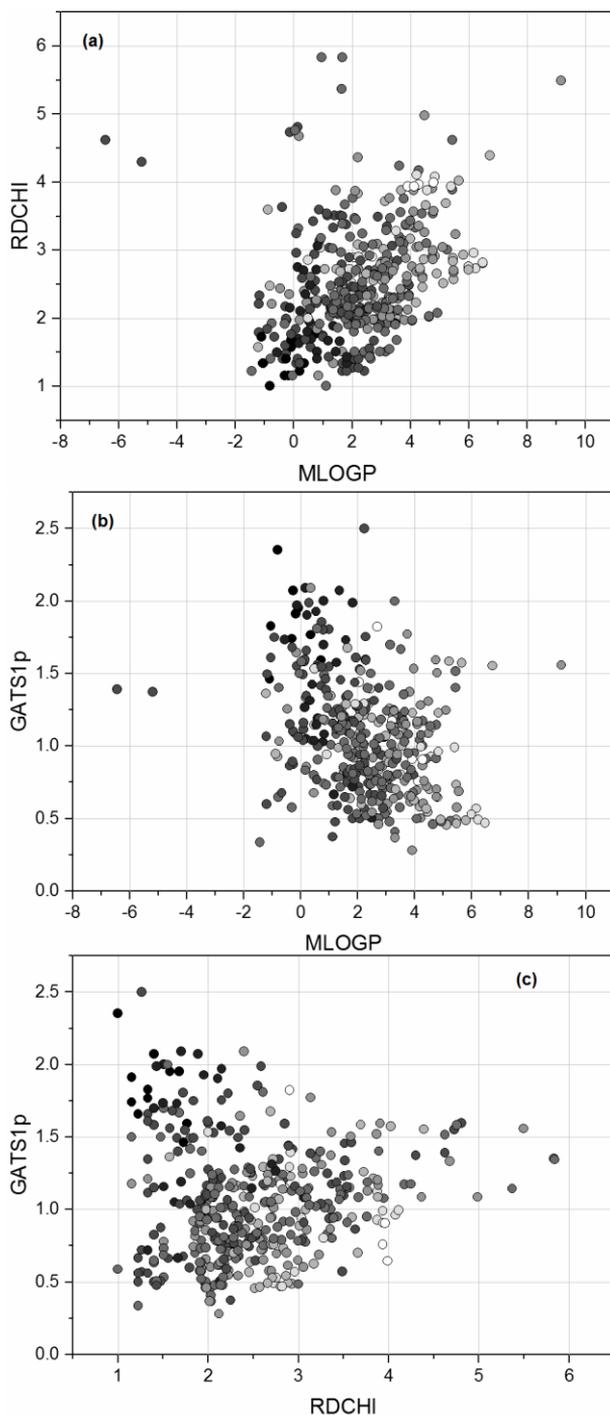
Figure 4.6. Scatterplots of the three descriptors most correlated with toxicity: MLOGP versus RDCHI (a); MLOGP versus GATS1p (b); RDCHI versus GATS1p (c). Molecules are coloured on the basis of the toxicity, which increases from black to white.

A further analysis of the relationships between descriptors and toxicity considered compounds in the first and last quartiles of toxicity (218 molecules). For each of the eight descriptors, histograms of the molecules in the first and last quartiles of toxicity were generated separately. The distribution of molecules confirmed that for *MLOGP, RDCHI* and *GATS1p* there is a trend between toxicity and descriptor values. In particular, high toxicity mainly corresponds to molecules with high *MLOGP* and *RDCHI*, as shown in Figure 4.7; the majority of molecules with *GATS1p* higher than 1.5 are instead associated to low toxicity.

In the scientific publication in Appendix I, it was hypothesised that the noticed relationship between polarisability, encoded by *GATS1p*, and toxicity could be explained in the light of the hard and soft acids and bases (HSAB) and the frontier molecular orbital (FMO) theories considering that soft species could react to form covalent bonds. Faucon et al., (2001) and Moosus and Maran, (2011) drawn similar considerations. In their cases, the molecular descriptors related to the softness were the hardness and the energy of the HOMO.

Regarding the interpretation of the roles of *MLOGP* and *H-050*, these two descriptors may help to distinguish between chemicals that cause narcosis I and narcosis II syndromes. This hypothesis was based on the evidence that type II narcosis was related also to the presence of hydrogen bond donor groups, whereas type I narcosis derived from hydrophobic interactions [Veith and Broderius, (1990)]. The contribution of the hydrogen-bonding group to toxicity was observed to decrease for high LogP.

In the last years, some parameters were used to predict absorption, among which hydrogen-bonding ability and polar surface area [Ertl et al., (2000)]. For instance, the polar surface area was found to be correlated with passive transport through membranes. Based on these considerations, it can be hypothesised that *TPSA(Tot)*, *H-050* and *SAacc*, accounting for the polar surface area and hydrogen-boding features, could be related to the ability of chemicals to cross membranes.
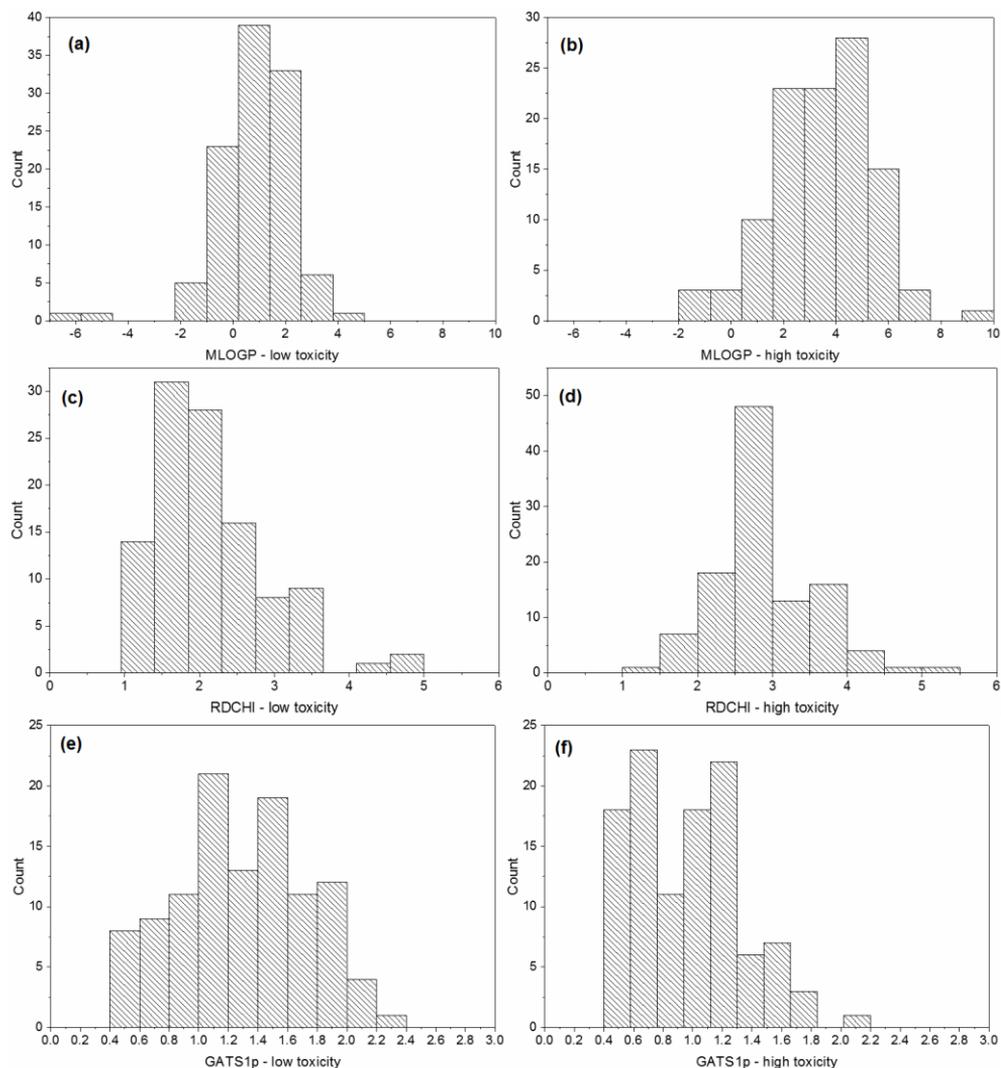
Figure 4.7. Histograms of individual descriptors for the first and last quartiles of toxicity. (a): MLOGP – first quartile; (b) MLOGP – last quartile; (c) RDCHI – first quartile; (d) RDCHI – last quartile; (e) GATS1p – first quartile; (f) GATS1p – last quartile.

4.6 Additional external validation and extension of the *k*NN model

Following the calibration of the MICHEM model, new data were gathered from the dataset of the model developed by Khüne et al., (2013) implemented in the ChemProp software [*ChemProp*, (2013)] and data provided by the QSAR group at the Technical University of Denmark [Niemelä et al., (2010)]. These data were

processed as described in paragraph 2.3.2 and three validation subsets were defined, namely:

- 'External to MICHEM': 1009 molecules from DTU and ChemProp not present in the MICHEM dataset.
- 'External to ChemProp': 228 molecules from MICHEM and DTU not present in the ChemProp dataset.
- 'External to both MICHEM and ChemProp': 128 molecules from DTU absent in both ChemProp and MICHEM datasets.

The validation subsets were used to validate the MICHEM model and compare its performance with that of the ChemProp model (Table 4.4). Moreover, the new data were used to recalibrate the previous MICHEM model and develop novel models as detailed in the scientific publication in Appendix II. A summary of the results is given in this paragraph.

The comparison of MICHEM and ChemProp models showed that they had comparable performance on the 'External to both MICHEM and ChemProp' subset. Slightly better statistics were provided by the MICHEM model on the 'External to MICHEM' subset compared to those provided by the ChemProp model on the 'External to ChemProp' subset. The lower statistics of the ChemProp model seemed due to few large errors, rather than to several medium-large errors. The MICHEM model had lower statistics on the 'External to MICHEM' validation data compared to the initial validation on the test set of 110 molecules. However, the performance was considered still satisfactory, especially considering the size of the training and validation sets (436 and 1009 molecules, respectively). The unbalance in the number of molecules between training and validation sets was reflected by the large percentage of molecules out of AD (51%). This was likely a consequence of the fact that the MICHEM model is based on a local approach and, therefore, the density of training compounds played a crucial role.

Table 4.4. Results of the external validation of MICHEM and ChemProp models.

Validation subset	No. mol.[a]	ChemProp			MICHEM		
		% out AD	Q^2_{ext}	RMSE	% out AD	Q^2_{ext}	RMSE
External to both MICHEM and ChemProp	128	47	0.60	1.073	45	0.56	1.100
External to MICHEM	1009				51	0.66	0.967
External to ChemProp	228	54	0.56	1.134			

[a] number of compounds.

4.6 Additional external validation and extension of the kNN model

Since the MICHEM model was considered to provide satisfactory results, the new data were used to define extended training (1331 molecules) and test sets (224 molecules). In doing this, the separation between training and test set chemicals of the original MICHEM dataset (546 compounds) was retained, therefore the extended test set comprised molecules never used for descriptor selection nor model calibration. The extended training set was used to recalibrate the MICHEM model by re-optimising the number of nearest neighbours and threshold on the average Mahalanobis distance. The resulting model will be referred to as the extended MICHEM model. The extended MICHEM model had slightly lower performance in internal validation but the predictivity on the extended test set was improved and the percentage of molecules out of AD was significantly reduced (Table 4.5). This was probably a consequence of the higher density of training compounds.

The extended training set was used also to calibrate a new model where binary fingerprints were used in place of the eight descriptors of the MICHEM model. Extended connectivity fingerprints and path fingerprints introduced in paragraph 3.1.3 were used. For both types of fingerprints the properties used to discriminate between fragments were (a) atom type; (b) aromaticity; (c) attached hydrogen; (d) connectivity; (e) charge; (f) bond order; and (g) rings. The maximum *radius* used for extended connectivity fingerprints was equal to two, whereas the maximum path length used for path fingerprints was set to six. In both cases, binary vectors constituted by 1024 bits were generated. The models calibrated on the binary fingerprints were based on the same approach as the MICHEM model (*k*NN with threshold on the average distance) but the Jaccard-Tanimoto similarity coefficient was used in order to find the nearest neighbours. The fingerprints models gave poorer results in terms of coefficients of determination in fitting, internal and external validation, but the percentage of molecules out of AD was also smaller. Better results were obtained with the extended connectivity fingerprints with respect to path fingerprints. The statistics of the model based on extended connectivity fingerprints are collated in Table 4.5 under the name 'Fingerprints'.

Eventually, the predictions provided by the extended MICHEM and fingerprints models were combined together in a *consensus* approach. Two *consensus* models were developed, 'Strict' and 'Loose', according to the strategies described in paragraph 3.4. The 'Loose' model allowed to broaden the AD by considering a lower percentage of molecules out of AD, whereas the 'Strict' model achieved more accurate predictions (Table 4.5).

Table 4.5. Summary of the statistics of the extended MICHEM, fingerprints and consensus models.

Model	k^a	Distance threshold	Fitting		Cross-validation[b]		External validation	
			R^2	% out AD	Q^2_{cv}	% out AD	Q^2_{ext}	% out AD
Extended MICHEM	5	1.136	0.71	36%	0.71	40%	0.69	31%
Fingerprints	6	0.664[c]	0.67	29%	0.67	33%	0.59	24%
Consensus 'Loose'	-	-	0.70	18%	0.70	20%	0.67	13%
Consensus 'Strict'	-	-	0.78	47%	0.78	52%	0.73	42%

[a] number of nearest neighbours; [b] five-fold cross-validation; [c] complement of Jaccard-Tanimoto similarity coefficient $(1-S_{jt})$.

Further details about the results of the validation of the MICHEM model and the comparison with the ChemProp model, the extended MICHEM and fingerprints models and the two *consensus* models are described in the scientific publication in Appendix II. The following paragraphs report additional analyses not included in the scientific article.

4.6.1 Analysis of neighbourhood behaviour

It was mentioned earlier that the Patterson plot and the Patterson ratio could be used to determine whether the data under analysis show neighbourhood behaviour and quantify its strength. The MICHEM dataset based on the eight molecular descriptors selected by means of GAs showed a relatively good neighbourhood behaviour (paragraph 4.5.1). The strength of the neighbourhood behaviour was determined also for the extended training set (1331 molecules) using both the eight molecular descriptors of the extended MICHEM model and the extended connectivity binary fingerprints. The Patterson plots of the training set for the extended MICHEM and fingerprints models are presented in Figure 4.8. In both cases, the lower right-hand side triangle is occupied, indicating a certain degree of neighbourhood behaviour. The Patterson plot on the binary fingerprints (Figure 4.8b) clearly shows the presence of pairs of molecules with high similarity (left-hand side of the plot), whose absolute differences of toxicity are quite large. As it was noted in the scientific publication in Appendix II, this type of binary fingerprints does not properly encode the information regarding the presence of long carbon chains, which is related to lipophilicity. Since lipophilicity plays a crucial role in aquatic toxicity, molecules regarded as similar by this type of fingerprints can have different lipophilicity and,

consequently, different toxicity: the absolute difference of toxicity for these pairs is therefore large. The weaker neighbourhood behaviour of the binary fingerprints is reflected in the Patterson ratio, which is equal to 2.99 and 1.76 for the extended MICHEM and fingerprints models, respectively. The Patterson ratio for the extended MICHEM model (2.99) is higher than the value on the original MICHEM model (2.59).
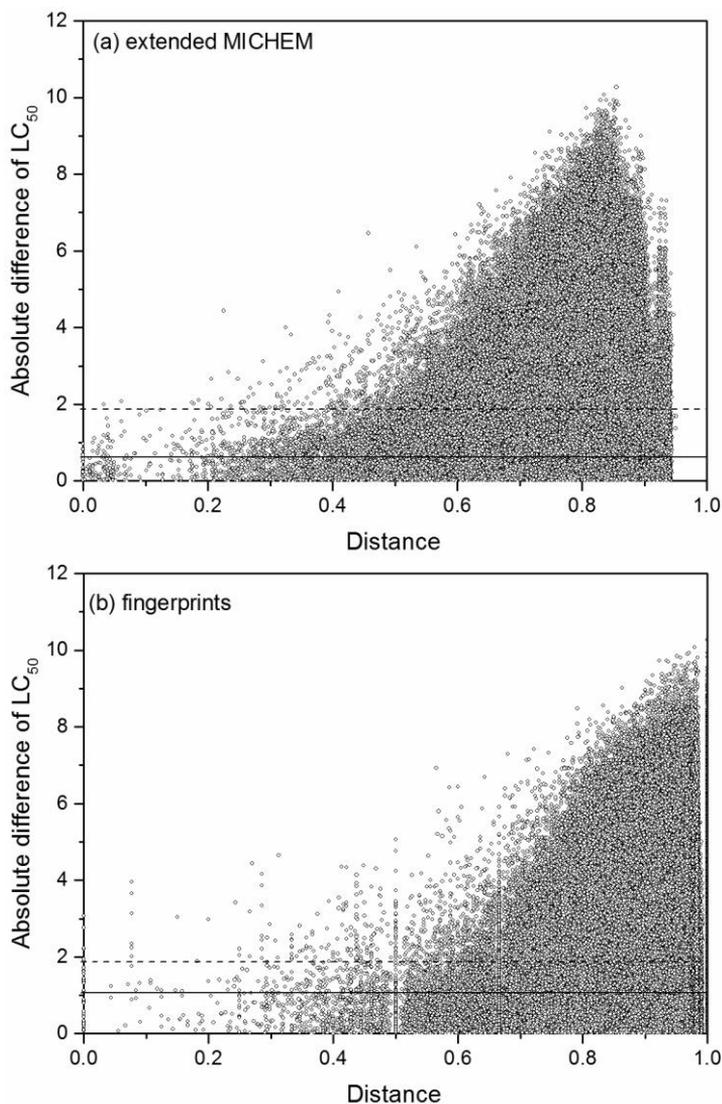
*Figure 4.8. Patterson plots of the training set for the extended MICHEM model (a) and the fingerprints model (b). X-axis: pairwise distance (1-similarity); y-axis: absolute difference between toxicity values. Dashed line: mean absolute difference of the toxicity on all pairs in the training set; solid line: mean absolute difference of the toxicity on pairs of molecules with distance smaller than or equal to 0.3.*

The strength of the neighbourhood behaviour was also checked on the extended test set (224 molecules). To this end, for each compound in the extended test set, the average absolute difference of its experimental toxicity with respect to the nearest neighbours and the average distance from the nearest neighbours were evaluated. Figure 4.9 reports the resulting scatterplots for the extended MICHEM (Figure 4.9a) and fingerprints models (Figure 4.9b). On the extended MICHEM model there is a trend showing that similarity in the descriptors space (x-axis) is related to the similarity in the toxicity space (y-axis). The situation on the fingerprints model shows a worse relationship. In particular, there is a relationship between chemical (x-axis) and toxicological (y-axis) similarities on the right-hand side of the plot (medium-low similarity in the descriptors space). This relationship is not valid in the left-hand side of the plot (high similarity in the descriptors space). This is probably still a consequence of the fact that this type of fingerprints does not properly account for the molecular lipophilicity.

4.7 Comparison with literature models

In paragraph 1.4.1, a number of published QSAR models for the prediction of acute toxicity towards *Daphnia magna* were reported. For the comparison described in this paragraph, when several models had been presented in a scientific publication, only the one suggested by the authors was used; in case no indication of this type had been given by the authors, the model associated with the highest and most balanced statistics for both internal and external validation was chosen. Table 4.6 collates information regarding the dataset and the statistics of models published in the scientific literature as well as QSAR models developed during this project.

The models developed in this study on the extended dataset and the model by Kühne et al., (2013) were derived from considerably larger datasets compared to the other literature models. The model by Kühne et al., (2013) achieved high accuracy in fitting and leave-one-out cross-validation (R^2 and Q^2_{cv} equal to 0.85 and 0.84, respectively). The external validation carried out in this study on the 'External to ChemProp' subset indicated a lower predictive power (Q^2_{ext} equal to 0.56) due mainly to few large errors rather than average modest accuracy. The percentage of molecules outside the AD was 45% in the training set and 54% on the test set. The extended MICHEM, fingerprints and *consensus* models developed in this study were characterised by more balanced statistics in fitting, internal and external validation.

4.7 Comparison with literature models

The fingerprints model had comparable performance on the test set (Q^2_{ext} equal to 0.59) with the model of Kühne et al., (2013) but the percentage of molecules outside the AD was smaller (24%). The extended MICHEM and *consensus* models achieved considerably higher statistics, the highest accuracy being provided by the *consensus* 'Strict' model (Q^2_{ext} equal to 0.73). The drawback associated to this model is the narrow applicability domain (42% of test molecules lie outside the AD).
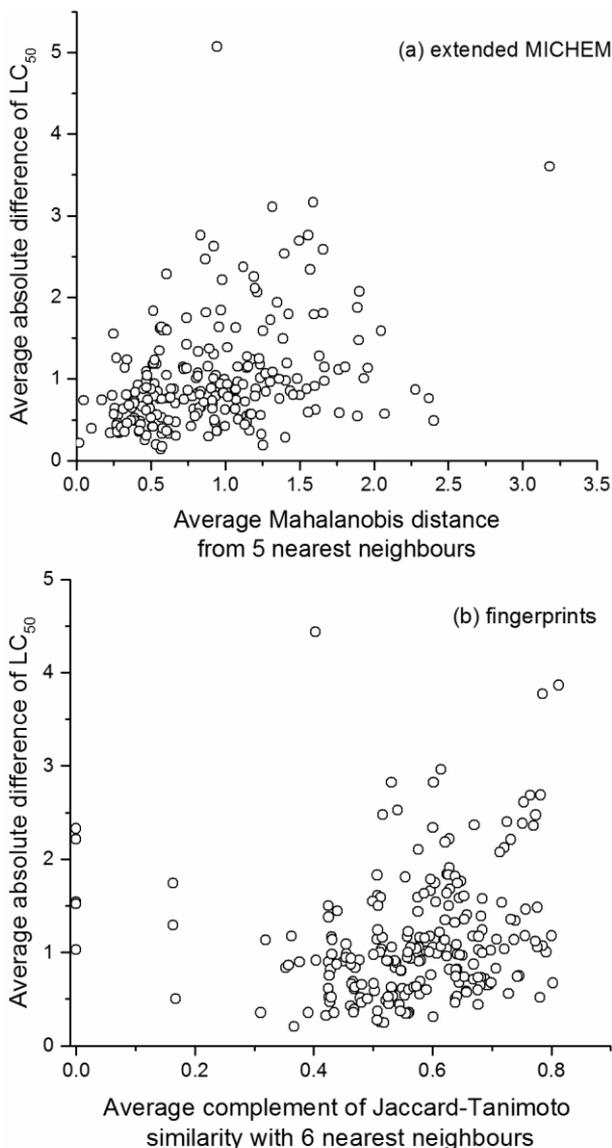

Figure 4.9. Average distance from the nearest neighbours versus average absolute difference of toxicity from the nearest neighbours for the test set. (a) extended MICHEM model; (b) fingerprints model.

The model of Kaiser and Niculescu, (2001) showed very good fitting ability (R^2 equal to 0.88) and good predictivity (Q^2_{ext} equal to 0.76) on a dataset whose size is roughly half of the ones modelled by Kühne et al., (2013) and in this study. On the other side, this model could provide a prediction for all the molecules in the dataset, because no evaluation of the AD was implemented. The fact that the predictions for all the compounds were considered, on one hand, underlines the high accuracy, but, on the other hand, does not allow to evaluate the reliability of new predictions because the space where such high accuracy is valid is not defined. Despite the good performance, this model might encounter limitations to its application for regulatory purposes also due to the complex modelling algorithm (probabilistic neural network).

Table 4.6 clearly shows that satisfactory statistics on the largest heterogeneous datasets were obtained by means of non-linear and local methods (kNN, PNN and a combination of MLR and kNN). On the contrary, modelling of smaller (yet heterogeneous) datasets could satisfactorily be carried out by means of linear methods (MLR and PLS). The hypothesis to explain this phenomenon could be that smaller datasets include less heterogeneity in the structures, and consequently in the mechanisms of toxicity. Therefore, linear and simple approaches can be suitable to model the underlying phenomena and obtain accurate predictions.

Recently, Golbamaki et al., (2014) carried out a validation of eight *in silico* software packages by means of a dataset comprising 480 compounds. The results showed that the accuracy on compounds not present in the training set was in general not very high (maximum Q^2_{ext} equal to 0.54). The MICHEM model was validated on an external test set of 1009 molecules (larger than the training set) and gave considerably better results (Q^2_{ext} equal to 0.66) on the compounds inside the AD (49%). Based on these considerations, it can be said that the models developed in this study showed satisfactory robustness and predictivity, especially considering the challenges presented by the size of the dataset (training and test sets comprising 1331 and 224 compounds, respectively) and the simplicity of the algorithm. Additional advantages are the use of 2D molecular descriptors that did not require geometry optimisation and quantum-chemical calculations and the systematic AD assessment procedure, which is lacking in several literature models.

Table 4.6. Details of QSAR models for the prediction of acute toxicity towards Daphnia magna published in the literature and developed during this project. Hyphens are used for lacking information.

Reference	Method[a]	n train[b]	n test[c]	R^2	Q^2_{cv}	Q^2_{ext}
MICHEM	kNN	436	110	0.78	0.78[d]	0.72
			1009			0.66
Extended MICHEM	kNN	1331	224	0.71	0.71[d]	0.69
Fingerprints	kNN	1331	224	0.67	0.67[d]	0.59
Consensus 'Loose'	*Consensus*	1331	224	0.70	0.70[d]	0.67
Consensus 'Strict'	*Consensus*	1331	224	0.78	0.78[d]	0.73
Kühne et al., (2013)	LR+kNN (tree)	1365	228	0.84	0.85[e]	0.56[f]
Kaiser and Niculescu, (2001)	PNN	700	76	0.88	-	0.76
Devillers et al., (1987)	MLR	57	44	0.89	-	0.70
Todeschini et al., (1996)	MLR	49	0	0.82	0.71[g]	-
Faucon et al., (2001)	MLR	61	35	0.54	0.49	0.57
Tao et al., (2002)	MLR	217	-	0.97	0.97[h]	-
Toropov and Benfenati, (2006)	MLR	220	42	0.78	-	0.74
Moosus and Maran, (2011)	MLR	118	117	0.74	0.74[i]	0.56
Toropova et al., (2012a)	MLR	114+108[j]	75	0.71	0.72[k]	0.78
Toropova et al., (2012b)	MLR	107+115[j]	75	0.73	0.80[k]	0.75
Katritzky et al., (2009)	MLR	86	44	0.70	0.64	0.74
Kar and Roy, (2010)	PLS	222	75	0.70	0.68[e]	0.74
Martin et al., (2012)	*Consensus*	283	70	-	-	0.74
			333[l]			0.46[l]
VEGA, EPA	MLR	269	68	0.71	-	0.75
			374[l]			0.54[l]
VEGA, DEMETRA	neuro-fuzzy	220	43	0.93	-	0.97
			220[l]			0.14[l]
			135[m]			0.63[m]

[a] kNN: k-Nearest Neighbours; LR+kNN(tree): Linear regression coupled with k-Nearest Neighbours with similarity thresholds; PNN: Probabilistic Neural Network; MLR: Multiple Linear Regression; PLS: Partial Least Squares regression; neuro-fuzzy: hybrid system that combines one PLS and two NN models; [b] number of compounds in the training set; [c] number of compounds in the test set; [d] 5-fold cv; [e] leave-one-out; [f] performed in this study; [g] leave-20%-out; [h] leave-20 molecules-out repeated 30 times; [i] iterated leave-30%-out; [j] training set + validation set; [k] Q^2 on the validation set; [l] additional validation reported by Golbamaki et al., (2014); [m] additional validation reported by Porcelli et al., (2008).

4.8 Compliance with the OECD principles

The applicability of QSAR models in the framework of REACH regulation depends on the fulfilment of the five OECD principles for the validation outlined in paragraph 1.3. The endpoint addressed by the models developed in this study for the prediction of acute toxicity towards *Daphnia magna* was defined (LC_{50} over a test duration of 48 hours). However, it should be recalled that data obtained under different experimental conditions and employing different populations of daphnids were merged together.

The models were based on the *k*-nearest neighbours method (*k*NN) where the prediction was computed as similarity-weighted average of the toxicities of the *k* nearest neighbours in the training set. The distance (or similarity) measure was specified: Mahalanobis distance for the MICHEM and extended MICHEM models, and the Jaccard-Tanimoto similarity coefficient for the fingerprints model. Additional characteristics that should facilitate the acceptability of the models for regulatory application are:

- The simplicity of the algorithm.
- The fact that a local method should, in principle, be able to deal with the different underlying mechanisms of toxicity because only a local region of the space is considered to make a prediction (presumably molecules acting via the same mechanism).

The assessment of the applicability domain was carried out for each prediction by the analysis of the average distance from the *k* nearest neighbours, which was compared to a fixed distance threshold. The distance threshold was selected to optimise the performance but it can be changed to tune the strictness of the AD criterion. The average distance itself could be used as a quantitative estimate of the reliability of the prediction, together with information regarding the structure and experimental toxicities of the nearest neighbours and the performance of the model in the chemical region of interest. All this information can help the evaluation of the uncertainty associated with each prediction.

The predictivity of the models was evaluated by means of test sets not used to calibrate nor optimise the models. The original MICHEM model was also subject to a thorough external validation carried out on a large set of 1009 molecules. The robustness was evaluated by means of five-fold cross-validation.

4.8 Compliance with the OECD principles

      The eight molecular descriptors of the MICHEM and extended MICHEM models showed evident trends with toxicity. Hypotheses regarding the interpretation of these descriptors in the light of current knowledge on aquatic toxicity were made. No interpretation was given for the binary fingerprints because this description of the molecular structure compresses the information regarding the presence of fragments in a vector of predefined length where there is no univocal association bit-fragment. However, flaws of the type of binary fingerprints used were highlighted.

      To conclude, the models developed for the prediction of acute toxicity towards *Daphnia magna* comply with the OECD principles and their applicability in the framework of REACH should be possible.

Results on Pimephales promelas

'Science may be described as the art of systematic over-simplification.'

Karl Popper, The Open Universe, 1982

5.1 Explorative analysis

The final dataset for acute toxicity towards *Pimephales promelas* comprised 908 organic chemicals (paragraph 2.4). The relationship between the *n*-octanol-water partition coefficient (LogP), which was shown to be correlated with narcosis, and toxicity was checked similarly to the case of *Daphnia magna*. The plot of the calculated LogP by means of KOWWIN software [*KOWWIN*, (2010)] *versus* the experimental toxicity is depicted in Figure 5.1. The solid line corresponds to the baseline toxicity model used in [Schüürmann et al., (2011)]. Like the case of *Daphnia magna*, many compounds lie along the baseline, while eleven molecules are placed far below (more than two logarithmic units), thus indicating that their toxicity is lower than what predicted based on the sole narcosis model (black circles in Figure 5.1). These large deviations from the baseline seem to occur mainly at medium-high LogP values. The same three hypotheses drawn for the case of *Daphnia magna* (paragraph 4.1) can apply here to explain this phenomenon, i.e. (a) variability in the toxicity values (one of the 11 chemicals, dehydroabietic acid, has one of the largest experimental variability (paragraph 2.2)); (b) accuracy of LogP estimates; and (c) accuracy of measurements or LC_{50} determinations. Therefore, experimental LogP values were searched for in the KOWWIN dataset, the OECD QSAR Toolbox [*The OECD QSAR Toolbox*, (2013)] and the ECHA registration database [*ECHA*]. Since experimental data were not found for all chemicals, predictions were also taken by the *MLOGP* model [Moriguchi et al., (1992); Moriguchi et al., (1994)] implemented in DRAGON software (Table 5.1). The comparison of LogP for the six chemicals

with available experimental values showed again that KOWWIN tends to overestimate and *MLOGP* to underestimate the LogP. The third hypothesis concerns the likelihood of inaccurate measurements, for instance if nominal concentrations were used. The analysis of water solubility in Table 5.1 (experimental and predicted with WATERNT [*WATERNT*, (2010)] and T.E.S.T. [Martin et al., (2012)]) highlighted that almost all LC_{50} values are lower (in negative logarithmic scale) than the water solubility, thus corroborating the hypothesis of inaccurate measurements. The experimental water solubility of dehydroabietic acid is, instead, slightly lower than the LC_{50} (-Log(mol/L)); the same applies to the predicted solubility of monochloro- and dichlorodehydroabietic acids by T.E.S.T.. Based on these considerations, the use of experimental LogP and *MLOGP* in place of the estimates of KOWWIN, and of water solubility in place of LC_{50} (when water solubility is higher in negative logarithm of molarity) would shift the eleven chemicals towards the baseline.
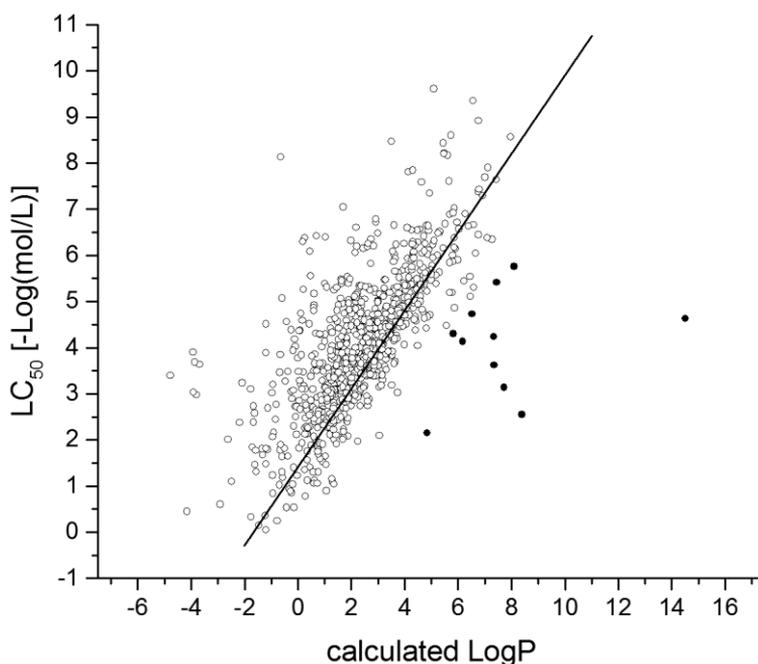
*Figure 5.1. Calculated LogP versus LC_{50} 96 hours (-Log(mol/L)) towards Pimephales promelas. Solid line: baseline toxicity used in Schüürmann et al., (2011). Black circles: eleven molecules whose LC_{50} are lower than 2 Log units from the baseline.*

Table 5.1. LogP, water solubility and LC₅₀ for the eleven molecules with toxicity lower than 2 logarithmic units from the baseline in Figure 5.1. Water solubility and LC₅₀ are reported as –Log(mol/L). Hyphens are used for lacking information.

CAS-RN	Name	LogP KOWWIN	*MLOGP*	Exp. LogP[a]	Water sol.	LC₅₀
107-64-2	Dimethyldioctadecyl ammonium chloride	14.52	6.20	-	12.00[b];6.03[c]	4.63
112-80-1	Oleic acid	7.73	5.58	7.64	7.14[b];4.96[c]	3.14
117-81-7	Bis(2-ethylhexyl) phthalate	8.39	5.43	[7.14, 7.94]	6.16[d]	2.55
138-86-3	Dipentene	4.83	3.27	4.57	3.99[d]	2.15
1740-19-8	Dehydroabietic acid	6.52	4.55	4.80	4.66[d]	4.73
2385-85-5	Mirex	7.35	6.52	6.89	6.81[d]	3.62
32534-95-5	4-Methylheptyl 2-(2,4,5-trichlorophenoxy)propanoate	7.33	5.18	-	7.98[b];6.76[c]	4.24
3383-96-8	Temephos	6.17	3.17	5.96	6.24[d]	4.14
475-20-7	Longifolene	5.82	4.74	-	6.81[b];5.21[c]	4.30
57055-38-6	Monochlorodehydroabietic acid	7.45	5.26	-	7.97[b];5.13[c]	5.41
57055-39-7	Dichlorodehydroabietic acid	8.10	5.74	-	8.66[b];5.43[c]	5.76

[a] experimental LogP; [b] predicted by WATERNT; [c] predicted by T.E.S.T.; [d] experimental.

From Figure 5.1 it is evident that also for the analysis of the toxicity towards the *Pimephales promelas* a linear model based only on the *n*-octanol-water partition coefficient would not give accurate predictions for all chemicals. Regression analysis was carried out by including additional molecular descriptors in a model calibrated on the whole training set (paragraph 5.3) in order to avoid additional steps (e.g. initial clusterisation in homogeneous groups of compounds) that would render the modelling strategy more complex.

5.2 Modelling with Leadscope Enterprise™

Leadscope Enterprise™ software was used to derive QSAR models on the *Pimephales promelas* dataset. The file containing the SMILES strings of the molecules was converted to structure-data file (sdf) format for import into Leadscope Enterprise™. The dataset was screened against a library of 27,000 fragments and only fragments possessed by at least one compound were retained. A number of scaffolds (704) were generated starting from the structures included in the dataset. The nine descriptors of molecular properties implemented in Leadscope were also calculated. Since the final model was calibrated on all the molecules, no fixed external test set was extracted. The external validation was carried out by means of the 5 x 2 x 50% off strategy previously described (paragraph 3.9.1). Fragments, scaffolds and descriptors of properties were used both separately and together to calibrate PLS

regression models. The selection of relevant molecular descriptors was carried out by means of the implemented algorithm, which is based on a) the analysis of the frequency (unsupervised), and b) the analysis of the residuals in cross-validation and t^2 test (supervised). The evaluation of the applicability domain (AD) was carried out according to the two criteria implemented in the software, i.e. presence of fragments included in the model and similarity with the training molecules (paragraph 3.9.1). A summary of the model calibration strategies is given in Table 5.2.

5.3 Modelling with DRAGON descriptors

DRAGON descriptors were also used to derive QSAR models from the *Pimephales promelas* dataset. The SMILES strings of the compounds were imported in DRAGON in order to calculate 0D, 1D and 2D molecular descriptors. Some molecular descriptors were neglected, just like for the analysis of the toxicity towards *Daphnia magna*:

- Descriptors from the drug-like block because they were supposed not to be relevant for modelling acute aquatic toxicity.
- Intrinsic state and electrotopological state pseudo-connectivity indices [Pogliani, (2000); Pogliani, (2004)], Estrada-like indices (*EE_M(w)*), coefficient sum, average coefficient and logarithmic coefficient sum of the last eigenvector (*VE1_M(w)*, *VE2_M(w)*, *VE3_M(w)*), Randic-like, normalised Randic-like and logarithmic Randic-like eigenvector based-index (*VR1_M(w)*, *VR2_M(w)*, *VR3_M(w)*) [Balaban et al., (1991)], Ghose-Crippen-Viswanadhan octanol-water partition coefficient (*ALOGP*) and its squared value (*ALOGP2*) [Ghose and Crippen, (1986); Viswanadhan et al., (1993); Ghose et al., (1998)] and molar refractivity (*AMR*) [Ghose and Crippen, (1987)] because they may not be computable on some structures.
- Baseline toxicity to daphnia, fish and algae because they are trivially defined from the *MLOGP* ($\rho = -1.00$).

Constant, near-constant and descriptors with at least one missing value were removed in DRAGON leading to a set of 2221 molecular descriptors. Unsupervised variable reduction was carried out on the basis of the pairwise correlation providing a final set of 1218 molecular descriptors used for the subsequent modelling analysis. The dataset was then randomly split into training and test sets comprising 726 and 182 chemicals, respectively. The training set was used for the selection of relevant

molecular descriptors and the calibration of QSAR models, whereas the test set was only submitted to the final models to assess their predictivity. The selection of relevant DRAGON molecular descriptors was carried out by means of genetic algorithms (GAs) and reshaped sequential replacement (RSR), as described below.

- **GAs:** GAs were run separately on each block of molecular descriptors from DRAGON (18 blocks in total). The most frequently selected descriptors were gathered from the 18 blocks and merged together to form a pool of approximately 200 descriptors, which was input to GAs. Afterwards, a reduced set of, approximately, the most selected 100 descriptors was generated and GAs were run again on this set. In the end, all the possible combinations of a small set of descriptors with the largest selection frequencies were generated and evaluated by means of an all subset models strategy (ASM).

- **RSR**: one single run was carried out on the whole set of 1218 molecular descriptors. The tabu list and the roulette wheel were activated with the aim to exclude descriptors uncorrelated to the response and generate an initial population comprising the most promising descriptors.

GAs were used to derive linear (OLS and PLS), nonlinear (SVR and GPR) and local (kNN) models, whereas RSR was only used to derive linear models by means of OLS regression. Five distance measures were used to calibrate kNN models, namely Euclidean, Mahalanobis, Soergel, Lance-Williams and Jaccard-Tanimoto.

The leverage approach was used as detailed in paragraph 3.5 to assess the applicability domain of models based on OLS, PLS, SVM and GPR. For models based on kNN, the AD was assessed by comparing the average distance from the k nearest neighbours with a fixed threshold selected in cross-validation to optimise the performance. A constraint on the maximum percentage of molecules regarded out of AD (40%) was introduced in order to avoid the selection of extreme values of the threshold. Since the Jaccard-Tanimoto and Soergel distance measures (paragraph 3.3.2.1) require molecular descriptors to have positive values, range scaling between [0,1] was applied when using these metrics. In order to assure values in the range [0,1] also for test molecules, a preliminary check based on the bounding box was introduced: test molecules with descriptors values out of range of the training set were considered out of AD.

Some QSAR models were also calibrated only on the MED-Duluth fathead minnow database (566 compounds) [Russom et al., (1997)]. These models were then

externally validated on the subset of compounds not included in the MED-Duluth database (349 compounds). A summary of the model calibration strategies is given in Table 5.2.

5.4 Modelling with binary fingerprints

The *Pimephales promelas* dataset was modelled also using binary fingerprints to describe the molecular structure. Extended connectivity fingerprints and path fingerprints introduced in paragraph 3.1.3 were used. For both types of fingerprints the properties used to discriminate between fragments were (a) atom type; (b) aromaticity; (c) attached hydrogen; (d) connectivity; (e) charge; (f) bond order; and (g) rings. The maximum *radius* used for extended connectivity fingerprints was equal to two, whereas the maximum path length used for path fingerprints was set to six. In both cases, binary vectors constituted by 1024 bits were generated. Neither variable reduction, nor selection was carried out on binary fingerprints because there is no univocal correspondence descriptor-fragment and the structural information is distributed along the entire fingerprint.

Table 5.2. Strategies used to derive regression models for Pimephales promelas.

Descriptors	Variable reduction	Variable selection	Method	AD	*n* train[a]	*n* test[b]	Internal validation	External validation
Leadscope	frequency	t^2 test + residuals	PLS	fragments + distance thr on avg dist[e]	908	-	10 x 50% off[c]	5 x 2 x 50% off[d]
DRAGON	correlation	GAs	*k*NN		726	182	5-fold cv	test set
DRAGON	correlation	GAs	OLS	leverage	726	182	5-fold cv	test set
DRAGON	correlation	GAs	PLS	leverage	726	182	5-fold cv	test set
DRAGON	correlation	GAs	SVR	leverage	726	182	5-fold cv	test set
DRAGON	correlation	GAs	GPR	leverage	726	182	5-fold cv	test set
DRAGON	correlation	RSR	OLS	leverage	726	182	5-fold cv	test set
Fingerprints	-	-	*k*NN	thr on avg sim[f]	726	182	5-fold cv	test set

[a] number of compounds in training set; [b] number of compounds in test set; [c] random splitting 50%-50% repeated ten times; [d] random splitting 50%-50% repeated five times prior to variable selection; [e] threshold on the average distance from the *k* nearest neighbours; [f] threshold on the average similarity with the *k* nearest neighbours.

The same splitting training-test set defined for DRAGON descriptors was used here as well. Binary fingerprints were used to derive models by means of kNN, where the similarity between compounds was evaluated by means of the Jaccard-Tanimoto coefficient (paragraph 3.3.2.1). The applicability domain was assessed by comparing the average similarity with the k nearest neighbours in the training set with a fixed threshold. The threshold value was optimised in cross-validation, given a constraint on the maximum percentage of molecules regarded out of AD (40%). A summary of the model calibration strategies is given in Table 5.2.

5.5 Summary of results

The best results obtained with the different regression methods are collated in Table 5.3. The statistics in fitting, internal and external validation were balanced for almost all models suggesting lack of overfitting. The models for which the applicability domain (AD) was assessed by the leverage approach had quite comparable statistics, regardless of the algorithm. In fact, the values of the coefficient of determination in internal and external validation were in the range [0.60,0.68] and [0.55,0.63], respectively. The statistics were not high, but on the other side the percentages of compounds considered out of the AD were small, especially on the training set. In order to check whether a stricter AD criterion would give beneficial effects, these models were re-calibrated using three times (instead of five times) the mean leverage as a threshold to define the model AD. On average, the stricter criterion was accompanied by a decrease in the root mean square error (RMSE), which was however modest. For few models, on the contrary, the RMSE increased, thus indicating worsening of the performance.

Regarding OLS regression, RSR seemed to provide better models compared to GAs in terms of both statistics and simplicity (lower number of molecular descriptors). The first model in Table 5.3, in particular, seems a good result if compared with the results obtained with the other regression methods, which are more complex.

The PLS model calibrated with Leadscope Enterprise™ seemed to perform better than the PLS models based on DRAGON descriptors, especially considering that both internal and external validation had been carried out by means of more thorough approaches based on splits 50%-50% (paragraph 3.9.1). A remark could be made on the large number of descriptors (119). However, only eight descriptors are

molecular properties, whereas the remaining 111 account for fragments. Fragmental descriptors can be considered to contribute less to the increase of model complexity because (a) they provide a simple "local" (or "partial") information, and (b) they take on null values when the fragments are absent (no contribution to the calculated toxicity).

SVR and GPR models had comparable statistics in internal validation (Q_{cv}^2) with respect to OLS and PLS models, but showed slightly higher predictivity (Q_{ext}^2). The statistics of GPR models were more balanced than were those of SVR models. Additionally, SVR tended to identify a large number of support vectors (SVs). In fact, it had been reported that the data compression properties of SVR are negatively affected by high dimensional and noisy data [Smola and Schölkopf, (2004)]. The main concern related to GPR is the long computational time. Indeed, the calculation to run GAs using the covariance function reported in paragraph 3.3.4 lasted more than one month on an *Intel Xeon CPU E5-2620 @ 2.00 GHz* with 16 GB RAM.

*K*NN models showed considerably different behaviour, likewise the *Daphnia magna* case (paragraph 4.4). The values of the coefficient of determination in fitting, internal and external validation were usually above 0.70, at the expense of a large percentage of molecules considered out of AD. Binary fingerprints did not provide satisfactory models. Interestingly, better statistics were obtained with path fingerprints, whereas for the *Daphnia magna* case extended connectivity fingerprints resulted more suitable. A possible hypothesis to explain this phenomenon is based on the consideration that path fingerprints should better encode the presence of long aliphatic chains, which are known to affect lipophilicity. This could be interpreted as an indication that the contribution of lipophilicity to toxicity may be higher for fish than it is for daphnids. *K*NN models based on DRAGON descriptors were calibrated using five different distance measures. The Euclidean distance gave the poorest statistics, whereas the performance of the models that used the Mahalanobis, Soergel, Lance-Williams and Jaccard-Tanimoto distances were comparable.

The *n*-octanol-water partition coefficient (calculated) was selected in all models, either as *MLOGP*, *MLOGP2* (squared *MLOGP*) or *ALOGP*, thus confirming its relevance for modelling acute toxicity to fish. Interestingly, *MLOGP2* was usually present together with *MLOGP*, with the exception of the two GPR models where *MLOGP* was not present. Another frequently selected molecular descriptor was the Geary autocorrelation of lag 1 weighted by ionisation potential (*GATS1i*) [Todeschini and Consonni, (2009)].

Table 5.3. Best models obtained for LC$_{50}$ towards Pimephales promelas with different regression methods. NR: not reported.

Method[a]	p[b]	Scaling	distance / kernel[c]	AD	LVs/k/ SVs[d]	R²	Q²$_{cv}$	Q²$_{ext}$	% out AD train[e]	% out AD test[f]
RSR_OLS	11	-	-	5*$\bar{h}$[g]	-	0.69	0.68	0.60	4%	8%
RSR_OLS	10	-	-	5*$\bar{h}$[g]	-	0.67	0.67	0.56	3%	3%
GA_OLS	28	-	-	5*$\bar{h}$[g]	-	0.65	0.61	0.57	2%	3%
GA_OLS	12	-	-	5*$\bar{h}$[g]	-	0.63	0.61	0.55	2%	5%
Leadscope PLS	119	NR	-	fragm +dist[h]	5	0.67	0.64	0.58	3%	10%
GA_PLS	27	auto	-	5*$\bar{h}$[g]	3	0.63	0.61	0.58	2%	4%
GA_PLS	9	auto	-	5*$\bar{h}$[g]	3	0.61	0.60	0.52	2%	2%
GA_SVR	37	auto	rbf[i]	5*$\bar{h}$[g]	634	0.75	0.66	0.62	1%	3%
GA_SVR	8	auto	rbf[i]	5*$\bar{h}$[g]	386	0.68	0.63	0.60	2%	4%
GA_GPR	5	auto	-	5*$\bar{h}$[g]	-	0.64	0.63	0.59	2%	4%
GA_GPR	7	auto	-	5*$\bar{h}$[g]	-	0.64	0.61	0.63	1%	4%
GA_kNN	13	auto	Euclidean	1.62[j]	5	0.70	0.69	0.66	27%	26%
GA_kNN	6	range	Jaccard-Tanimoto[k]	0.152[j]	6	0.73	0.74	0.77	33%	27%
GA_kNN	14	auto	Mahalanobis	1.90[j]	3	0.74	0.74	0.73	36%	34%
GA_kNN	7	range	Lance-Williams	0.064[j]	3	0.71	0.73	0.76	30%	32%
GA_kNN	9	range	Soergel	0.132[j]	3	0.73	0.72	0.76	35%	31%
FP_kNN	1024	-	Jaccard-Tanimoto[l]	0.738[j]	9	0.59	0.58	0.61	25%	25%

[a] RSR_OLS: OLS by means of RSR; GA_OLS: OLS by means of GAs; Leadscope_PLS: PLS by means of Leadscope Enterprise™; GA_PLS: PLS by means of GAs; GA_SVR: ν-SVR by means of GAs; GA_GPR: GPR by means of GAs; GA_kNN: kNN by means of GAs; FP-kNN: kNN on binary path fingerprints; [b] number of model descriptors; [c] type of distance for kNN or type of kernel for SVR; [d] number of latent variables (LVs) for PLS, number of nearest neighbours (k) for kNN, number of support vectors (SVs) for SVR; [e] percentage of training compounds out of AD; [f] percentage of test compounds out of AD; [g] five times the mean leverage; [h] presence of fragments and distance from training molecules; [i] radial basis function; [j] threshold value on the average distance from the k nearest neighbours; [k] Jaccard-Tanimoto distance for real data; [l] complement of Jaccard-Tanimoto similarity coefficient for binary data (1-S$_{jt}$).

Considering both the statistics and the number of model descriptors, the kNN model based on the Jaccard-Tanimoto distance was chosen as the overall best result and is further analysed in the following paragraphs.

5.6 Discussion of the kNN model

The kNN model based on the Jaccard-Tanimoto distance calculated on six molecular descriptors, hereinafter referred to as the MICHEM model, was selected as the best model obtained. The scientific publication in Appendix III includes the description of (a) the procedure used to prepare the dataset; (b) the calibration of the model and optimisation of model parameters (number of nearest neighbours, k, and

5.6 Discussion of the kNN model

distance threshold); (c) the validation procedures and corresponding statistics; (d) the analysis of the residuals; (e) the interpretation of model descriptors and investigation of their correlation with toxicity; (f) comparison with literature models; and (g) an example of application. Here a summary is given. The optimisation of model parameters was carried out in five-fold cross-validation. The value of k was selected to be equal to six, whereas two distance threshold values were chosen ('Strict' = 0.152 and 'Soft' = 0.197). The two distance thresholds correspond to a different strictness of the criterion to assess the applicability domain (AD). The 'Strict value (0.152) demands a lower average distance from the six nearest neighbours in the training set (i.e. higher similarity). The 'Soft' distance threshold (0.197) corresponds to a less strict AD criterion and was chosen noticing that the coefficient of determination in internal validation (Q_{cv}^2) decreased smoothly, while the percentage of compounds out of AD dropped more sharply. A constraint on the maximum percentage of molecules considered out of AD (40%) was imposed in order to avoid the selection of extreme values. It should be recalled from the discussion of the QSAR model for *Daphnia magna* (paragraph 4.5) that the threshold can be changed to tune the strictness of the AD criterion. It is noteworthy that training molecules considered out of AD were still retained in the training set because they could be useful to predict future query compounds. The model had good performance in internal and external validation (Q_{cv}^2 = 0.67, Q_{ext}^2 = 0.73 with the 'Soft' threshold and Q_{cv}^2 = 0.74, Q_{ext}^2 = 0.77 with the 'Strict' threshold). In addition to the five-fold internal validation, the model was also tested by random splitting 80%-20% repeated 1000 times achieving good statistics (Q_{cv}^2 equal to 0.72 and 0.79 with the 'Soft' and 'Strict' criterion, respectively). The six molecular descriptors were Moriguchi octanol-water partition coefficient (*MLOGP*), complementary information content index of 0-order (*CIC0*), number of unsaturated sp^2 carbon atoms of the type =C< and =CH- (*NdssC* and *NdsCH*), Geary autocorrelation of lag 1 weighted by ionisation potential (*GATS1i*) and spectral moment of order 1 calculated from the Barysz matrix weighted by the atomic number (*SM1_Dz(Z)*). These descriptors encoded information on lipophilicity, heteroatoms and electrophilicity. The following paragraphs report additional investigations not included in the scientific article in Appendix III.

5.6.1 Analysis of the neighbourhood behaviour

The Patterson plot [Patterson et al., (1996)] was used to visually check whether the dataset based on the six molecular descriptors of the MICHEM model exhibited neighbourhood behaviour. The Jaccard-Tanimoto distance was used on the x-axis because it is already defined in the range [0,1]. The Patterson plot of the training set (Figure 5.2) seems to show a relatively good neighbourhood behaviour. Indeed, the absolute differences of toxicity increase along the x-axis from left to right. However, some pairs at null distance show differences up to two logarithmic units. Compared to the Patterson plots for the *Daphnia magna* case, the high similarity region (left-hand side) is more densely populated. The Patterson ratio was used to quantify the strength of the neighbourhood behaviour. The calculated value (1.66) indicates a moderate neighbourhood behaviour.
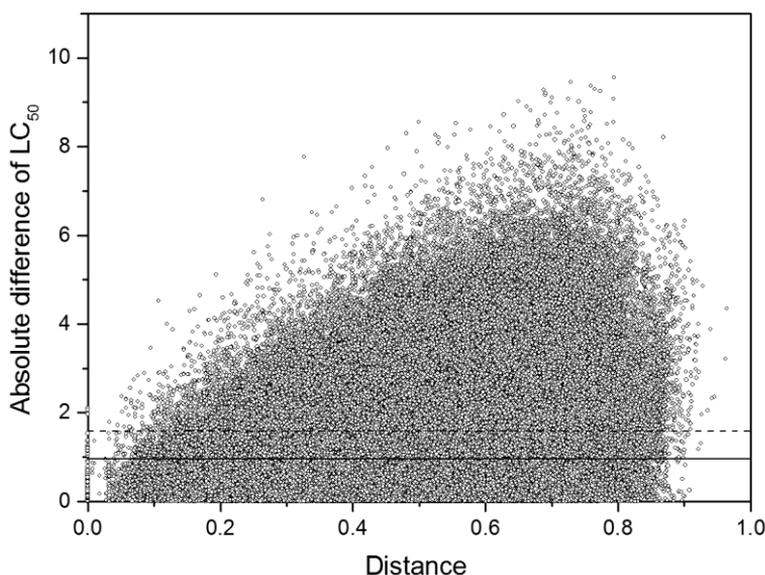

Figure 5.2. Patterson plot of the training set. X-axis: pairwise Jaccard-Tanimoto distance; y-axis: absolute difference between toxicity values. Dashed line: mean absolute difference of the toxicity on all pairs in the training set; solid line: mean absolute difference of the toxicity on pairs of compounds with distance smaller than or equal to 0.3.
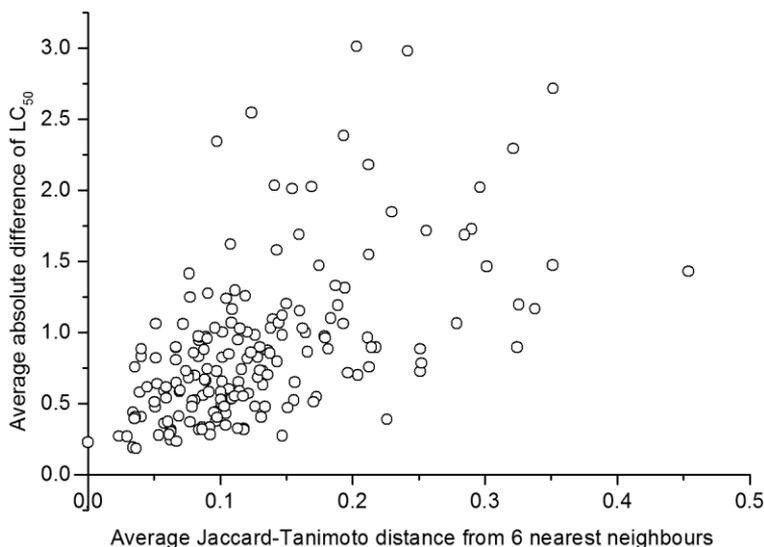
Figure 5.3. Average distance from the six nearest neighbours versus average absolute difference of toxicity from the six nearest neighbours for the test set.

An analysis of the relationship between similarity in the descriptors and in the response domains was carried out on the test set as well. For each compound of the test set, the average absolute difference of its experimental toxicity with respect to the six nearest neighbours and the average Jaccard-Tanimoto distance from the six nearest neighbours were evaluated. The resulting scatterplot (Figure 5.3) shows a pattern similar to that observed in the Patterson plot on the training set (Figure 5.2), but stronger. In fact, the mean absolute differences in the toxicity values are much smaller, which suggest higher prediction accuracy. The pattern observed in Figure 5.3 can further justify the introduction of a distance threshold on the average distance from the six nearest neighbours, so that predictions based on dissimilar molecules are hindered because likely to be inaccurate.

5.6.2 Correlation between model descriptors and toxicity

The meaning of the six molecular descriptors included in the MICHEM model is explained in the scientific publication in Appendix III, where also a principal component analysis (PCA) was carried out to investigate the relationships between model descriptors and toxicity. The PCA highlighted an evident trend between the first principal component and toxicity, which is mainly related to the effect of lipophilicity. The second principal component showed a weak correlation with toxicity, which was related to the presence of heteroatoms. Two model descriptors

(*NdssC* and *NdsCH*) had low loadings on both components and their contribution seemed limited to electrophilic compounds. However, from the score plot (Figure 5.4), where molecules are coloured based on the toxic potency, one compound of concern can be identified (highlighted by a black square). This compound is 2-Mercaptopyridine N-oxide sodium salt (CAS-RN 3811-73-2) and has a considerably higher LC_{50} (i.e. it is more toxic) than the surrounding compounds. This chemical was further investigated in order to understand the reason for its outlying behaviour. Two hypotheses were made: a) there might be a large variability (or an error) in the experimental LC_{50} used in the training set; or b) the compound might be misplaced in the score plot because of erroneous values of the molecular descriptors. In this regard, one model descriptor, i.e. *MLOGP*, is an estimate of a molecular property and could be affected by errors. Hence, additional experimental LC_{50} and LogP values were searched for. The results are collated in Table 5.4. The experimental LogP found in the ECHA database [*ECHA*] suggests a higher lipophilicity, which would slightly move the compound towards the left-hand side of the score plot. The additional LC_{50} values are almost five logarithmic units lower than the value in the dataset. This could indicate either a large variability in the measured toxicity values or the presence of erroneous values. With the new toxicity values, 2-Mercaptopyridine N-oxide sodium salt is no longer an outlier.
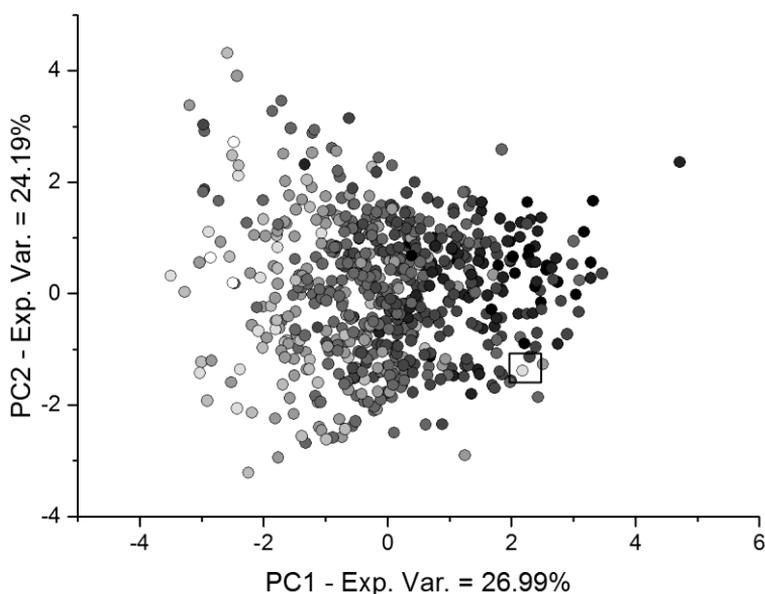

Figure 5.4. Score plot of the training set. Compounds are coloured based on the toxicity, which increases from black to white.

Table 5.4. LogP and toxicity values for 2-Mercaptopyridine N-oxide sodium salt.

	LogP	**LC$_{50}$ [-Log(mol/L)]**
Values in dataset	-1.26	8.13
Additional experimental values	0.0023	3.40-3.46

5.7 Investigating the effect of heterogeneity and experimental variability

The dataset for acute toxicity towards the *Pimephales promelas* (908 organic molecules) was generated gathering data from different sources, including the MED-Duluth fathead minnow database [Russom et al., (1997)]. This database represents a set of high quality data measured in the same division and often used alone to derive QSAR models. An investigation was undertaken to screen the two sets of data. In particular, the objective was to assess whether (and to what extent) QSAR models developed from the sole MED-Duluth data would achieve higher statistics because of the lower number of molecules (566 *versus* 908, respectively) and lower experimental variability.

After treating the MED-Duluth database with the dissociation algorithm and filtering out inorganic compounds, 566 molecules were retained. Models were developed on these data, used as training set, and then externally validated using the compounds from the dataset generated in this study not present in the MED-Duluth set. Models were developed using DRAGON descriptors and *k*NN following the same procedure outlined in paragraph 5.3. The results are collated in Table 5.5. The statistics in fitting and internal validation seemed to show that higher prediction accuracy could be obtained using only the MED-Duluth data. However, it was not determined whether this was due to the lower number of training compounds or the lower variability. The external validation gave different results. Indeed, the accuracy of the predictions of the model calibrated on the MED-Duluth data for external molecules was considerably lower. This might be an effect of the multiple sources of data for the molecules external to the MED-Duluth dataset. Additionally, the percentage of compounds out of AD was also higher, which could be an indication of limited applicability deriving from the smaller training set. On the contrary, the MICHEM model calibrated on the dataset defined in this study gave lower statistics

Table 5.5. Comparison of the best QSAR models calibrated on the MED-Duluth database and the MICHEM model.

Dataset[a]	p^b	distance	k^c	R^2	Q^2_{cv}	Q^2_{ext}	% out AD train[d]	% out AD test[e]
Pim.pro.	6	Jaccard-Tanimoto	6	0.73	0.74	0.77	33%	27%
Duluth	8	Jaccard-Tanimoto	4	0.78	0.79	0.62	32%	51%

[a] *Pim. pro.*: dataset defined in this study (726 training /182 test compounds); Duluth: MED-Duluth database (566 compounds, tested on the 349 compounds from the *Pim. pro.* dataset not present in the MED-Duluth; [b] number of model descriptors; [c] number of nearest neighbours; [d] percentage of training compounds out of AD; [e] percentage of test compounds out of AD.

in fitting and internal validation, but the performance was stable also on the test set, in terms of both accuracy (Q^2_{ext}) and applicability (percentage of molecules out of AD).

5.8 Comparison with literature models

A number of QSAR models from the literature for the prediction of acute toxicity towards *Pimephales promelas* were reported in paragraph 1.4.2. The complete discussion of literature models and comparison with the MICHEM model is carried out in the scientific publication in Appendix III. Here, a summary is given.

The MICHEM model was calibrated on a dataset that was larger than the ones used in the literature and combined data from different sources. These aspects, presumably, determined greater difficulty for modelling due to the higher structural diversity and the variability in the data. On the other side, the results reported in paragraph 5.7 seemed also to indicate that the larger training set granted a wider applicability to external molecules (lower percentage of molecules out of AD in the test set compared to the model developed on the MED-Duluth data).

Differences in datasets and validation procedures make the comparison of QSAR models not straightforward. The largest statistics in cross-validation were obtained through the definition of a strict AD assessed on a similarity basis (Q^2_{cv} equal to 0.87) [Schüürmann et al., (2011)]. Linear regression models achieved higher statistics on individual clusters of chemicals, but lacked an external validation to test the whole procedure (clustering and regression) [Klopman et al., (2000); Colombo et al., (2008)]. It was noticed that the statistical performance of MICHEM model was comparable to that of literature models calibrated on the largest datasets, especially regarding external validation. The highest prediction accuracies were provided by the SVR model of Wang et al., (2010) and the PNN model of Niculescu et al., (2004)

(Q^2_{ext} equal to 0.80 and 0.78, respectively), which are similar to the value achieved by the MICHEM model with the 'Strict' distance threshold (0.77) on a larger test set.

In the end, considering the challenges of the larger number of molecules and the variability of the data, the low number of molecular descriptors, the simple modelling algorithm and the statistical performance, it was concluded that the MICHEM model may be considered satisfactory.

5.9 Compliance with the OECD principles

In paragraph 1.3 it was reported that the applicability of QSAR models within REACH depends on the fulfilment of the five OECD principles. Thus, it is worth to check whether, and to what extent, the MICHEM model complies with the aforementioned principles.

The endpoint was defined as the concentration that kills 50% of test fish after 96 hours of exposure. It should be recalled that data were gathered from different sources where different conditions may have been used (this information was not available for all the records).

The algorithm of the model (kNN) was also described in its mathematical details. For each prediction, the Jaccard-Tanimoto distance for real data was used to identify the six nearest neighbours in the training set. The experimental toxicities of the neighbours were used to provide a prediction computed as weighted mean. Additional characteristics that should be valued by regulators, and therefore facilitate the acceptability of the model, are:

- The simplicity of the algorithm.
- The fact that by considering only the six most similar training molecules, the model should be able to deal with the different modes of action present in the dataset. The hypothesis is that the query compound and the nearest neighbours act via the same mechanism.

The applicability domain (AD) was assessed by comparing the average distance from the six nearest neighbours with a fixed threshold. The rationale behind this approach, justified by the analysis of the residuals, is that high distance (i.e. low similarity) is symptom of diverse structure and, consequently, different toxicity. The use of the distance threshold allows to avoid predictions based on dissimilar structures, likely to be inaccurate. The information regarding the structure and toxicity of the nearest neighbours and the performance of the model in the chemical

region of interest constitute additional elements that can support the evaluation of the uncertainty associated with each prediction

The robustness of the model was evaluated by means of two internal validation techniques: five-fold cross-validation and random splits 80%-20% repeated 1000 times. The predictivity was assessed by means of external validation on 182 compounds (20% of the initial dataset) never used for descriptor selection, nor model calibration.

Eventually, the six molecular descriptors of the model showed evident trend with the experimental toxicity in a principal component analysis (PCA). The descriptors encoded information on lipophilicity, presence of heteroatoms, and electrophilic functional groups.

In conclusion, the model seems to comply with the five OECD principles and, consequently, its application within REACH should be possible.

CHAPTER 6
Conclusions and perspectives

'There can be no ultimate statements in science: there can be no statements in science which can not be tested, and therefore none which cannot in principle be refuted, by falsifying some of the conclusions which can be deduced from them.'

Karl Popper, Logik Der Forschung, 1935

Conclusions

The European REACH regulation was introduced with the main goal of protecting both human health and the environment. Information on acute and chronic toxicity towards aquatic invertebrates (*Daphnia* preferred species), fish and algae is required by REACH for the registration of substances (Annexes VII-IX). Moreover, the information on the aquatic toxicity can be used for the PBT assessment in regards to the toxicity (T) criterion. In fact, chemical substances released into the environment can eventually end up in the aquatic ecosystem, where they can exert toxic effects that can spread from a single organism to the entire aquatic community, threatening the survival of other species as well.

Ethical issues were taken on by REACH and are visible in the strive to reduce animal testing. Among the tools made available to pursue this goal, *in silico* methods, including quantitative structure-activity relationships (QSARs), were foreseen. The applicability of QSAR models to fill data gaps of substances lacking experimental data depends on the compliance with the five OECD principles for validation. Existing QSAR models for the prediction of acute toxicity towards *Daphnia magna* and *Pimephales promelas* were not always developed to comply with the OECD principles. Consequently, some may face difficulties if seeking regulatory application, thus leaving space for the development of novel models.

Conclusions

This project aimed at developing QSARs for the prediction of the acute toxicity towards *Daphnia magna* and *Pimephales promelas* that can be applied in the framework of REACH to fill data gaps. Model development was, consequently, carried out with the intent to comply to the full extent with the OECD principles in order to guarantee acceptance of model predictions from the regulators.

The data preparation phase highlighted a large variability of the experimental toxicity values, which inevitably affected the performance of the QSAR models. Highly consistent data, especially regarding the toxicity towards *Daphnia magna*, are needed in order to obtain more accurate models. Nevertheless, data from different sources were merged together with the aim to define large heterogeneous datasets, needed to grant a wide applicability of the models. This was motivated by the will to expand the spectrum of potential users and do not restrict the applicability to some specific chemical classes.

The heterogeneity of the structures in the dataset translated into concomitant presence of different modes of action (MoAs). Since the datasets were modelled altogether, the selected molecular descriptors did not provide precise mechanistic information for each mode of action, but described aspects that are more general.

Different regression methods (linear, nonlinear, local) were used to establish the relationships. Global models, i.e. models were the estimation of parameters and/or coefficients is based on the entire training set, did not achieve satisfactory statistics regardless of their complexity (linear or nonlinear). The hypothesis to explain this behaviour is related mainly to the simultaneous presence of different modes of action (MoAs) and the fact that the description of the molecular structure and its properties provided by molecular descriptors is an approximation. Theoretically, if one could perfectly describe the molecular structure and its properties, a unique functional relationship to explain the toxicity might be derived. Since this is not the case, one can only build approximated relationships. In this context, global models are more affected by the presence of diverse MoAs because all the compounds contribute to the estimation of model parameters. Local similarity-based models, on the other side, still feel the effect of the diverse dataset (the optimisation of model parameters depends on the overall performance) but allow a better estimation of the local relationship between response and descriptors because only a small neighbourhood is considered. Indeed, the best results were obtained by means of the local k-nearest neighbours (kNN) method.

The most suited approach to assess the applicability domain (AD) was based on the comparison of the average distance (or similarity) from the k closest molecules in the training set with a fixed threshold. This approach is appropriate for use in combination with the local similarity-based regression method. On the contrary, the combination of the local kNN with global AD approaches (i.e. approaches that always consider the entire training set), such as the leverage, did not seem appropriate.

The developed models achieved satisfactory performance in both internal and external validation (taking into account the experimental variability), thus proving the efficacy of the local similarity-based approach. These performances were obtained by means of the definition of a relatively narrow applicability domain.

Regarding the study of acute toxicity towards *Daphnia magna*, the appropriateness of the selected descriptors was highlighted by the satisfactory results of the thorough external validation. The extended MICHEM model was more stable and had a wider applicability than the original one by virtue of the greater number of compounds in the training set. Binary fingerprints were less effective presumably because they could not properly encode the effect of long aliphatic chains on lipophilicity, which, in turn, largely affects the toxicity. The *consensus* models allowed broadening the applicability domain and achieving predictions that were more accurate.

Regarding the study of acute toxicity towards *Pimephales promelas*, it was shown that data with lower experimental variability (the MED-Duluth database) allowed calibrating a model characterised by higher statistics in internal validation. This model, however, suffered of (a) a low applicability to external molecules (large percentage of test molecules out of AD), presumably as a consequence of the small training set and (b) low predictivity (lower statistics for the external validation) because the test set included data from different sources affected by experimental variability. The MICHEM model achieved more balanced performance in internal and external validation and showed a greater applicability to test compounds.

The models developed in this study comply with the OECD principles:

- OECD principle one: the endpoint was specified (LC_{50} over a test duration of 48 and 96 hours for *Daphnia magna* and *Pimephales promelas*, respectively), even though data measured under different conditions were merged together.

- OECD principle two: the algorithm (kNN) was described by specifying all the parameters used to calibrate the models (data scaling, distance, number of

nearest neighbours (k), weights used to compute the prediction as weighted mean).

- OECD principle three: the applicability domain was systematically assessed by comparing the average distance from the nearest neighbours with a fixed threshold.

- OECD principle four: the robustness of the models was estimated by means of appropriate internal validation techniques (five-fold and repeated random splits cross-validation) and the predictivity by means of a test set.

- OECD principle five: the correlations between model descriptors and toxicity were investigated by means of principal component analyses (PCA) and the interpretation of model descriptors was made trying to relate it with general knowledge on aquatic toxicity. However, since the datasets were heterogeneous, a precise mechanistic interpretation was not possible because the molecular descriptors described aspects that are more general.

Additional beneficial characteristics of the developed models are related to the simplicity in terms of both algorithm (kNN) and type of molecular descriptors (two-dimensional), which are calculated on the SMILES strings and do not require geometry optimisation, which can be time consuming and introduce inconsistency if different algorithms are used. Thanks to the nature of the models, a number of information can be provided in order to further assess the uncertainty associated with each prediction, e.g. structure and experimental toxicity of the neighbours in the training set, performance of the model in the chemical region of interest, information on the performance of the model on individual functional groups. All these elements should enhance the confidence in the use of the models and facilitate their acceptance for regulatory purposes.

Future perspectives

A number of actions can be devised as a follow up of the present study. The most urgent action, needed to respond the issues behind the project and give practical application to the results obtained, is the implementation and distribution of the models in a software (or online platform) that can be used by registrants to meet the 2018 REACH deadline. In doing this, the output of the software can be tailored to the needs of registrants and regulators in order to provide all the information useful to assess the reliability of each prediction. This information should then be elaborated

by means of expert judgement in order to obtain a final assessment. For instance, predictions could be converted from –Log(mol/L) (used for modelling) to mg/L and then compared with the thresholds used in the REACH and CLP regulations to define categories of toxicity. This means that quantitative predictions would be discretised and used to classify compounds in classes of toxicity. In doing this, attention should be paid to avoid erroneous classifications. For example, a compound considered not harmful but whose quantitative prediction is close to the threshold for classification in the harmful class should be treated with care. On the other hand, predictions far from class thresholds could be considered reliable even if potentially affected by large errors, if the classification would remain unchanged. The estimation of the mean prediction errors of the model (RMSEP) can help in this sense to quantify the potential error that affects each prediction.

The reliability of each prediction (both quantitative and categorical) could be assessed by means of different information, such as the structure and experimental toxicity of the neighbours, the overall average error of the model, the average error on the neighbours and on compounds possessing the same functional groups as the query molecule. Additional information could come from existing classification schemes on categories of toxicity or modes of action applied to both the query chemical and the neighbours. Low average errors, similarity in the structures and classification to the same mode of action should enhance confidence in the prediction.

Regarding further modelling, different directions may be followed. An investigation can be made to check whether the selected descriptors with more difficult interpretation may be replaced by more simple ones, without considerably affecting the performance of the models. Models based on fragmental descriptors, or combinations of relevant molecular properties (such as the LogP) and fragmental descriptors, could be investigated. The advantage would be a more direct relationship between model descriptors and chemical structure. Due to the heterogeneous nature of the datasets, it is expected that a large number of fragments need to be included to obtain a proper coverage of the structures in the dataset. In this case, attention should be paid to overfitting.

Fragmental descriptors could also be used to define correction factors that could intervene in specific cases to modify the predictions given by a model that accounts for trends that are more general.

The information on the mode of action could be explicitly taken into account by defining local models associated with each MoA. This strategy would then require

a preliminary step consisting in the assignation of query molecules to one or more MoAs. The approach was already tried in the scientific literature, but it does not seem clear how inaccurate a prediction can be if an erroneous assignation is made in the first step.

Even though simple models are preferred, further attempts could be made with methods that are more complex. One possibility would be to couple them with AD approaches more suitable than the leverage to describe the distribution of compounds in the descriptors space. However, the use of such methods should be limited to cases where evidence of considerably higher prediction accuracy is provided.

Bibliography

Andersson, M., 2009. A comparison of nine PLS1 algorithms. *J. Chemometrics* 23, 518–529.

Ankley, G.T., Villeneuve, D.L., 2006. The fathead minnow in aquatic toxicology: past, present and future. *Aquat. Toxicol.* 78, 91–102.

ASTM. American Society for Testing and Materials, 2007. Guide for conducting acute toxicity tests on test materials with fishes, macroinvertebrates, and amphibians (No. E729).

Balaban, A.T., Ciubotariu, D., Medeleanu, M., 1991. Topological indices and real number vertex invariants based on graph eigenvalues or eigenvectors. *J. Chem. Inf. Comput. Sci.* 31, 517–523.

Barron, M.G., Jackson, C.R., Awkerman, J.A., 2012. Evaluation of in silico development of aquatic toxicity species sensitivity distributions. *Aquat. Toxicol.* 116–117, 1–7.

Basak, S.C., Gieschen, D.P., Magnuson, V.R., 1984. A quantitative correlation of the LC_{50} values of esters in *Pimephales promelas* using physicochemical and topological parameters. *Environ. Toxicol. Chem.* 3, 191–199.

Basak, S.C., Magnuson, V.R., 1983. Molecular topology and narcosis. A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC). *Arzneim.-ForschungDrug Res.* 33, 501–503.

Bearden, A.P., Schultz, T.W., 1997. Structure-activity relationships for *Pimephales* and *Tetrahymena*: A mechanism of action approach. *Environ. Toxicol. Chem.* 16, 1311–1317.

Benfenati, E., 2010. The CAESAR project for *in silico* models for the REACH legislation. *Chem. Cent. J.* 4, I1.

Bernot, R.J., Brueseke, M.A., Evans-White, M.A., Lamberti, G.A., 2009. Acute and chronic toxicity of imidazolium-based ionic liquids on *Daphnia magna. Environ. Toxicol. Chem.* 24, 87–92.

Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B., 2007. KNIME: The Konstanz Information Miner, in: *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer, 319–326.

Boeije, G.M., Cano, M.L., Marshall, S.J., Belanger, S.E., Van Compernolle, R., Dorn, P.B., Gümbel, H., Toy, R., Wind, T., 2006. Ecotoxicity quantitative structure–activity relationships for alcohol ethoxylate mixtures based on substance-specific toxicity predictions. *Ecotoxicol. Environ. Saf.* 64, 75–84.

Bolton, E.E., Wang, Y., Thiessen, P.A., Bryant, S.H., 2008. PubChem: integrated platform of small molecules and biological activities. *Annu. Rep. Comput. Chem.* 4, 217–241.

Bondy, J.A., Murty, U.S.R., 2008. Graph Theory. Springer.

Bradbury, S.P., Carlson, R.W., Niemi, G.J., Henry, T.R., 1991. Use of respiratory-cardiovascular responses of rainbow trout (*Oncorhynchus mykiss*) in identifying acute toxicity syndromes in fish: Part 4. Central nervous system seizure agents. *Environ. Toxicol. Chem.* 10, 115–131.

Bradbury, S.P., Henry, T.R., Niemi, G.J., Carlson, R.W., Snarski, V.M., 1989. Use of respiratory-cardiovascular responses of rainbow trout (*Salmo gairdneri*) in identifying acute toxicity syndromes in fish: Part 3. Polar narcotics. *Environ. Toxicol. Chem.* 8, 247–261.

Broderius, S.J., Kahl, M.D., 1985. Acute toxicity of organic chemical mixtures to the fathead minnow. *Aquat. Toxicol.* 6, 307–322.

Broderius, S.J., Kahl, M.D., Hoglund, M.D., 1995. Use of joint toxic response to define primary mode of toxic action for diverse industrial organic chemicals. *Environ. Toxicol. Chem.* 14, 1591 – 1605.

Casalegno, M., Benfenati, E., Sello, G., 2005. An automated group contribution method in predicting aquatic toxicity: the diatomic fragment approach. *Chem. Res. Toxicol.* 18, 740–746.

Casalegno, M., Sello, G., 2013. Determination of toxicant mode of action by augmented top priority fragment class. *J. Chem. Inf. Model.* 53, 1113–1126.

Cassani, S., Kovarich, S., Papa, E., Roy, P.P., Rahmberg, M., Nilsson, S., Sahlin, U., Jeliazkova, N., Kochev, N., Pukalov, O., 2013a. Evaluation of CADASTER QSAR models for the aquatic toxicity of (benzo) triazoles and prioritisation by consensus prediction. *Altern. Lab. Anim. ATLA* 41, 49–64.

Cassani, S., Kovarich, S., Papa, E., Roy, P.P., van der Wal, L., Gramatica, P., 2013b. Daphnia and fish toxicity of (benzo)triazoles: validated QSAR models, and interspecies quantitative activity-activity modelling. *J. Hazard. Mater.* 258-259, 50-60.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: a library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol*. 2, 27:1–27:27.

ChemProp, Chemical Properties Estimation Software System, 2013. UFZ Department of Ecological Chemistry, Leipzig.

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

Chen, J., Liao, Y., Zhao, Y., Wang, L., Lu, G., Zhao, T., 1996. Quantitative structure-activity relationships and mixture toxicity studies of heterocyclic nitrogen compounds. *Bull. Environ. Contam. Toxicol.* 57, 77–83.

Chen, T., Morris, J., Martin, E., 2007. Gaussian process regression for multivariate spectroscopic calibration. *Chemom. Intell. Lab. Syst.* 87, 59–71.

Coe, T.S., Hamilton, P.B., Griffiths, A.M., Hodgson, D.J., Wahab, M.A., Tyler, C.R., 2009. Genetic variation in strains of zebrafish (*Danio rerio*) and the implications for ecotoxicology studies. *Ecotoxicology* 18, 144–150.

Colombo, A., Benfenati, E., Karelson, M., Maran, U., 2008. The proposal of architecture for chemical splitting to optimize QSAR models for aquatic toxicity. *Chemosphere* 72, 772–780.

Consonni, V., Ballabio, D., Todeschini, R., 2009. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* 49, 1669–1678.

Consonni, V., Todeschini, R., Basic requirements for valid molecular descriptors. Tutorial on *www.moleculardescriptors.eu.*

Costanzo, S.D., Watkinson, A.J., Murby, E.J., Kolpin, D.W., Sandstrom, M.W., 2007. Is there a risk associated with the insect repellent DEET (*N,N*-diethyl-*m*-toluamide) commonly found in aquatic environments? *Sci. Total Environ.* 384, 214 – 220.

Cronin, M.T.D., Netzeva, T.I., Dearden, J.C., Edwards, R., Worgan, A.D.P., 2004. Assessment and modeling of the toxicity of organic chemicals to *Chlorella vulgaris*: development of a novel database. *Chem. Res. Toxicol.* 17, 545–554.

Cronin, M.T.D., Zhao, Y.H., Yu, R.L., 2000. pH-Dependence and QSAR analysis of the toxicity of phenols and anilines to *Daphnia magna*. *Environ. Toxicol.* 15, 140–148.

Cui, X., Wang, Z., Zhang, Z., Yuan, X., Harrington, P. de B., 2008. QSAR Study on the toxicity of phenols for fathead minnows by using support vector machine and neural networks. *IEEE*, 134–138.

Dai, J., Wang, L., 2000. Quantitative structure–toxicity relationships for derivates of benzanilides to *Daphnia magna*. *Bull. Environ. Contam. Toxicol.* 65, 366–374.

Davies, J., Ward, R.S., Hodges, G., Roberts, D.W., 2004. Quantitative structure-activity relationship modeling of acute toxicity of quaternary alkylammonium sulfobetaines to *Daphnia magna*. *Environ. Toxicol. Chem.* 23, 2111–2115.

Daylight Chemical Information Systems, Inc., SMILES - A Simplified Chemical Language. *http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html*

Daylight Chemical Information Systems, Inc., Fingerprints - Screening and Similarity. *http://www.daylight.com/dayhtml/doc/theory/theory.finger.html*

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

Deneer, J.W., Van Leeuwen, C.J., Seinen, W., Maas-Diepeveen, J.L., Hermens, J.L.M., 1989. QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna, Chlorella pyrenoidosa* and *Photobacterium phosphoreum*. *Aquat. Toxicol.* 15, 83–98.

Devillers, J., 2005. A new strategy for using supervised artificial neural networks in QSAR. *SAR QSAR Environ. Res.* 16, 433–442.

Devillers, J., Chambon, P., 1986. Acute toxicity and QSAR of chlorophenols on *Daphnia magna*. *Bull. Environ. Contam. Toxicol.* 37, 599–605.

Devillers, J., Chambon, P., Zakarya, D., 1987. A predictive structure-toxicity model with *Daphnia magna*. *Chemosphere* 16, 1149–1163.

Di Delupis, G.D., Macrí, A., Civitareale, C., Migliore, L., 1992. Antibiotics of zootechnical use: effects of acute high and low dose contamination on *Daphnia magna* Straus. *Aquat. Toxicol.* 22, 53–59.

DRAGON 6 (Software for Molecular Descriptor Calculation), 2012. Talete srl.

Drummond, R.A., Russom, C.L., 1990. Behavioral toxicity syndromes: a promising tool for assessing toxicity mechanisms in juvenile fathead minnows. *Environ. Toxicol. Chem.* 9, 37–46.

Drummond, R.A., Russom, C.L., Geiger, D.L., DeFoe, D.L., 1986. Behavioral and morphological changes in fathead minnow (*Pimephales promelas*) as diagnostic endpoints for screening chemicals according to mode of action, in: *Aquatic Toxicology and Environmental Fate*. American Society for Testing and Materials, Philadelphia, PA, 415–435.

ECETOC. European Centre for Ecotoxicology and Toxicology Of Chemicals, 2003. TR 091 - ECETOC Aquatic Toxicity (EAT) database.

ECHA, European CHemical Agency (ECHA). *http://echa.europa.eu/*.

ECHA, European Chemicals Agency, 2008. Guidance on information requirements and chemical safety assessment Chapter R.6: QSARs and grouping of chemicals.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. *Ann. Stat.* 32, 407–499.

Efroymson, M.A., 1960. Multiple regression analysis, in: *Mathematical Methods for Digital Computers*. Wiley, New York.

Eldred, D.V., Weikel, C.L., Jurs, P.C., Kaiser, K.L.E., 1999. Prediction of fathead minnow acute toxicity of organic compounds from molecular structure. *Chem. Res. Toxicol.* 12, 670–678.

Environment Canada, 1996. Biological Test Method - Acute lethality test using Daphnia spp. (No. EPS 1/RM/11).

Environment Canada, 2011. Biological Test Method - Test of larval growth and survival using fathead minnows (No. EPS 1/RM/22).

EPI Suite, 2012. U.S. E.P.A.

Ertl, P., Rohde, B., Selzer, P., 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.* 43, 3714–3717.

Esbensen, K.H., Geladi, P., 2010. Principles of proper validation: use and abuse of re-sampling for validation. *J. Chemometrics* 24, 168–187.

Faucon, J.C., Bureau, R., Faisant, J., Briens, F., Rault, S., 2001. Prediction of the *Daphnia* acute toxicity from heterogeneous data. *Chemosphere* 44, 407–422.

Ferrari, B., Mons, R., Vollat, B., Fraysse, B., Paxēaus, N., Giudice, R.L., Pollio, A., Garric, J., 2004. Environmental risk assessment of six human pharmaceuticals: are the current environmental risk assessment procedures sufficient for the protection of the aquatic environment? *Environ. Toxicol. Chem.* 23, 1344–1354.

Foit, K., Kaske, O., Liess, M., 2012. Competition increases toxicant sensitivity and delays the recovery of two interacting populations. *Aquat. Toxicol.* 106–107, 25–31.

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

Freidig, A.P., Hermens, J.L., 2000. Narcosis and chemical reactivity in acute fish toxicity QSARs. *Quant. Struct.-Act. Relat.* 19, 547-553.

Furuhama, A., Aoki, Y., Shiraishi, H., 2012. Development of ecotoxicity QSAR models based on partial charge descriptors for acrylate and related compounds. *SAR QSAR Environ. Res.* 23, 731–749.

Ghose, A.K., Crippen, G.M., 1986. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* 7, 565–577.

Ghose, A.K., Crippen, G.M., 1987. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. *J. Chem. Inf. Comput. Sci.* 27, 21–35.

Ghose, A.K., Viswanadhan, V.N., Wendoloski, J.J., 1998. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem.* A 102, 3762–3772.

Gini, G., Craciun, M.V., Konig, C., Benfenati, E., 2004. Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity. *J. Chem. Inf. Model.* 44, 1897–1902.

Golbamaki, A., Cassano, A., Lombardo, A., Moggio, Y., Colafranceschi, M., Benfenati, E., 2014. Comparison of *in silico* models for prediction of *Daphnia magna* acute toxicity. *SAR QSAR Environ. Res.* 25, 673–694.

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

Goodarzi, M., Freitas, M.P., Jensen, R., 2009. 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. *Chemom. Intell. Lab. Syst.* 98, 123–129.

Grigor'ev, V.Y., Razdol'skii, A.N., Zagrebin, A.O., Tonkopii, V.D., Raevskii, O.A., 2014. QSAR classification models of acute toxicity of organic compounds with respect to *Daphnia magna*. *Pharm. Chem. J.* 48, 242–245.

Gunatilleka, A.D., Poole, C.F., 1999. Models for estimating the non-specific aquatic toxicity of organic compounds. *Anal. Commun.* 36, 235–242.

Hall, L., Kier, L.B., Phipps, G., 1984. Structure-activity relationships studies on the toxicities of benzene derivatives: I. an additivity model. *Environ. Toxicol. Chem.* 3, 355–365.

Hansch, C., Maloney, P.P., Fujita, T., Muir, R.M., 1962. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature* 194, 178–180.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning, 2nd ed. Springer.

Hermens, J., Canton, H., Janssen, P., De Jong, R., 1984. Quantitative structure-activity relationships and toxicity studies of mixtures of chemicals with anaesthetic potency: acute lethal and sublethal toxicity to *Daphnia magna*. *Aquat. Toxicol.* 5, 143–154.

Hewitt, M., Cronin, M.T.D., Madden, J.C., Rowe, P.H., Johnson, C., Obi, A., Enoch, S.J., 2007. Consensus QSAR models: do the benefits outweigh the complexity? *J. Chem. Inf. Model.* 47, 1460–1468.

He, Y., Wang, L., 1996. Quantitative structure-activity relationships for studying alkyl (1-phenylsulfonyl) cyckloalkane-carboxylates, *J. Environ. Sci.* 8, 157–166.

Hodges, G., Roberts, D.W., Marshall, S.J., Dearden, J.C., 2006. The aquatic toxicity of anionic surfactants to *Daphnia magna* - A comparative QSAR study of linear alkylbenzene sulphonates and ester sulphonates. *Chemosphere* 63, 1443 – 1450.

Holland, J.H., 1992. Adaptation in natural and artificial systems. MIT Press.

Horn, O., Nalli, S., Cooper, D., Nicell, J., 2004. Plasticizer metabolites in the environment. *Water Res.* 38, 3693–3698.

In, Y.-Y., Lee, S.-K., Kim, P.-J., No, K.-T., 2012. Prediction of acute toxicity to fathead minnow by local model based QSAR and global QSAR approaches. *Bull. Korean Chem. Soc.* 33, 613–619.

Ismail Hossain, M., Samir, B.B., El-Harbawi, M., Masri, A.N., Abdul Mutalib, M.I., Hefter, G., Yin, C.-Y., 2011. Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities. *Chemosphere* 85, 990–994.

ISO. International Organization for Standardization, 2010. ISO 7346:1996 -- Determination of the acute lethal toxicity of substances to a freshwater fish [*Brachydanio rerio* Hamilton-Buchanan (Teleostei, Cyprinidae)]. Parts 1-2-3.

ISO. International Organization for Standardization, 2012. ISO 6341:2012 -- Determination of the inhibition of the mobility of *Daphnia magna* Straus (Cladocera, Crustacea) -- Acute toxicity test.

Jaccard, P., 1908. Nouvelles recherches sur la distribution florale. *Bull. Société Vaudoise Sci. Nat.* 44, 223–270.

Jemec, A., Tišler, T., Drobne, D., Sepčić, K., Fournier, D., Trebše, P., 2007. Comparative toxicity of imidacloprid, of its commercial liquid formulation and of diazinon to a non-target arthropod, the microcrustacean *Daphnia magna*. *Chemosphere* 68, 1408–1418.

Jensen, G.E., Nikolov, N.G., Wedebye, E.B., Ringsted, T., Niemelä, J.R., 2011. QSAR models for anti-androgenic effect – a preliminary study. *SAR QSAR Environ. Res.* 22, 35–49.

Jolliffe, I., 2005. Principal Component Analysis, in: Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd.

Kaiser, K.L.E., Niculescu, S.P., 2001. Modeling acute toxicity of chemicals to *Daphnia magna*: A probabilistic neural network approach. *Environ. Toxicol. Chem.* 20, 420–431.

Karabunarliev, S., Mekenyan, O.G., Karcher, W., Russom, C.L., Bradbury, S.P., 1996a. Quantum-chemical descriptors for estimating the acute toxicity of

electrophiles to the fathed minnow (*Pimephales promelas*): an analysis based on molecular mechanisms. *Quant. Struct.-Act. Relat.* 15, 302–310.

Karabunarliev, S., Mekenyan, O.G., Karcher, W., Russom, C.L., Bradbury, S.P., 1996b. Quantum-chemical descriptors for estimating the acute toxicity of substituted benzenes to the guppy (*Poecilia reticulata*) and fathead minnow (*Pimephales promelas*). *Quant. Struct.-Act. Relat.* 15, 311–320.

Kar, S., Roy, K., 2010. QSAR modeling of toxicity of diverse organic chemicals to *Daphnia magna* using 2D and 3D descriptors. *J. Hazard. Mater.* 177, 344–351.

Katritzky, A.R., Slavov, S.H., Stoyanova-Slavova, I.S., Kahn, I., Karelson, M., 2009a. Quantitative structure–activity relationship (QSAR) modeling of EC_{50} of aquatic toxicities for *Daphnia magna*. *J. Toxicol. Environ. Health A* 72, 1181–1190.

Kennard, R.W., Stone, L.A., 1969. Computer Aided Design of Experiments. *Technometrics* 11, 137–148.

Kim, K.J., Cho, S.B., 2006. A comprehensive overview of the applications of artificial life. *Artif. Life* 12, 153–182.

Klaassen, C., 2001. Casarett and Doull's Toxicology: the basic science of poisons, sixth. ed. The McGraw-Hill Copanies, Inc.

Klopman, G., Saiakhov, R., Rosenkranz, H.S., 2000. Multiple computer-automated structure evaluation study of aquatic toxicity II. Fathead minnow. *Environ. Toxicol. Chem.* 19, 441–447.

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

KOWWIN, 2010. U.S. E.P.A.

Kühne, R., Ebert, R.-U., von der Ohe, P.C., Ulrich, N., Brack, W., Schüürmann, G., 2013. Read-across prediction of the acute toxicity of organic compounds toward the water flea *Daphnia magna*. *Mol. Inform.* 32, 108–120.

Kyriakopoulou, K., Anastasiadou, P., Machera, K., 2009. comparative toxicities of fungicide and herbicide formulations on freshwater and marine species. *Bull. Environ. Contam. Toxicol.* 82, 290–295.

Leadscope Enterprise^{TM}. Leadscope Inc., Columbus, OH, USA.

Leardi, R., 2003. Nature-inspired methods in chemometrics: genetic algorithms and artificial neural networks. Elsevier.

Leardi, R., González, A.L., 1998. Genetic algorithms applied to feature selection in PLS regression: how and when to use them. *Chemom. Intell. Lab. Syst.* 41, 195 – 207.

Lindgren, F., Hansen, B., Karcher, W., Sjöström, M., Eriksson, L., 1996. Model validation by permutation tests: applications to variable selection. *J. Chemometrics* 10, 521–532.

Liu, X., Wang, B., Huang, Z., Han, S., Wang, L., 2003. Acute toxicity and quantitative structure–activity relationships of α-branched phenylsulfonyl acetates to *Daphnia magna*. *Chemosphere* 50, 403–408.

Lozano, S., Halm-Lemeille, M.-P., Lepailleur, A., Rault, S., Bureau, R., 2010. Consensus QSAR related to global or MOA models: application to acute toxicity for fish. *Mol. Inform.* 29, 803–813.

Maes, G.E., Raeymaekers, J.A.M., Pampoulie, C., Seynaeve, A., Goemans, G., Belpaire, C., Volckaert, F.A.M., 2005. The catadromous European eel *Anguilla anguilla* (L.) as a model for freshwater evolutionary ecotoxicology: relationship between heavy metal bioaccumulation, condition and genetic variability. *Aquat. Toxicol.* 73, 99–114.

Mansouri, K., Ringsted, T., Ballabio, D., Todeschini, R., Consonni, V., 2013. Quantitative structure-activity relationship models for ready biodegradability of chemicals. *J. Chem. Inf. Model.* 53, 867–878.

Maran, U., Sild, S., Mazzatorta, P., Casalegno, M., Benfenati, E., Romberg, M., 2007. Grid computing for the estimation of toxicity: acute toxicity on fathead minnow (*Pimephales promelas*), in: *Distributed, High-Performance and Grid Computing in Computational Biology*. Springer, 60–74.

Marchini, S., Passerini, L., Hoglund, M.D., Pino, A., Nendza, M., 1999. Toxicity of aryl-and benzylhalides to *Daphnia magna* and classification of their mode of action based on quantitative structure–activity relationship. *Environ. Toxicol. Chem.* 18, 2759–2766.

Marengo, E., Todeschini, R., 1992. A new algorithm for optimal, distance-based experimental design. *Chemom. Intell. Lab. Syst.* 16, 37–44.

Martins, J.C., Saker, M.L., Teles, L.F.., Vasconcelos, V.M., 2007. Oxygen consumption by *Daphnia magna* Straus as a marker of chemical stress in the aquatic environment. *Environ. Toxicol. Chem.* 26, 1987–1991.

Martin, T., Harten, P., Venkatapathy, R., Young, D., 2012. *T.E.S.T.* (Toxicity Estimation Software Tool). U.S. E.P.A.

Marvin, 2012. ChemAxon Ltd.

MATLAB, 2012. MathWorks Inc., Natick, MA, USA.

Mayo-Bean, K., 2012. *ECOSAR*. U.S. E.P.A.

Mazzatorta, P., Benfenati, E., Neagu, C.-D., Gini, G., 2003a. tuning neural and fuzzy-neural networks for toxicity modeling. *J. Chem. Inf. Model.* 43, 513–518.

Mazzatorta, P., Vracko, M., Jezierska, A., Benfenati, E., 2003b. Modeling toxicity by using supervised Kohonen neural networks. *J. Chem. Inf. Model.* 43, 485–492.

McFarland, J.W., 1970. On the Parabolic relation between drug potency and hydrophobicity. *J. Med. Chem.* 13, 1192–1196.

McKim, J.M., Schmieder, P.K., Carlson, R.W., Hunt, E.P., Niemi, G.J., 1987a. Use of respiratory-cardiovascular responses of rainbow trout (*Salmo gairdneri*) in identifying acute toxicity syndromes in fish: Part 1. pentachlorophenol, 2, 4-dinitrophenol, tricaine methanesulfonate and 1-octanol. *Environ. Toxicol. Chem.* 6, 295–312.

McKim, J.M., Schmieder, P.K., Niemi, G.J., Carlson, R.W., Henry, T.R., 1987b. Use of respiratory-cardiovascular responses of rainbow trout (*Salmo gairdneri*) in

identifying acute toxicity syndromes in fish: Part 2. malathion, carbaryl, acrolein and benzaldehyde. *Environ. Toxicol. Chem.* 6, 313–328.

Meyer, H., 1899. Zur Theorie der Alkoholnarkose. *Arch. Für Exp. Pathol. Pharmakol.* 42, 109–118.

Michielan, L., Pireddu, L., Floris, M., Moro, S., 2010. Support vector machine (SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals. *Mol. Inform.* 29, 51–64.

Miller, A.J., 1984. Selection of subsets of regression variables. *J. R. Statist. Soc. A.* 147, 389–425.

Miller, A.J., 2002. Subset selection in regression, 2nd ed. Chapman & Hall/CRC.

Moosus, M., Maran, U., 2011. Quantitative structure–activity relationship analysis of acute toxicity of diverse chemicals to *Daphnia magna* with whole molecule descriptors. *SAR QSAR Environ. Res.* 22, 757–774.

Morgan, H.L., 1965. The generation of a unique machine description for chemical structures-a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* 5, 107–113.

Moriguchi, I., Hirono, S., Liu, Q., Nakagome, I., Matsushita, Y., 1992. Simple method of calculating octanol/water partition coefficient. *Chem. Pharm. Bull.* 40, 127–130.

Moriguchi, I, Hirono, S., Nakagome, I., Hirano, H., 1994. Comparison of reliability of log P values for drugs calculated by several methods. *Chem. Pharm. Bull.* 42, 976–978.

Morrall, D., Belanger, S.., Dunphy, J.., 2003. Acute and chronic aquatic toxicity structure–activity relationships for alcohol ethoxylates. *Ecotoxicol. Environ. Saf.* 56, 381–389.

Mount, D.R., Gulley, D.D., Hockett, J.R., Garrison, T.D., Evans, J.M., 1997. Statistical models to predict the toxicity of major ions to *Ceriodaphnia dubia, Daphnia magna* and *Pimephales promelas* (fathead minnows). *Environ. Toxicol. Chem.* 16, 2009–2019.

NCI/CADD Group, Chemical Identifier Resolver. *http://cactus.nci.nih.gov*

Nendza, M., Müller, M., Wenzel, A., 2014. Discriminating toxicant classes by mode of action: 4. Baseline and excess toxicity. *SAR QSAR Environ. Res.* 25, 393–405.

Nendza, M., Russom, C.L., 1991. QSAR modelling of the ERL-D fathead minnow acute toxicity database. *Xenobiotica* 21, 147 – 170.

Netzeva, T.I., Aptula, A.O., Benfenati, E., Cronin, M.T.D., Gini, G., Lessigiarska, I., Maran, U., Vračko, M., Schüürmann, G., 2005. Description of the electronic structure of organic chemicals using semiempirical and ab initio methods for development of toxicological QSARs. *J. Chem. Inf. Model.* 45, 106–114.

Newsome, L.D., Johnson, D.E., Lipnick, R.L., Broderius, S.J., Russom, C.L., 1991. A QSAR study of the toxicity of amines to the fathead minnow. *Sci. Total Environ.* 109, 537–551.

Newsome, L.D., Johnson, D.E., Nabholz, J.V., 1993. Quantitative structure-activity predictions for amine toxicity to algae and daphnids, in: *Environmental Toxicology and Risk Assessment*: 2nd Volume. STP 1216. American Society for Testing and Materials, Philadelphia, 591–609.

Newsome, L.D., Nabholz, J.V., Kim, A., 1996. Designing aquatically safer chemicals, in: *Designing Safer Chemicals: Green Chemistry for Pollution Prevention*. American Chemical Society, Washington D.C., USA, 172–192.

Niculescu, S.P., Atkinson, A., Hammond, G., Lewis, M., 2004. Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow. *SAR QSAR Environ. Res.* 15, 293–309.

Niemelä, J.R., Wedebye, E.B., Nikolov, N.G., Jensen, G.E., Ringsted, T., 2010. The Advisory list for self-classification of dangerous substances (No. Environmental Project No. 1322 2010). The Danish Environmental Protection Agency.

Nikolov, N., Grancharov, V., Stoyanova, G., Pavlov, T., Mekenyan, O., 2006. Representation of chemical information in OASIS centralized 3D database for existing chemicals. *J. Chem. Inf. Model.* 46, 2537–2551.

Nørgaard, K., Cedergreen, N., 2010. Pesticide cocktails can interact synergistically on aquatic crustaceans. *Environ. Sci. Pollut. Res.* 17, 957–967.

Öberg, T., 2004. A QSAR for baseline toxicity: validation, domain of application, and prediction. *Chem. Res. Toxicol.* 17, 1630–1637.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. *J. Cheminformatics* 3, 33.

Ochoa-Acuña, H., Bialkowski, W., Yale, G., Hahn, L., 2009. Toxicity of soybean rust fungicides to freshwater algae and *Daphnia magna. Ecotoxicology* 18, 440–446.

OECD. The Organization for Economic Development and Co-operation, 1992. Test No. 203: fish, acute toxicity test.

OECD. The Organization for Economic Development and Co-operation, 2004. Test No. 202: Daphnia sp. acute immobilisation test.

OECD. The Organization for Economic Development and Co-operation, 2007. Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models (No. ENV/JM/MONO(2007)2), Series on Testing and Assessment.

O'Hagan, A., Kingman, J.F.C., 1978. Curve fitting and optimal design for prediction. *J. R. Statist. Soc. B.* 40, 1–42.

Overton, C.E., 1901. Studien über die Narkose zugleich ein Beitrag zur allgemeinen Pharmakologie. Gustav Fischer, Jena.

Padmanabhan, J., Parthasarathi, R., Subramanian, V., Chattaraj, P.K., 2006. Group philicity and electrophilicity as possible descriptors for modeling ecotoxicity applied to chlorophenols. *Chem. Res. Toxicol.* 19, 356–364.

Papa, E., Villa, F., Gramatica, P., 2005. Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in *Pimephales promelas* (Fathead Minnow). *J. Chem. Inf. Model.* 45, 1256–1266.

Parkerton, T.F., Konkel, W.J., 2000. Application of quantitative structure-activity relationships for assessing the aquatic toxicity of phthalate esters. *Ecotoxicol. Environ. Saf.* 45, 61 – 78.

Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D., Weinberger, L.E., 1996. Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* 39, 3049–3059.

Pavan, M., Netzeva, T.I., Worth, A.P., 2006. Validation of a QSAR model for acute toxicity. *SAR QSAR Environ. Res.* 17, 147–171.

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. *Philos. Mag. Ser.* 6 2, 559–572.

Picado, A., Chankova, S., Fernandes, A., Simões, F., Leverett, D., Johnson, I., Hernan, R., Pires, A.M., Matos, J., 2007. Genetic variability in *Daphnia magna* and ecotoxicological evaluation. *Ecotoxicol. Environ. Saf.* 67, 406–410.

Pogliani, L., 2000. Modeling with molecular pseudoconnectivity descriptors. A useful extension of the intrinsic I-state concept. *J. Phys. Chem. A* 104, 9029–9045.

Pogliani, L., 2004. Modeling with indices obtained from complete graphs. *Croat. Chem. Acta* 77, 193–201.

Porcelli, C., Boriani, E., Roncaglioni, A., Chana, A., Benfenati, E., 2008a. Regulatory perspectives in the use and validation of QSAR. A case study: DEMETRA model for *Daphnia* toxicity. *Environ. Sci. Technol.* 42, 491–496.

Pritchard, J.B., 1993. Aquatic toxicology: past, present, and prospects. *Environ. Health Perspect.* 100, 249–257.

Qin, W.C., Su, L.M., Zhang, X.J., Qin, H.W., Wen, Y., Guo, Z., Sun, F.T., Sheng, L.X., Zhao, Y.H., Abraham, M.H., 2010. Toxicity of organic pollutants to seven aquatic organisms: effect of polarity and ionization. *SAR QSAR Environ. Res.* 21, 389–401.

Raevskii, O.A., Razdol'skii, A.N., Tonkopii, V.D., Iofina, I.V., Zagrebin, A.O., 2008. Classificatory and quantitative models of the relationship between the structures of chemical compounds and their toxicity for *Daphnia magna*. *Pharm. Chem. J.* 42, 329–334.

Raimondo, S., Vivian, D.N., Delos, C., Barron, M.G., 2008. Protectiveness of species sensitivity distribution hazard concentrations for acute toxicity used in endangered species risk assessment. *Environ. Toxicol. Chem.* 27, 2599–2607.

Randall, W.F., Dennis, W.H., Warner, M.C., 1979. Acute toxicity of dechlorinated DDT, chlordane and lindane to bluegill (*Lepomis macrochirus*) and *Daphnia magna. Bull. Environ. Contam. Toxicol.* 21, 849–854.

Rand, G.M., 1995. Fundamentals of aquatic toxicology: effects, environmental fate and risk assessment, 2nd ed. CRC Press.

Randić, M., 1996. Molecular bonding profiles. *J. Math. Chem.* 19, 375–392.

Rasmussen, C.E., 1996. Evaluation of Gaussian processes and other methods for non-linear regression (PhD thesis). University of Toronto.

Rasmussen, C.E., Williams, C., 2006. GPML.

Ren, S., 2002. Predicting three narcosis mechanisms of aquatic toxicity. *Toxicol. Lett.* 133, 127–139.

Ren, S., Schultz, T.W., 2002. Identifying the mechanism of aquatic toxicity of selected compounds by hydrophobicity and electrophilicity descriptors. *Toxicol. Lett.* 129, 151–160.

Ringsted, T., Nikolov, N., Jensen, G.E., Wedebye, E.B., Niemelä, J., 2009. QSAR models for P450 (2D6) substrate activity. *SAR QSAR Environ. Res.* 20, 309–325.

Roberts, D.W., Roberts, J.F., Hodges, G., Gutsell, S., Ward, R.S., Llewellyn, C., 2013. Aquatic toxicity of cationic surfactants to *Daphnia magna*. *SAR QSAR Environ. Res.* 24, 417–427.

Rogers, D., Hahn, M., 2010. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754.

Royal Society of Chemistry, ChemSpider. *http://www.chemspider.com*

Roy, K., Das, R.N., 2012. QSTR with extended topochemical atom (ETA) indices. 15. Development of predictive models for toxicity of organic chemicals against

fathead minnow using second-generation ETA indices. *SAR QSAR Environ. Res.* 23, 125–140.

Roy, K., Das, R.N., 2013. QSTR with extended topochemical atom (ETA) indices. 16. Development of predictive classification and regression models for toxicity of ionic liquids towards *Daphnia magna*. *J. Hazard. Mater.* 254-255, 166–178.

Roy, K., Das, R.N., Popelier, P.L.A., 2014. Quantitative structure–activity relationship for toxicity of ionic liquids to *Daphnia magna*: aromaticity vs. lipophilicity. *Chemosphere* 112, 120–127.

Russom, C.L., Bradbury, S.P., Broderius, S.J., Hammermeister, D.E., Drummond, R.A., 1997. Predicting modes of toxic action from chemical structure: acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Toxicol. Chem.* 16, 948–967.

Sahigara, F., Ballabio, D., Todeschini, R., Consonni, V., 2013. Defining a novel *k*-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions. *J Cheminformatics* 5, 27.

Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R., 2012. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* 17, 4791–4810.

Sanderson, H., Thomsen, M., 2009. Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q) SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action. *Toxicol. Lett.* 187, 84–93.

Schölkopf, B., Bartlett, P., Smola, A., Wiliamson, R., 1999. Shrinking the tube: a new support vector regression algorithm, in: *Advances in Neural Information Processing Systems 11*.

Schölkopf, B., Smola, A., Williamson, R., Bartlett, P., 2000. New support vector algorithms. *Neural Comput.* 12, 1207–1245.

Schüürmann, G., Ebert, R.-U., Kühne, R., 2011. Quantitative read-across for predicting the acute fish toxicity of organic compounds. *Environ. Sci. Technol.* 45, 4616–4622.

Shen, Q., Jiang, J.-H., Jiao, C.-X., Shen, G., Yu, R.-Q., 2004. Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists. *Eur. J. Pharm. Sci.* 22, 145–152.

Sheridan, R.P., Feuston, B.P., Maiorov, V.N., Kearsley, S.K., 2004. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Model.* 44, 1912–1928.

Sigma-Aldrich Co. *http://www.sigmaaldrich.com*

Smola, A., Schölkopf, B., 2004. A tutorial on support vector regression. *Stat. Comput.* 14, 199–222.

Song, W., Guo, J., Ding, F., Hu, W., Li, Z., Gao, M., 2011. Study on acute toxicity and structure–activity relationship of *Daphnia magna* exposed to naphthoquinones. *Environ. Toxicol. Pharmacol.* 32, 102–106.

Staples, C.A., Davis, J.W., 2002. An examination of the physical properties, fate, ecotoxicity and potential environmental risks for a series of propylene glycol ethers. *Chemosphere* 49, 61–73.

Stoyanova-Slavova, I.B., Slavov, S.H., Pearce, B., Buzatu, D.A., Beger, R.D., Wilkes, J.G., 2014. Partial least square and k-nearest neighbor algorithms for improved 3D quantitative spectral data–activity relationship consensus modeling of acute toxicity. *Environ. Toxicol. Chem.* 33, 1271–1282.

Tao, S., Xi, X., Xu, F., Li, B., Cao, J., Dawson, R., 2002. A fragment constant QSAR model for evaluating the EC_{50} values of organic chemicals to *Daphnia Magna*. *Environ. Pollut.* 116, 57–64.

The OECD QSAR Toolbox for Grouping Chemicals into Categories, 2010. Organisation for Economic Co-operation and Development.

The OECD QSAR Toolbox for Grouping Chemicals into Categories, 2013. Organisation for Economic Co-operation and Development.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* 58, 267–288.

Todeschini, R., 1997. Data correlation, number of significant principal components and shape of molecules. The *K* correlation index. *Anal. Chim. Acta* 348, 419–430.

Todeschini, R., Ballabio, D., Consonni, V., 2015. Distances and other dissimilarity measures in chemometrics, in: *Encyclopedia of Analytical Chemistry*. John Wiley & Sons.

Todeschini, R., Ballabio, D., Consonni, V., Manganaro, A., Mauri, A., 2009. Canonical measure of correlation (CMC) and canonical measure of distance (CMD) between sets of data. Part 1. Theory and simple chemometric applications. *Anal. Chim. Acta* 648, 45–51.

Todeschini, R., Consonni, V., 2000. Handbook of molecular descriptors. Wiley-VCH Verlag GmbH, Weinheim, Germany.

Todeschini, R., Consonni, V., 2009. Molecular descriptors for chemoinformatics, 2nd ed. Wiley-VCH.

Todeschini, R., Consonni, V., Maiocchi, A., 1999. The *K* correlation index: theory development and its application in chemometrics. *Chemom. Intell. Lab. Syst.* 46, 13–29.

Todeschini, R., Consonni, V., Mauri, A., Pavan, M., 2004. Detecting "bad" regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* 515, 199–208.

Todeschini, R., Vighi, M., Provenzani, R., Finizio, A., Gramatica, P., 1996. Modeling and prediction by using whim descriptors in QSAR studies: toxicity of heterogeneous chemicals on *Daphnia magna*. *Chemosphere* 32, 1527–1545.

Toropov, A.A., Benfenati, E., 2006. QSAR models for Daphnia toxicity of pesticides based on combinations of topological parameters of molecular structures. *Bioorg. Med. Chem.* 14, 2779–2788.

Toropova, A.P., Toropov, A.A., Benfenati, E., Gini, G., 2012a. QSAR models for toxicity of organic substances to *Daphnia magna* built up by using the CORAL freeware. *Chem. Biol. Drug Des.* 79, 332–338.

Toropova, A.P., Toropov, A.A., Martyanov, S.E., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J., 2012b. CORAL: QSAR modeling of toxicity of organic chemicals towards *Daphnia magna*. *Chemom. Intell. Lab. Syst.* 110, 177–181.

Toropova, A.P., Toropov, A.A., Lombardo, A., Roncaglioni, A., Benfenati, E., Gini, G., 2012c. CORAL: QSAR models for acute toxicity in fathead minnow (*Pimephales promelas*). *J. Comput. Chem.* 1218–1223.

Tosato, M.L., Pino, A., Passerini, L., Marchini, S., Viganò, L., Hoglund, M.D., 1993. Updating and validation of a daphnia toxicity model for benzene derivatives. *Sci. Total Environ.* 134, 1479–1490.

Urrestarazu Ramos, E., Vaes, W.H.J., Verhaar, H.J.M., Hermens, J.L.M., 1998. Quantitative structure−activity relationships for the aquatic toxicity of polar and nonpolar narcotic pollutants. *J. Chem. Inf. Comput. Sci.* 38, 845–852.

US EPA. U.S. Environmental Protection Agency, 1996a. OPPTS 850.1010. Aquatic Invertebrate Acute Toxicity Test, Freshwater Daphnids (No. EPA 712–C–96–114), Ecological Effects Test Guidelines.

US EPA. U.S. Environmental Protection Agency, 1996b. OPPTS 850.1075. Fish Acute Toxicity Test, Freshwater and Marine (No. EPA 712–C–96–118), Ecological Effects Test Guidelines.

US EPA. U.S. Environmental Protection Agency. ECOTOX Database, Release 4.0.

Vaes, W.H., Ramos, E.U., Verhaar, H.J., Hermens, J.L., 1998. Acute toxicity of nonpolar versus polar narcosis: is there a difference? *Environ. Toxicol. Chem.* 17, 1380–1384.

Van Leeuwen, C.J., Van Der Zandt, P.T.J., Aldenberg, T., Verhaar, H.J.M., Hermens, J.L.M., 1992. Application of QSARs, extrapolation and equilibrium partitioning in aquatic effects assessment. I. Narcotic industrial pollutants. *Environ. Toxicol. Chem.* 11, 267–282.

Vapnik, V.N., Chervonenkis, A., 1964. A note on one class perceptrons. *Autom. Remote Control* 25.

Vapnik, V.N., Lerner, A.Y., 1963. Pattern recognition using generalized portrait method. *Autom. Remote Control* 24, 774–780.

VEGA Non-Interactive Client. Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy.

Veith, G.D., Broderius, S.J., 1987. Structure-toxicity relationships for industrial chemicals causing type (II) narcosis syndrome, in: *QSAR in Environmental Toxicology - II*. Springer, The Netherlands, 385–391.

Veith, G.D., Broderius, S.J., 1990. Rules for distinguishing toxicants that cause type I and type II narcosis syndromes. *Environ. Health Perspect.* 87, 207–211.

Veith, G.D., Call, D.J., Brooke, L.T., 1983. Structure-toxicity relationships for the fathead minnow, *Pimephales promelas*: narcotic industrial chemicals. *Can. J. Fish. Aquat. Sci.* 40, 743–748.

Verhaar, H.J.M., van Leeuwen, C.J., Hermens, J.L.M., 1992. Classifying environmental pollutants. *Chemosphere* 25, 471–491.

Vighi, M., Calamari, D., 1985. QSARs for organotin compounds on *Daphnia magna*. *Chemosphere* 14, 1925 – 1932.

Vighi, M., Garlanda, M.M., Calamari, D., 1991. QSARs for toxicity of organophosphorous pesticides to *Daphnia* and honeybees. *Sci. Total Environ.* 109–110, 605–622.

Viswanadhan, V.N., Reddy, M.R., Bacquet, R.J., Erion, M.D., 1993. Assessment of methods used for predicting lipophilicity: application to nucleosides and nucleoside bases. *J. Comput. Chem.* 14, 1019–1026.

Von der Ohe, P.C., Kühne, R., Ebert, R.-U., Altenburger, R., Liess, M., Schüürmann, G., 2005. Structural alerts - a new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. *Chem. Res. Toxicol.* 18, 536–555.

Wang, Y., Zheng, M., Xiao, J., Lu, Y., Wang, F., Lu, J., Luo, X., Zhu, W., Jiang, H., Chen, K., 2010. Using support vector regression coupled with the genetic algorithm

for predicting acute toxicity to the fathead minnow. *SAR QSAR Environ. Res.* 21, 559–570.

WATERNT, 2010. U.S. E.P.A.

Wedekind, C., von Siebenthal, B., Gingold, R., 2007. The weaker points of fish acute toxicity tests and how tests on embryos can solve some issues. *Environ. Pollut.* 148, 385–389.

Wei, D.B., Zhang, A.Q., Han, S.K., Wang, L.S., 2001. Joint QSAR analysis using the Free-Wilson approach and quantum chemical parameters. *SAR QSAR Environ. Res.* 12, 471–479.

Williams, E.S., Berninger, J.P., Brooks, B.W., 2011. Application of chemical toxicity distributions to ecotoxicology data requirements under REACH. *Environ. Toxicol. Chem.* 30, 1943–1954.

Wong, D.C., Dorn, P.B., Chai, E.Y., 1997. Acute toxicity and structure-activity relationships of nine alcohol ethoxylate surfactants to fathead minnow and *Daphnia magna. Environ. Toxicol. Chem.* 16, 1970–1976.

Yang, S., Ye, R., Han, B., Wei, C., Yang, X., 2014. Ecotoxicological effect of nano-silicon dioxide particles on *Daphnia Magna. Integr. Ferroelectr.* 154, 64–72.

Yap, C.W., 2010. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* 32, 1466–1474.

Yi, Z., Zhang, A., 2012. A QSAR study of environmental estrogens based on a novel variable selection method. *Molecules* 17, 6126–6145.

Yuan, H., Wang, Y.-Y., Cheng, Y.-Y., 2007. Mode of action-based local QSAR modeling for the prediction of acute toxicity in the fathead minnow. *J. Mol. Graph. Model.* 26, 327–335.

Zhao, Y.H., Cronin, M.T.D., Dearden, J.C., 1998. Quantitative structure-activity relationships of chemicals acting by non-polar narcosis—theoretical considerations. *Quant. Struct.-Act. Relat.* 17, 131–138.

Zhu, C.M., Kong, L.R., Hong, H., Huang, Q.G., Wang, L.S., 1999. Acute toxicity of substituted biphenyls to *Daphnia magna* and quantitative structure-activity relationship study. *Toxicol. Environ. Chem.* 68, 267–273.

Zou, E., Fingerman, M., 1997. Effects of estrogenic xenobiotics on molting of the water flea,*Daphnia magna*. *Ecotoxicol. Environ. Saf.* 38, 281 – 285.

Zvinavashe, E., Du, T., Griff, T., Berg, H.H.J. van den, Soffers, A.E.M.F., Vervoort, J., Murk, A.J., Rietjens, I.M.C.M., 2009. Quantitative structure-activity relationship modeling of the toxicity of organothiophosphate pesticides to *Daphnia magna* and *Cyprinus carpio*. *Chemosphere* 75, 1531–1538.

List of Tables

List of Figures

List of publications

Scientific articles

1. Cassotti, M., Ballabio, D., Consonni, V., Mauri, A., Tetko, I.V., Todeschini, R., 2014. Prediction of acute aquatic toxicity towards *Daphnia magna* by using the GA-*k*NN method. *ATLA-Altern. Lab. Anim.* 42, 31–41.

2. Cassotti, M., Consonni, V., Mauri, A., Ballabio, D., 2014. Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards *Daphnia magna*. *SAR QSAR Environ. Res.* 25 (12), 1013-1036.

3. Cassotti, M., Ballabio, D., Todeschini, R., Consonni, V., 2015. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*). *SAR QSAR Environ. Res.*, 26 (3), 217-243.

4. Cassotti, M., Grisoni, F., Todeschini, R., 2014. Reshaped Sequential Replacement algorithm: an efficient approach to variable selection. *Chemometr. Intell. Lab. Syst.* 133, 136-148.

5. Grisoni, F., Cassotti, M., Todeschini, R., 2014. Reshaped Sequential Replacement for variable selection in QSPR: comparison with other reference methods. *J. Chemometrics* 28, 249-259.

6. Todeschini, R., Consonni, V., Ballabio, D., Mauri, A., Cassotti, M., Lee, S., West, A., Cartlidge, D., 2014. QSPR study of rheological and mechanical properties of Chloroprene rubber accelerators. *Rubber Chem. Technol.* 87 (2), 219-238.

Posters

1. Cassotti, M., Ballabio, D., Consonni, V., Mauri, A., Tetko, I.V., Todeschini, R., 2013. Modelling of acute aquatic toxicity towards *Daphnia magna* using GA-*k*NN method. *Final Conference of the ECO Project*, Prien am Chiemsee (Germany).

2. Ringsted, T., Luecke, S., Poellinger, L., Cassotti, M., Andersson, P. L., 2013. QSAR models on the effect of dioxins and dioxin-like chemicals to human keratinocytes. *Final Conference of the ECO Project*, Prien am Chiemsee (Germany).

3. Grisoni, F., Cassotti, M., Todeschini, R., 2013. Reshaped Sequential Replacement algorithm (RSR) for variable selection. *VIII Colloquium Chemiometricum Mediterraneum*, Bevagna (Italy).

4. Consonni, V., Ballabio, D., Sahigara, F., Mauri, A., Cassotti, M., Grisoni, F., Todeschini, R., 2013. A comparative study on different methods for applicability domain assessment. *VIII Colloquium Chemiometricum Mediterraneum*, Bevagna (Italy).

5. Grisoni, F., Cassotti, M., Todeschini, R., 2012. Comparison of variable selection methods. *3rd Summer School on Chemoinformatics*, Strasbourg (France).

6. Ringsted, T., Giagloglou, E., Ballabio, D., Mauri, A., Cassotti, M., Consonni, V., Todeschini, R., 2012. Read-across methodology in aquatic ecotoxicology and ready biodegradation. *2nd Winter School of the ECO Project*, Madrid (Spain).

Oral communications

1. A QSAR model for acute aquatic toxicity towards *Daphnia magna*. *16th International Workshop on Quantitative Structure-Activity Relationships in Environmental and Health Sciences (QSAR2014)* (2014), Milan (Italy). Awarded as best presentation by a student.

2. Modelling of acute aquatic toxicity on *Daphnia magna* and fish. *2nd Summer School of the ECO Project* (2012), Verona (Italy).

Deliverables

Deliverables produced during this Ph.D. project:

1. Acute aquatic toxicity to *Daphnia Magna* dataset. Available at:

    http://michem.disat.unimib.it/chm/download/toxicity.htm

2. Acute aquatic toxicity to *Pimephales promelas* dataset. Available at:

    http://michem.disat.unimib.it/chm/download/toxicityfish.htm

3. Reshaped Sequential Replacement toolbox for MATLAB. Available at:

    http://michem.disat.unimib.it/chm/download/rsrinfo.htm

Appendix I

Cassotti, M., Ballabio, D., Consonni, V., Mauri, A., Tetko, I.V., Todeschini, R., 2014. Prediction of acute aquatic toxicity towards *Daphnia magna* by using the GA-*k*NN method. *ATLA-Altern. Lab. Anim.* 42, 31–41.

Available at:
http://www.atla.org.uk/prediction-of-acute-aquatic-toxicity-toward-daphnia-magna-by-using-the-ga-knn-method/

Prediction of Acute Aquatic Toxicity Toward *Daphnia magna* by using the GA-*k*NN Method

Matteo Cassotti,[1] Davide Ballabio,[1] Viviana Consonni,[1] Andrea Mauri,[1] Igor V. Tetko[2,3,4] and Roberto Todeschini[1]

[1]University of Milano-Bicocca, Department of Earth and Environmental Sciences, Milano, Italy; [2]Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Munich, Germany; [3]Chemistry Department, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia; [4]eADMET GmbH, Garching, Germany

Summary — In this study, a QSAR model was developed from a data set consisting of 546 organic molecules, to predict acute aquatic toxicity toward *Daphnia magna*. A modified *k*-Nearest Neighbour (*k*NN) strategy was used as the regression method, which provided prediction only for those molecules with an average distance from the *k* nearest neighbours lower than a selected threshold. The final model showed good performance (R^2 and Q^2_{cv} equal to 0.78, Q^2_{ext} equal to 0.72). It comprised eight molecular descriptors that encoded information about lipophilicity, the formation of H-bonds, polar surface area, polarisability, nucleophilicity and electrophilicity.

Key words: *aquatic toxicity,* Daphnia magna*, genetic algorithms,* kNN, QSAR.

Address for correspondence: *Davide Ballabio, University of Milano-Bicocca, Department of Earth and Environmental Sciences, Piazza della Scienza 1, Milano 20126, Italy.*
E-mail: *davide.ballabio@unimib.it*

Introduction

Many chemicals partition in water and can exert adverse effects on aquatic systems, damaging aquatic species and food webs, and threatening the survival of other members of these ecosystems, such as birds and mammals (1). The adverse effects of toxicants can be induced by means of both non-specific and specific mechanisms of action. Non-specific interactions, e.g. narcosis and general reactivity, derive from high concentrations of the toxicants within the cell or cellular membrane, and thus are strongly related to the ability of chemicals to enter the organism.

Some chemicals are able to directly interact with biological targets within the aquatic organism, causing higher toxicity (compared to the baseline set by narcosis). These interactions, or reactions, usually take place between the toxicant (or its metabolites) and critical cellular macromolecules. The assessment of the aquatic toxicity of chemicals is a primary aspect to be addressed. Toxicity tests are typically divided into acute and chronic tests (2), according to the duration of the exposure. Information about the acute aquatic toxicity of chemicals is required for all substances subject to the European Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) regulation (3). In particular, Annex VII of REACH suggests that *Daphnia* is used as the preferred organism for short-term aquatic toxicity testing.

REACH promotes the use of alternative test methods, such as *in vitro* and computer-based methods, including Quantitative Structure–Activity Relationships (QSAR; 3), which are mathematical models that relate the structure of chemical compounds to their activities/properties by using molecular descriptors (4). The Organisation for Economic Co-operation and Development (OECD; 5) set five principles that should be fulfilled by a QSAR model, in order for it to be applicable for regulatory purposes.

Several QSAR models which address acute toxicity toward *D. magna* were calibrated both on heterogeneous and homogeneous data sets, the latter comprising only one specific class of chemical. A list of published QSAR models is reported in Table 1. In general, QSAR models developed on homogeneous (6–14) data sets had higher performances than models calibrated on heterogeneous data. When dealing with QSAR models calibrated on large heterogeneous data sets (8, 15–19), model statistics are lower than those of models calibrated on homogeneous data sets. This is probably due to non-linearity introduced by different mechanisms of action.

To the best of our knowledge, three published QSAR models demonstrated good performance on large heterogeneous data sets. Kaiser (17) developed four QSAR models by using Probabilistic Neural Networks (PNN) coupled with linear corrections with 57 molecular descriptors, calibrated on 700

Table 1: Published QSAR models for acute toxicity toward *D. magna*

Reference	Chemical class	No. of models	*n* training	*n* test	R^2	p	Q^2_{cv}	Q^2_{ext}
Homogeneous data sets								
Vighi (6)	Organophosphorus	1	22	—	0.89	6	—	—
Vighi (7)	Organotin	14	< 15	—	[0.44–0.99]	[1–3]	—	—
Todeschini (8)	Amines	1	8	—	1.00	4	1.00	—
Todeschini (8)	Chlorobenzenes	1	6	—	1.00	3	1.00	—
Todeschini (8)	Organotin	1	15	—	0.99	6	0.97	—
Todeschini (8)	Organophosphorus	1	20	—	0.92	5	0.85	—
Deneer (9)	Nitroaromatics	3	[15–22]	—	[0.60–0.75]	[1–2]	—	—
Hossain (10)	Ionic liquids	1	64	—	0.97	25	—	—
Zvinavashe (11)	Organothiophosphate	3	10	5	[0.80–0.82]	[1–2]	[0.62–0.73]	[0.61–0.71]
Cassani (12)	Triazoles and benzotriazoles	2	97	NR	[0.73–0.77]	[5–5]	[0.70–0.74]	[0.68–0.83]
Cassani (13)	Triazoles and benzotriazoles	7	90	—	[0.59–0.82]	[5–245]	[0.70–0.75]	—
Tetko (14)	Polybrominated diphenyl ethers	1	46	—	0.70	150	0.70	—
Heterogeneous data sets								
Todeschini (8)	—	5	49	—	[0.68–0.82]	[3–7]	[0.64–0.74]	—
Faucon (15)	—	1	61	35	0.54	2	0.49	0.57
Katritzki (16)	—	1	86	44	0.70	5	0.64	0.74
Katritzki (16)	—	2	87	43	[0.72–0.78]	5	[0.67–0.75]	[0.66–0.54]
Katritzki (16)	—	1	130	—	0.71	5	0.68	—
Kaiser (17)	—	4	700	76	[0.87–0.88]	57	—	[0.76–0.76]
Kar (18)	—	1	222	75	0.69	7	0.68	0.74
Kühne (19)	—	1	1365	—	0.85	NR	0.84	—

p = number of molecular descriptors in the model. NR = not reported; Q^2_{cv} = coefficient of determination in cross-validation; Q^2_{ext} = coefficient of determination in external validation; R^2 = coefficient of determination in fitting.
Bibliographic reference, chemical class (where relevant), number of developed models, number of molecules in training and external test sets are reported. In the case of multiple models, the range of the statistics is reported in square brackets.

training compounds and externally validated with 76 molecules (Q^2_{ext} equal to 0.76). Kar (18) collected experimental data on 297 chemicals, and the best QSAR model was obtained by using Partial Least Squares (PLS) regression with seven molecular descriptors (Q^2_{ext} equal to 0.74). Kühne (19) developed a decision-tree model based on linear regression for the prediction of narcosis-level toxicity; read-across was then used to estimate the toxicity enhancement. Models were calibrated on 1,365 organic compounds and the final decision-tree provided a quantitative estimation for 757 compounds (56% of the data set) with a Q^2_{LOO} equal to 0.84.

Published models of acute toxicity toward *D. magna* have some drawbacks that can limit their actual application for regulatory purposes. One drawback, for instance, for PNNs and decision-tree models, is the complex modelling strategy. This can result in a difficult implementation, while

OECD Principle 2 requires the "use of an unambiguous algorithm" in order to give transparency in the equations. Moreover, OECD Principle 5 requires a mechanistic interpretation, if possible: the model based on the PNN strategy lacks a direct mechanistic interpretation, due to both the large number of molecular descriptors (57 fragments) and the intrinsic complexity of the modelling algorithm. Also, OECD Principle 4, requires a correct validation procedure of the QSAR models, which, in some cases, is not properly fulfilled, since several published models were validated by optimistic procedures, such as the *ad hoc* selection of test molecules and leave-one-out cross validation (20).

In order to overcome the drawbacks and limitations of existing models, the aim of this study was to develop a QSAR model for the toxicity of organic chemicals toward *D. magna*, characterised by: a) a simple modelling method based on local structural

similarities; b) interpretable descriptors; c) an appropriate validation procedure to estimate the real predictivity and reliability of the model; and d) an implicit definition of the Applicability Domain (AD; 21, 22). In addition, attention was paid to data screening, in order to detect erroneous chemical structures and reduce the influence of anomalous toxicity values.

Materials and Methods

Experimental data

Experimental data on aquatic toxicity were retrieved from three databases (ECOTOX [23], EAT5 [24] and OASIS) and available scientific publications (25–41). The OASIS database was downloaded from the OECD QSAR Toolbox (42). The downloaded databases were imported into the Konstanz information miner (KNIME; 43), and *ad hoc*-designed workflows were used to extract LC50 data, which is the concentration that causes death in 50% of test *D. magna* over a test duration of 48 hours. Data were obtained under different experimental conditions, such as composition and characteristics (e.g. pH and temperature) of test water, test locations (laboratory or field), exposure types (e.g. static, flow-through, renewal). In the EAT5 database, LC50 data were reported as EC50 (effective concentration), with lethality as the observed effect. Records of the ECOTOX database indicating ranges or thresholds of experimental values were removed.

Data curation and filtering

In order to guarantee data consistency, data were checked, and ambiguous molecular structures and anomalous experimental values were disregarded.

Curation of molecular structures

Chemical names and CAS registry numbers (CASRNs) were available for every record in the data set. Web services to the chemical database, ChemSpider (44), and the Chemical Identifier Resolver (CIR; 45) of the CADD Group at NCI/NIH, were set up in the KNIME environment, to check the correctness of the molecular structures and the correspondence of CASRNs and names. CASRNs and chemical names were independently used as queries to retrieve the standard InChI codes and the Simplified Molecular-Input Line-Entry System (SMILES). The retrieved InChI codes were then compared. Out of 2,640 records (corresponding to 693 different CAS numbers), 1,577 (378 CAS numbers) presented mismatches. All the records that

had at least one mismatch were manually checked by using the PubChem (46) and ChemSpider databases and the Sigma-Aldrich website (47). During this phase, some records were deleted for the following reasons: a) a chemical name–CASRN mismatch was not possible to resolve — for example, because the original publication was not found or was not accessible; b) the CASRN was non-existent; c) the molecular structure was not available, as it was a commercially-named chemical; d) information about which isomer was used was missing; and e) the record pertained to a chemical mixture. In total, 2,410 records, corresponding to 628 different CAS numbers were retained and merged with the data taken from scientific publications (195 records for a total of 183 different CAS numbers).

Filtering

The data set contained a certain number of disconnected structures, i.e. salts and mixtures. In particular, 733 records for a total of 118 disconnected structures were present. All the disconnected structures were removed from the data set, since toxic effects could arise from any of the chemical species present, either behaving independently or interacting to give additive, synergistic or antagonistic effects. Moreover, the calculation of molecular descriptors is limited when dealing with disconnected structures.

Inorganic compounds were removed, since the goal was to develop a model for acute toxicity that was limited to organic molecules. A total of 141 records, corresponding to 28 different inorganic compounds, were therefore removed.

Handling stereochemistry

Some stereoisomers were present in the data set. Since the majority of two-dimensional (2-D) molecular descriptors does not discriminate stereoisomers, the information about stereochemistry was removed from the SMILES before the calculation of molecular descriptors.

Curation of experimental values

Lethal concentrations were first converted to molarity and then transformed to a logarithmic scale (–Log mol/L). For several molecules, multiple values of LC50 were available, and in some cases, differences of a few orders of magnitude were observed for the same chemical. In order to avoid an excessive dependence on outlying data, the median value was calculated, as it is a more robust estimator than the mean value. The standard deviation was also calculated and used as an alert for

inconsistent data. The pooled standard deviation over the data set was equal to 0.37. Therefore, if the standard deviation of a molecule was larger than 0.7 log units (approximately twice the standard deviation over the entire data set), the original scientific publications were consulted in order to detect errors in the compilation of the databases. If the original study was not accessible or not found, the corresponding value was removed.

Experimental data for some Polycyclic Aromatic Hydrocarbons (PAHs) from a scientific publication (48) were removed, because toxicity had been photo-induced in the experimental tests.

The final data set included 546 organic molecules and is freely available (49, 50).

Molecular descriptors

The SMILES of the 546 organic molecules were used to calculate molecular descriptors. Three-dimensional (3-D) descriptors were not calculated, since the optimisation of molecular geometry may be a time-consuming step, and could also limit the future application of the model due to inconsistencies with the generation of 3-D structures (51).

One-dimensional and 2-D molecular descriptors implemented in the software DRAGON (52) were calculated. Constant, near-constant and descriptors with at least one missing value were removed, resulting in a total of 2,187 molecular descriptors.

Modelling methods

Due to the nature of the problem, non-linear regression methods were assumed to give better results than the classical linear regression. Methods based on local similarity are expected to be able to deal with non-linear responses, while still retaining a simple algorithm. This is the case for the k-Nearest Neighbour (kNN; 53) strategy, which was used to calibrate the models. The predicted value for a molecule is computed from the values of its k nearest neighbours, typically as a mean or weighted mean. In this study, the similarity between two molecules was calculated as:

$$S_{st} = \frac{1}{1 + d_{st}} = \frac{1}{1 + \sqrt{(x_s - x_t)^T \; S^{-1} \, (x_s - x_t)}} \quad 0 \leq S_{st} \leq 1$$

[Equation 1]

where d_{st} is the Mahalanobis distance between molecules s and t, x_s and x_t are the descriptor vectors for molecules s and t, and S^{-1} is the inverse of the covariance matrix of the training set. The predicted response, y_s, was computed as the weighted mean over the k neighbours, where the weights were calculated as a function of the similarity, as:

$$y_s = \sum_{t=1}^{k} y_t . w_t = \sum_{t=1}^{k} y_t . \frac{S_{st}}{\sum_{t=1}^{k} S_{st}} \qquad \text{[Equation 2]}$$

where y_t and w_t are the response and the weight of the t–th neighbour, respectively, and the sum runs over the k neighbours. The term S_{st} is the similarity between molecules s and t, and the sum runs again over the k nearest neighbours.

A threshold value on the average distance from the k nearest neighbours was also adopted, in order to detect test molecules that are dissimilar from their k nearest neighbours. Hence, only molecules with an average distance from their neighbours lower than a defined threshold were predicted, while those exceeding the threshold were regarded as outliers on the assumption that their predictions may be influenced by dissimilar neighbours and therefore might not be reliable. The training molecules exceeding the threshold did not contribute to the model's statistics, but were not removed from the data set, since they still contributed to define the model's domain and, in principle, can be useful to predict the responses of external compounds.

Genetic Algorithms (GA) were coupled with kNN method to select the relevant molecular descriptors. The GA strategy described by Leardi and González was used in this study (54). For each combination of molecular descriptors (model), values of k (number of nearest neighbours) from 1 to 10 were tested. For each k value, the distance threshold from the k neighbours was automatically chosen during GA runs as the average distance value giving the largest coefficient of determination in cross-validation (Q^2_{cv}), with a constraint on the maximum allowed percentage of unpredicted molecules of 40%. This value was selected as a reasonable value to carry out the selection of molecular descriptors during model optimisation. Eventually, for each combination of molecular descriptors, the pair of k values and similarity threshold giving the largest Q^2_{cv} was chosen as the optimal one.

Model validation

In order to thoroughly validate the developed models, the 546 molecules of the data set were randomly divided into a training set (436 molecules) and an external test set (110 molecules). The training set was used to calibrate models and select the optimal molecular descriptors by means of GA, while the test set was used only to test the predictive power of the calibrated models. During the GA runs, model performance was evaluated by means of internal five-fold cross-validation (55). The predictive ability on the external test set was evaluated by means of the Q^2_{ext} function reported in the literature (56).

Software

KNIME (43) was used to process the databases, in order to extract the relevant data and check the molecular structures. Molecular descriptors were calculated by means of DRAGON 6 (52). Variable selection by means of GA, model fitting and validation were carried out in MATLAB (57), by using toolboxes and functions written by the authors.

Results and Discussion

The GA selection was organised into two subsequent steps, in order to handle the large number of calculated descriptors, i.e. 2,187, and to avoid potential over-fitting. Initially, GAs were run on each descriptor block separately. For each block, molecular descriptors with the largest frequencies of selection were chosen and merged together to form a set of 201 descriptors. GAs were then carried out on this reduced set, to find the most appropriate subset of descriptors.

Only one molecular descriptor, *TPSA(tot)* (topological polar surface area with N, O, S and P contributions; 58), had a selection frequency significantly larger than the others. In order to avoid selection based on small differences in the descriptor frequencies and to obtain a consistent solution, models based on the 15 most frequent descriptors were explored by means of an all-subset strategy, with two constraints: the maximum number of descriptors included in the models was set to 10; and *TPSA(tot)* was always included, since it proved to be relevant for the toxicity modelling. The best models were finally judged on the basis of both their predictive power and their complexity, also taking into consideration descriptor interpretability. This procedure resulted in a *k*NN model (*k* equal to three) constituted by eight molecular descriptors, which are briefly described below:

a) *MLOGP* is the octanol–water partition coefficient (LogP) calculated from the Moriguchi model (59, 60). LogP expresses the lipophilicity of a molecule, this being the driving force of narcosis.

b) *RDCHI* is a topological index (61) that encodes information about molecular size and branching, but does not account for heteroatoms. Since molecular size affects lipophilicity, it is reasonable that this descriptor also accounts, to a certain extent, for lipophilicity.

c) *SAacc* (62) describes the Van der Waals surface area (VSA) of atoms that are acceptors of hydrogen bonds.

d) *TPSA(tot)* (58) represents the topological polar surface area calculated by means of a contribution method that takes into account N, O, P and S. The two descriptors, *SAacc* and

TPSA(tot), taken together, account for the exposed molecular polar surface area that can interact with biological targets, where *SAacc* specifically takes into account the formation of hydrogen bonds, while the main contribution of *TPSA(tot)* is toward the calculation of the responses of P-containing and S-containing molecules (such as pesticides and herbicides).

e) *H-050* (63) represents the number of hydrogen atoms bonded to heteroatoms. Therefore, this descriptor still partly contains information related to the possibility of H-bond formation, but focuses on the number rather than on the surface area.

f) *nN* (4) is the number of nitrogen atoms present in the molecule. It is known that many nitrogen-containing functional groups are nucleophiles, due to the presence of a lone pair on the nitrogen atom (typically amines). Therefore, it is hypothesised that *nN* encodes information on the nucleophilicity, deriving from the presence of nitrogen atoms in the toxicants.

g) *C-040* (63) represents the number of carbon atoms of the type R–C(=X)–X / R–C#X / X=C=X, where X represents any electronegative atom (O, N, S, P, Se, halogens). In other words, *C-040* codifies specific functional groups such as esters, carboxylic acids, thioesters, carbamic acids, nitriles, etc. Since all of these groups are electron-poor on the carbon atoms, *C-040* seems to be able to account for electrophilic features.

h) *GATS1p* (4) encodes information on molecular polarisability, and tends to have low values for molecules with pairs of bonded atoms with comparable polarisabilities such as –OH, –NH and –NO. Moreover, *GATS1p* has smaller values when the polarisabilities of bonded atoms are large. In other words, the more polarisable a bond, the lower the value of *GATS1p*.

To conclude, the interpretation of the molecular descriptors was demonstrated to be in agreement with previous knowledge on the structural and electronic features that determine acute aquatic toxicity. It was confirmed that toxicity increases with lipophilicity, as a consequence of the enhanced ability of toxicants to enter the organism (narcosis). Moreover, a relationship was found between molecular polarisability and toxicity. This relationship was linked to the HSAB (64) and FMO (65) theories and the Klopman–Salem equation (66, 67), on the basis of the consideration that polarisable molecules are 'soft' species, which therefore tend to react with other soft species. In fact, it seems that more-polarisable molecules tend to have higher toxicities, and this might be due to the formation of covalent bonds that involve the HOMO and LUMO of soft acids and bases.

Figure 1 shows the Q^2_{cv} and percentage of unpredicted molecules as a function of the threshold. The percentage of unpredicted molecules decreased linearly with increasing threshold values, as was expected. On the other hand, model performance remained stable (Q^2_{cv} around 0.80) for threshold values in the range 0.8–1.4. A threshold value equal to 1.26 was finally selected as a reasonable trade-off between model predictivity and applicability limitation. Therefore, predictions for molecules with an

Figure 1: Q^2_{cv} and percentage of unpredicted samples as the function of the threshold value on the average Mahalanobis distance from $k = 3$ neighbours
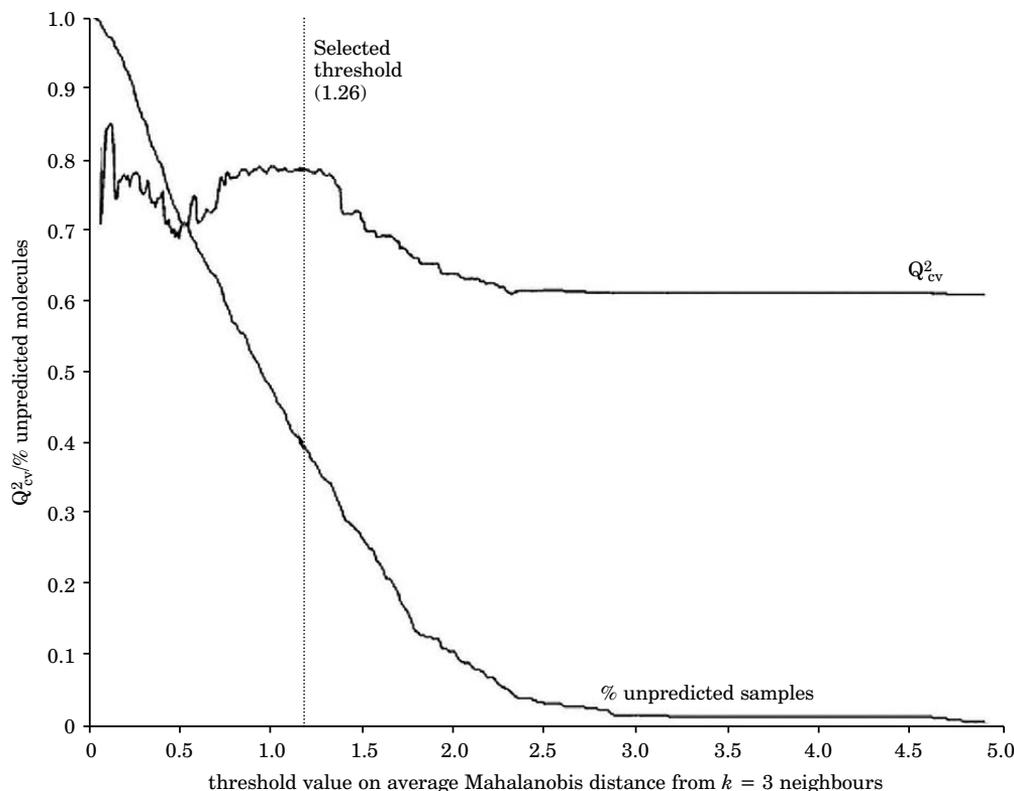
The vertical line corresponds to the selected threshold value (1.26).

Table 2:   Regression statistics of the *k*NN model

Model statistics

k	Av. dist threshold	R^2	Q^2_{cv}	Q^2_{ext}	% unpredicted fit	% unpredicted cv	% unpredicted test
3	—	0.60	0.61	0.43	0	0	0
3	1.26	0.78	0.78	0.72	38	39	31

Q^2_{cv} = *coefficient of determination in cross-validation; Q^2_{ext} = coefficient of determination in external validation; R^2 = coefficient of determination in fitting.*

average distance from their three neighbours greater than 1.26 were regarded as unreliable and were not considered. If no threshold was considered, the classical *k*NN approach would be obtained with Q^2_{cv} equal to 0.61.

The threshold can also be user-defined to best suit the purposes of a specific study. For example, for high-throughput screening, where high reliability is not a strict requirement, one can increase the threshold value in order to have predictions for most of the molecules.

The developed QSAR model was finally validated on the external test set that was not part of the descriptor selection and model calibration. The regression statistics of the *k*NN model are collated in Table 2. The developed model was compared with a 'classical' *k*NN model where no molecule was left unpredicted. It is possible to see that the introduction of the threshold on the average distance enhanced the model's performance, since R^2,

Q^2_{cv} and Q^2_{ext} were improved with 0.18, 0.17 and 0.29 points, respectively, to the detriment of the increase in the number of unpredicted molecules. Moreover, the performance of the model (as well as the percentage of unpredicted molecules) in fitting, cross-validation and external validation, gave similar values. This balance between model performance on the training and test sets indicates the absence of over-fitting, which can occur when dealing with variable selection on high-dimensional data.

Figure 2 shows the experimental *versus* predicted responses in cross-validation for the training set (Figure 2a) and for the external test set (Figure 2b). Black circles indicate compounds with average Mahalanobis distance from the three nearest neighbours which is lower than the selected threshold (i.e. 1.26). White circles indicate molecules with an average distance larger than the threshold. The introduction of the threshold per-

Figure 2: Experimental *versus* predicted responses for the training set and the external test set*
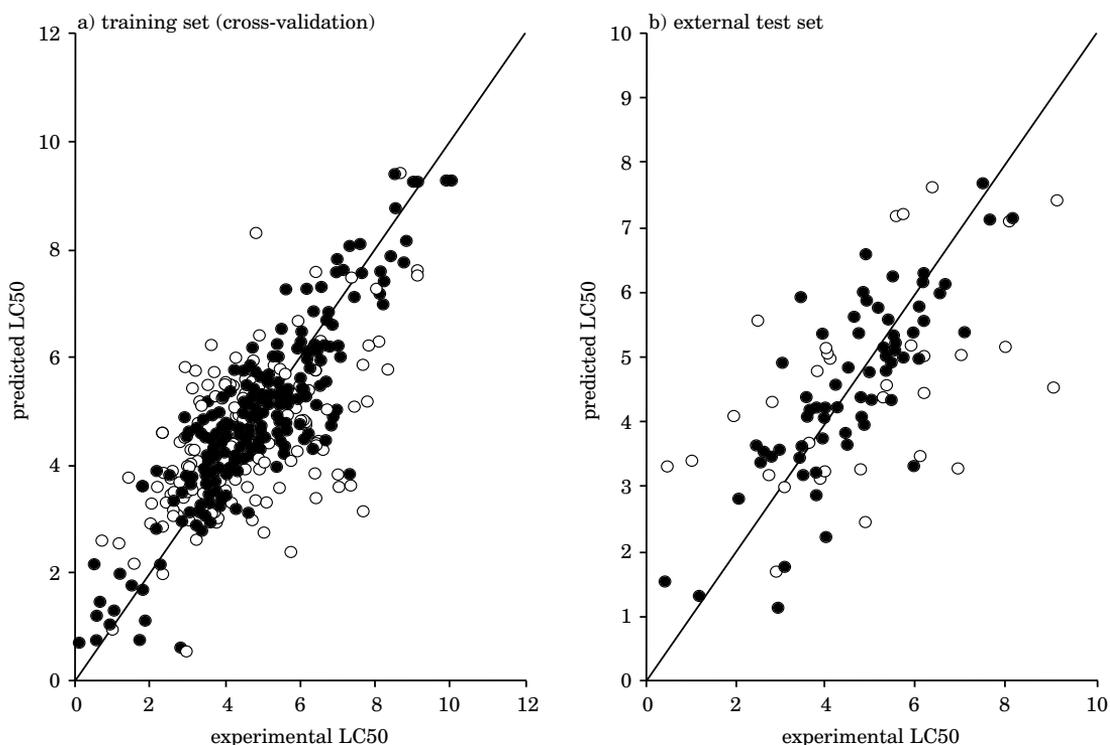
Black circles indicate compounds with average Mahalanobis distance from the three neighbours lower than the fixed threshold (1.26). White circles indicate molecules with average distance higher than the threshold.

*This shows a corrected version of the graph in Figure 2a.

mitted the identification of most of the molecules that were well predicted, while molecules very dissimilar from their neighbours also showed greater residuals in the response, especially in the case of the test set. Nevertheless, there are some molecules, with no structurally similar compounds, that are instead characterised by small residuals. This is likely to be the case of structural cliffs, i.e. molecules with different structures (small similarity) but similar toxicity.

Figure 3 shows, for each training and test molecule, the studentised residuals in cross-validation and external prediction *versus* the average distance from the three nearest neighbours. Six molecules had average Mahalanobis distances larger than 3.5, with a maximum value of 13.5. In order to make the plot more readable, the x-axis was cut

at a value of 3.5. Predictions for molecules placed on the right hand-side of the vertical line (threshold value) were regarded as unreliable. A general trend of increasing residuals as the average distance increases can be observed.

Comparing the statistics of the proposed kNN model (Table 2) with those of other models calibrated on large heterogeneous data sets (Table 1), the proposed kNN model showed comparable performance, but was advantageous in the simplicity of its algorithm (OECD Principle 2), as well as its interpretability (OECD Principle 5). In fact, the proposed kNN model is based on only eight descriptors, while, for instance, the PNNs were based on 57 fragments. An additional important aspect for a QSAR model, especially when applied for regulatory purposes, is the definition of its AD,

Figure 3: Studentised residuals of the training set in cross-validation and the external test set *versus* the average Mahalanobis distance from the three neighbours
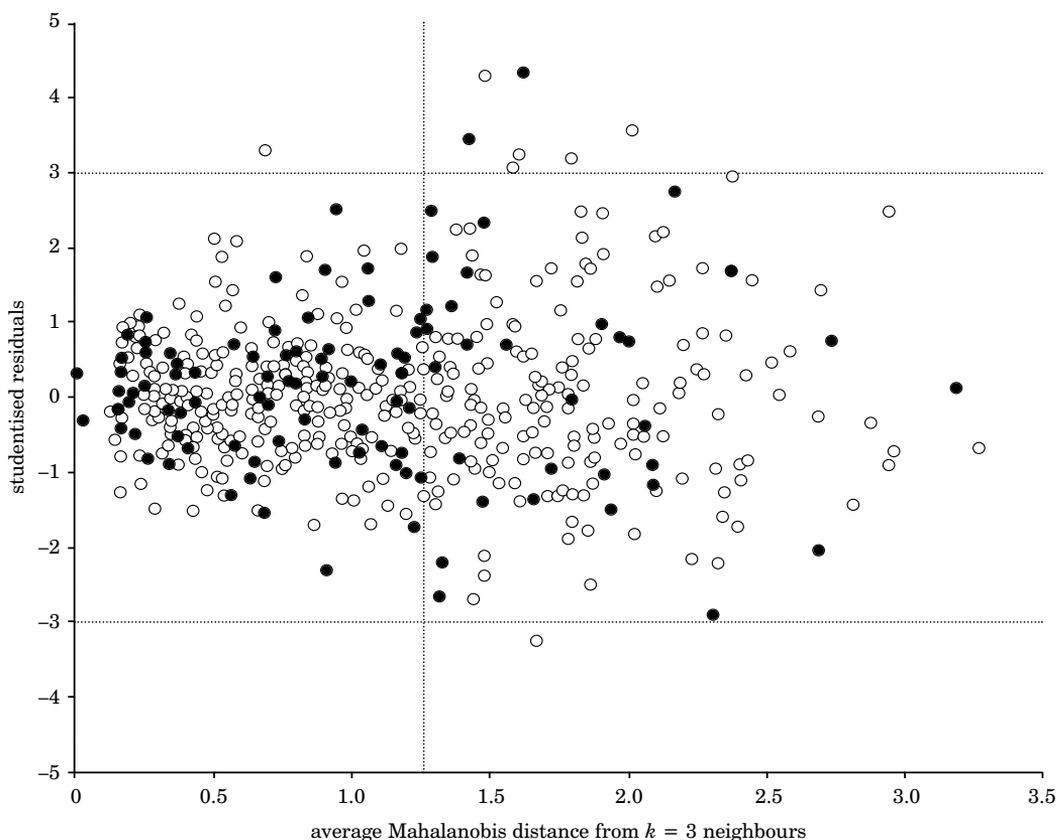
The vertical line represents the threshold value (1.26); the horizontal lines represent warning values on the residuals (3σ). ○ = training set (cross-validation); ● = external test set.

which is the chemical space where it can provide reliable predictions (OECD Principle 3). The introduction of a threshold on the average distance allowed the model to self-determine its own AD, because molecules with distances larger than the threshold are not predicted, on the assumption that their predictions are less (or not) reliable. Additional advantages that the *k*NN model can provide are that it allows a local similarity analysis based on the nearest neighbours for each molecule to be predicted, and it can include new molecules in the training set without the need for recalculation of model parameters, except for the covariance matrix.

Conclusions

This study addressed the problem of predicting the toxicity of organic chemicals toward *D. magna* by means of a QSAR model that was developed to comply with the OECD principles required for the model to be applicable for regulatory purposes.

Data on aquatic toxicity (LC50 on *D. magna* over a test duration of 48 hours) were taken from three databases and 17 additional scientific publications (25–41). *Ad hoc*-designed workflows were used for data curation and filtering. The final data set comprised 546 organic molecules, randomly divided into a training set and an external test set. The GA-*k*NN strategy was implemented with a threshold on the average Mahalanobis distance from the *K* nearest neighbours, so that only molecules satisfying the threshold criterion were predicted. The final QSAR model showed good performance in fitting (R^2 equal to 0.78), cross-validation (Q^2_{cv} equal to 0.78) and external prediction (Q^2_{ext} equal to 0.72), with percentages of unpredicted molecules equal to 38%, 39%, and 31% in fitting, cross-validation and external validation, respectively. An analysis of the residuals on both the training and test sets showed that high residuals were associated with large average distances from the neighbours, thus justifying the introduction of the threshold. The model comprised eight molecular descriptors that encoded information about lipophilicity, formation of H-bonds, polar surface area, polarisability, nucleophilicity and electrophilicity.

Acknowledgment

References

1. Newsome, L.D., Nabholz, J.V. & Kim, A. (1996). Designing aquatically safer chemicals. In *Designing Safer Chemicals: Green Chemistry for Pollution Prevention* (ed. S.C. DeVito & R.L. Garrett), pp. 172–192. Washington, DC, USA: American Chemical Society.

2. Rand, G.M. & Petrocelli S.R. (1985). *Fundamentals of Aquatic Toxicology: Methods and Applications*, 666pp. Washington, DC, USA: Hemisphere Publishing.

3. European Parliament (2006). *Regulation (EC) No 1907/2006* of the European Parliament and of the Council of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending *Directive 1999/45/EC* and repealing *Council Regulation (EEC) No 793/93* and *Commission Regulation (EC) No 1488/94* as well as *Council Directive 76/769/EEC* and *Commission Directives 91/155/EEC, 93/67/EEC, 93/105/EC* and *2000/21/EC. Official Journal of the European Union* **L396**, 30.12.2006, 1–849.

4. Todeschini, R. & Consonni, V. (2009). *Molecular Descriptors for Chemoinformatics*, 1257pp. Weinheim, Germany: Wiley-VCH.

5. OECD (undated). *The Organisation for Economic Co-operation and Development (OECD)*. [Homepage.] Available at: http://www.oecd.org (Accessed 10.01.14).

6. Vighi, M., Masoero Garlanda, M. & Calamari, D. (1991). QSARs for toxicity of organophosphorous pesticides to *Daphnia* and honeybees. *Science of the Total Environment* **109/110**, 605–622.

7. Vighi, M. & Calamari, D. (1985). QSARs for organotin compounds on *Daphnia magna*. *Chemosphere* **14**, 1925–1932.

8. Todeschini, R., Vighi, M., Provenzani, R., Finizio, A. & Gramatica, P. (1996). Modeling and prediction by using WHIM descriptors in QSAR studies: Toxicity of heterogeneous chemicals on *Daphnia magna*. *Chemosphere* **32**, 1527–1545.

9. Deneer, L.W., van Leeuwen, C.J., Seinen, W., Maas-Diepveen, J.L. & Hermens, J.L.M. (1989). QSAR study of the toxicity of nitrobenzene derivatives towards *Daphnia magna*, *Chlorella pyrenoidosa* and *Photobacterium phosphoreum*. *Aquatic Toxicology* **15**, 83–98.

10. Hossain, M.I., Samir, B.B., El-Harbawi, M., Masri, A.N., Mutalib, M.I.A., Hefter, G. & Yin, C.Y. (2011). Development of a novel mathematical model using a group contribution method for prediction of ionic liquid toxicities. *Chemosphere* **85**, 990–994.

11. Zvinavashe, E., Du, T., Griff, T., van den Berg, H.H.J., Soffers, A.E.M.F., Vervoort, J., Murk, A.J. & Rietjens, I.M.C.M. (2009). Quantitative structure–activity relationship modeling of the toxicity of organothiophosphate pesticides to *Daphnia magna* and *Cyprinus carpio*. *Chemosphere* **75**, 1531–1538.

12. Cassani, S., Kovarich, S., Papa, E., Roy, P.P., van der Wal, L. & Gramatica, P. (2013). *Daphnia* and fish toxicity of (benzo)triazoles: Validated QSAR models, and interspecies quantitative activity–activity modelling. *Journal of Hazardous Materials* **258/259**, 50–60.

13. Cassani, S., Kovarich, S., Papa, E., Roy, P.P., Rahmberg, M., Nilsson, S., Sahlin, U., Jeliazkova, N., Kochev, N., Pukalov, O., Tetko, I.V., Brandmaier, S.,

Durjava, M.K., Kolar, B., Peijnenburg, W. & Gramatica, P. (2013). Evaluation of CADASTER QSAR models for aquatic toxicity of (benzo)triazoles and prioritisation by consensus. *ATLA* **41**, 49–64.

14. Tetko, I.V., Sopasakis, P., Kunwar, P., Brandmaier, S., Novotarskyi, S., Charochkina, L., Prokopenko, C. & Peijnenburg, W.J.G.M. (2013). Prioritization of polybrominated diphenyl ethers (PBDEs) using the QSPR-THESAURUS web tool. *ATLA* **41**, 127–135.

15. Faucon, J.C., Bureau, R., Faisant, J., Briens, F. & Rault, S. (2001). Prediction of the *Daphnia* acute toxicity from heterogeneous data. *Chemosphere* **44**, 407–422.

16. Katritzky, A.R., Slavov, S.H., Stoyanova-Slavova, I.S., Kahn, I. & Karelson, M. (2009). Quantitative structure–activity relationship (QSAR) modeling of EC50 of aquatic toxicities for *Daphnia magna*. *Journal of Toxicology & Environmental Health, Part A* **72**, 1181–1190.

17. Kaiser, K.L.E. & Niculescu, S.P. (2001). Modeling acute toxicity of chemicals to *Daphnia magna*: A probabilistic neural network approach. *Environmental Toxicology & Chemistry* **20**, 420–431.

18. Kar, S. & Roy, K. (2010). QSAR modeling of toxicity of diverse organic chemicals to *Daphnia magna* using 2D and 3D descriptors. *Journal of Hazardous Materials* **177**, 344–351.

19. Kühne, R., Ebert, R.U., von der Ohe, P.C., Ulrich, N., Brack, W. & Schüürmann, G. (2013). Read-across prediction of the acute toxicity of organic compounds toward the water flea *Daphnia magna*. *Molecular Informatics* **32**, 108–120.

20. Golbraikh, A. & Tropsha, A. (2002). Beware of Q2! *Journal of Molecular Graphics & Modelling* **20**, 269–276.

21. Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V. & Todeschini, R. (2012). Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **17**, 4791–4810.

22. Tetko, I.V., Bruneau, P., Mewes, H.W., Rohrer, D.C. & Poda, G.I. (2006). Can we estimate the accuracy of ADME-Tox predictions? *Drug Discovery Today* **11**, 700–707.

23. US EPA (2013). *ECOTOX Database, Version 4.0.* Washington, DC, USA: US Environmental Protection Agency. Available at: http:/www.epa.gov/ecotox/ (Accessed 20.12.13)

24. ECETOC (2003). *Technical Report 091: Aquatic Hazard Assessment II*, pp. 1–164. Brussels, Belgium: European Centre for Ecotoxicology and Toxicology of Chemicals. Available at: http://www.ecetoc.org/technical-reports (04.01.14).

25. Bernot, R.J., Brueseke, M.A., Evans-White, M.A. & Lamberti, G.A. (2005). Acute and chronic toxicity of imidazolium-based ionic liquids on *Daphnia magna*. *Environmental Toxicology & Chemistry* **24**, 87–92.

26. Randall, W.F., Dennis, W.H. & Warner, M.C. (1979). Acute toxicity of dechlorinated DDT, chlordane and lindane to bluegill (*Lepomis macrochirus*) and *Daphnia magna*. *Bulletin of Environmental Contamination & Toxicology* **21**, 849–854.

27. Sanderson, H. & Thomsen, M. (2009). Comparative analysis of pharmaceuticals *versus* industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q)SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action. *Toxicology*

28. Jemec, A., Tisler, T., Drobne, D., Sepcic, K., Fournier, D. & Trebse, P. (2007). Comparative toxicity of imidacloprid, of its commercial liquid formulation and of diazinon to a non-target arthropod, the microcrustacean *Daphnia magna*. *Chemosphere* **68**, 1408–1418.

29. Zou, E. & Fingerman, M. (1997). Effects of estrogenic xenobiotics on molting of the water flea, *Daphnia magna*. *Ecotoxicology & Environmental Safety* **38**, 281–285.

30. Costanzo, S.D., Watkinson, A.J., Murby, E.J., Kolpin, D.W. & Sandstrom, M.W. (2007). Is there a risk associated with the insect repellent DEET (*N*,*N*-diethyl-*m*-toluamide) commonly found in aquatic environments? *Science of the Total Environment* **384**, 214–220.

31. Staples, C.A. & Davis, J.W. (2002). An examination of the physical properties, fate, ecotoxicity and potential environmental risks for a series of propylene glycol ethers. *Chemosphere* **49**, 61–73.

32. Martins, J.C., Saker, M.L., Oliva Teles, L.F. & Vasconcelos, V.M. (2007). Oxygen consumption by *Daphnia magna* Straus as a marker of chemical stress in the aquatic environment. *Environmental Toxicology & Chemistry* **26**, 1987–1991.

33. Von der Ohe, P.C., Kuhne, R., Ebert, R., Altenburger, R., Liess, M. & Schuurmann, G. (2005). Structural alerts — A new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay. *Chemical Research in Toxicology* **18**, 536–555.

34. Williams, E.S., Berninger, J.P. & Brooks, B. (2011). Application of chemical toxicity distributions to ecotoxicology data requirements under REACH. *Environmental Toxicology & Chemistry* **30**, 1943–1954.

35. Nørgaard, K.B. & Cedergreen, N. (2010). Pesticide cocktails can interact synergistically on aquatic crustaceans. *Environmental Science & Pollution Research* **17**, 957–967.

36. Dojmi di Delupis, G., Macri, A., Civitareale, C. & Migliore, L. (1992). Antibiotics of zootechnical use: Effects of acute high and low dose contamination on *Daphnia magna* Straus. *Aquatic Toxicology* **22**, 53–60.

37. Ferrari, B., Mons, R., Vollat, B., Fraysse, B., Paxeus, N., Lo Giudice, R., Pollio, A. & Garric, J. (2004). Environmental risk assessment of six human pharmaceuticals: Are the current environmental risk assessment procedures sufficient for the protection of the aquatic environment? *Environmental Toxicology & Chemistry* **23**, 1344–1354.

38. Foit, K., Kaske, O. & Liess, M. (2012). Competition increases toxicant sensitivity and delays the recovery of two interacting populations. *Aquatic Toxicology* **106/107**, 25–31.

39. Ochoa-Acuna, H.G., Bialkowski, W., Yale, G. & Hahn, L. (2009). Toxicity of soybean rust fungicides to freshwater algae and *Daphnia magna*. *Ecotoxicology* **18**, 440–446.

40. Horn, O., Nalli, S., Cooper, D. & Nicell, J. (2004). Plasticizer metabolites in the environment. *Water Research* **38**, 3693–3698.

41. Kyriakopoulou, K., Anastasiadou, P. & Machera, K. (2009). Comparative toxicities of fungicide and herbicide formulations on freshwater and marine species. *Bulletin of Environmental Contamination &*

Toxicology **82**, 290–295.

42. OECD (2012). *The OECD QSAR Toolbox for Grouping Chemicals into Categories, Version 2.3.* Paris, France: Organisation for Economic Co-operation and Development. Available at: http://www.qsartoolbox.org/ (Accessed 04.12.13).

43. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K. & Wiswedel, B. (2007). KNIME: The Konstanz information miner. In *Studies in Classification, Data Analysis and Knowledge Organization*, pp. 319–326. London, UK: Springer. [ISSN: 1431-8814.]

44. RSC (2013). *ChemSpider*. Cambridge, UK: Royal Society of Chemistry. Available at: http://www.chemspider.com/ (Accessed 20.12.13).

45. NCI/CADD Group (2013). *Chemical Identifier Resolver*. Available at: http://cactus.nci.nih.gov/chemical/structure (Accessed 20.12.13).

46. Bolton, E., Wang, Y., Thiessen, P.A. & Bryant, S.H. (2008). Chapter 12 PubChem: Integrated platform of small molecules and biological activities. *Annual Reports in Computational Chemistry* **4**, 217–241.

47. Anon. (2012). *Sigma-Aldrich Co.* [Homepage.] Available at: http://www.sigmaaldrich.com (Accessed 20.12.13).

48. Lampi, M.A., Gurska, J., McDonald, K.I.C., Xie, F., Huang, X.D., Dixon, D.G. & Greenberg, B.M. (2005). Photoinduced toxicity of polycyclic aromatic hydrocarbons to *Daphnia magna*: Ultraviolet-mediated effects and the toxicity of polycyclic aromatic hydrocarbon photoproducts. *Environmental Toxicology & Chemistry* **25**, 1079–1087.

49. Todeschini, R. (undated). *Acute Aquatic Toxicity Dataset*. Milan, Italy: Milano Chemometrics and QSAR Research Group. Available at: http://michem.disat.unimib.it/chm/download/toxicity.htm (Accessed 20.12.13).

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

51. Brandmaier, S., Peijnenburg, W., Durjava, M.K., Kolar, B., Gramatica, P., Papa, E., Bhhatarai, B., Kovarich, S., Cassani, S., Roy, P.P., Rahmberg, M., Öberg, T., Jeliazkova, N., Golsteijn, L., Comber, M., Charochkina, L., Novotarskyi, S., Sushko, I., Abdelaziz, A., D'Onofrio, E., Kunwar, P., Ruggiu, F. & Tetko, I.V. (2014). The QSPR-THESAURUS: The online platform of the CADASTER project. *ATLA* **42**, 13–24.

52. Talete srl (2013). *Talete srl, Dragon (Software for Molecular Descriptor Calculation) Version 6.0 — 2013*. Available at: http://www.talete.mi.it/ (Accessed 12.03.14).

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

54. Leardi, R. & González, A.L. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemometrics & Intelligent Laboratory Systems* **41**, 195–207.

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

56. Consonni, V., Ballabio, D. & Todeschini, R. (2009). Comments on the definition of the Q^2 parameter for QSAR validation. *Journal of Chemical Information & Modeling* **49**, 1669–1678.

57. Anon. (undated). *MATLAB R2012a (64-bit)*. Natick, MA, USA: MathWorks Inc.

58. Ertl, P., Rohde, B. & Selzer P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry* **43**, 3714–3717.

59. Moriguchi, I., Hirono, S., Nakagome, I. & Hirano, H. (1994). Comparison of reliability of log P values for drugs calculated by several methods. *Chemical & Pharmaceutical Bullettin* **42**, 976–978.

60. Moriguchi, I., Hirono, S., Liu, Q., Nakagome, I. & Matsushita, Y. (1992). Simple method of calculating octanol/water partition coefficient. *Chemical & Pharmaceutical Bullettin* **40**, 127–130.

61. Ivanciuc, O., Balaban, T.S. & Balaban, A.T. (1993). Design of topological indices. Part 4. Reciprocal distance matrix, related local vertex invariants and topological indices. *Journal of Mathematical Chemistry* **12**, 309–318.

62. Labute, P. (2000). A widely applicable set of descriptors. *Journal of Molecular Graphics & Modelling* **18**, 464–477.

63. Viswanadhan, V.N., Ghose, A.K., Revankar, G.R. & Robins, R.K. (1989). 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. *Journal of Chemical Information & Modelling* **29**, 163–172.

64. Pearson, R.G. (1963). Hard and soft acids and bases. *Journal of the American Chemical Society* **85**, 3533–3539.

65. Fukui, K., Yonezawa, T. & Shingu, H. (1952). A molecular orbital theory of reactivity in aromatic hydrocarbons. *Journal of Chemical Physics* **20**, 722.

66. Klopman, G. (1968). Chemical reactivity and the concept of charge- and frontier-controlled reactions. *Journal of the American Chemical Society* **90**, 223–234.

67. Salem, L. (1968). Intermolecular orbital theory of the interaction between conjugated systems. I. General theory. *Journal of the American Chemical Society* **90**, 543–552.

Appendix II

Cassotti, M., Consonni, V., Mauri, A., Ballabio, D., 2014. Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards *Daphnia magna*. *SAR QSAR Environ. Res.* 25 (12), 1013-1036.

Available at:

DOI: 10.1080/1062936X.2014.977818

Taylor & Francis
Taylor & Francis Group

Validation and extension of a similarity-based approach for prediction of acute aquatic toxicity towards *Daphnia magna*[£]

M. Cassotti*, V. Consonni, A. Mauri and D. Ballabio

Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan, Italy

(*Received 17 July 2014; in final form 15 September 2014*)

Quantitative structure–activity relationship (QSAR) models for predicting acute toxicity to *Daphnia magna* are often associated with poor performances, urging the need for improvement to meet REACH requirements. The aim of this study was to evaluate the accuracy, stability and reliability of a previously published QSAR model by means of further external validation and to optimize its performance by means of extension to new data as well as a *consensus* approach. The previously published model was validated with a large set of new molecules and then compared with ChemProp model, from which most of the validation data were taken. Results showed better performance of the proposed model in terms of accuracy and percentage of molecules outside the applicability domain. The model was re-calibrated on all the available data to confirm the efficacy of the similarity-based approach. The extended dataset was also used to develop a novel model based on the same similarity approach but using binary fingerprints to describe the chemical structures. The fingerprint-based model gave lower regression statistics, but also less unpredicted compounds. Eventually, *consensus* modelling was successfully used to enhance the accuracy of the predictions and to halve the percentage of molecules outside the applicability domain.

Keywords: QSAR; *Daphnia magna*; toxicity; similarity; REACH; validation

1. Introduction

According to the Registration, Evaluation, Authorization and restriction of CHemicals (REACH) regulation, chemical substances imported or manufactured in quantities higher than one tonne per year need to be registered at the European Chemicals Agency (ECHA) [1]. Registration dossiers should include information regarding the ecotoxicological, toxicological, environmental fate and physical–chemical properties of the addressed compounds. Among ecotoxicological properties, short-term toxicity to *Daphnia* is required for all substances subject to REACH.

Since REACH promotes the adoption of alternative testing approaches such as *in vitro* and *in silico* methods, several (quantitative) structure–activity relationship [(Q)SAR] models were calibrated to address the problem of predicting short-term toxicity to *Daphnia magna* on both small homogeneous and large heterogeneous datasets [2]. Some models were explicitly developed to comply with the five Organisation for Economic Co-operation and Development (OECD) principles for the validation of QSAR models [3]. However, these models can also

present drawbacks leading to an incomplete fulfilment of the OECD principles, which in turn can cause limitations in the real applicability of the models themselves or in the acceptance of their predictions from regulatory bodies. In this regard, particular stress should be given to the validation of QSAR models, as addressed by the fourth OECD principle. Validation is, indeed, a procedure intended to estimate the real predictive power of a model and to quantify this by means of statistical parameters that are assumed also valid for real applications of the model to new compounds. Given that the main goal of a QSAR model is to obtain reliable estimates (predictions) for new compounds, it is extremely important to carefully and thoroughly validate the model.

Moreover, the importance of validation for QSAR models to predict short-term toxicity towards *D. magna* was pointed out in a recent study where several software programs were compared [4]. Considerably lower statistics were obtained when the models were tested on new external data. As a result, the authors stressed the need for high quality data and models able to provide estimates that are more accurate.

Recently, we proposed a QSAR model for the prediction of short-term toxicity (LC_{50} 48 hours) to *D. magna* [2]. The entire development process was carried out with the aim to comply with the five OECD principles. The outcome was a QSAR model based on local similarities that implements an implicit definition of the applicability domain (AD), which is assessed by analysing the distance of each test molecule from its nearest neighbours. The model was originally validated with 110 test molecules randomly chosen from the initial dataset (546 compounds). Validation results on the subset of test molecules inside the AD (69%) showed a satisfactory predictive power (Q^2_{ext} equal to 0.72), which is in agreement with the results in cross-validation (Q^2_{cv} equal to 0.78 for the 61% of training molecules inside the AD).

The aim of this study was (a) to evaluate the accuracy, stability and reliability of the previously published QSAR model by means of further external validation and (b) to optimize its performances by means of the extension to the new data and the implementation of the same similarity-based approach using binary fingerprints. In particular, a thorough validation of the developed QSAR model was carried out by means of a large set of new external data (1009 molecules). The performance of the model was evaluated in comparison with the model implemented in the REACH-oriented ChemProp [5,6] software, from which most of the validation data were taken. The model was finally re-calibrated on the new extended dataset and the same similarity-based approach was implemented on binary fingerprints used in place of the previous molecular descriptors. Finally, *consensus* modelling based on the two available models was used to improve the accuracy of the predictions and to broaden the applicability domain of the two single models.

2. Materials and methods

2.1 *Aquatic toxicity models*

Our previous aquatic toxicity model [2], hereinafter referred to as the MICHEM model, was calibrated on a training set of 436 organic molecules and tested on an external set of 110 compounds, randomly selected. Toxicity data were retrieved from three databases (ECOTOX [7], ECETOC [8] and OASIS [9]) and available scientific publications [10–26]. The model is based on eight molecular descriptors calculated by DRAGON 6 software [27], namely: topological polar surface area with N, O, S, P polar contributions [*TPSA(tot)*]; surface area of acceptor atoms from P_VSA-like descriptors (*SAacc*); Moriguchi octanol–water partition coefficient (*MLOGP*); reciprocal distance sum Randic-like index (*RDCHI*); number of nitrogen

atoms (nN); atom-centred fragments of the type R-C(=X)-X / R-C#X / X=C=X (C-040) with X being either O, N, S, P, Se or halogens; number of hydrogen atoms attached to an hetero-atom (H-050); and Geary autocorrelation of lag one weighted by polarisability ($GATS1p$). These descriptors encode information on lipophilicity ($MLOGP$ and $RDCHI$), formation of H-bonds (H-050 and $SAacc$), polar surface area [$TPSA(tot)$ and $SAacc$], polarisability ($GATS1p$), nucleophilicity (nN) and electrophilicity (C-040).

The MICHEM model implements a modified kNN approach that uses a threshold on the average Mahalanobis distance from the first three nearest neighbours. According to this approach a molecule is predicted only if its average distance from the first three neighbours is lower than a fixed threshold value; otherwise it is unpredicted and considered out of the applicability domain. In this way, the evaluation of the applicability domain is implicitly defined on a similarity-based approach [28–30] and carried out on the fly for each prediction. The introduction of the threshold is meant to identify test molecules that are dissimilar from their nearest neighbours and whose predictions are consequently supposed to be unreliable. If the distance threshold is fulfilled, the prediction is taken as similarity-weighted average of the values of the neighbours. The MICHEM model provided satisfactory performance in fitting, cross-validation and external validation to the detriment of a percentage of molecules that were unpredicted, i.e. considered outside the applicability domain (Table 1, MICHEM).

The model developed by Kühne et al. [6] and implemented in ChemProp [5] is based on a decision tree and is capable of providing either a quantitative or a qualitative estimation of short-term toxicity to *D. magna*. The quantitative approach employs a linear regression to estimate narcosis-level toxicity. Then, atom-centred fragments (ACFs) are used in a read across approach to estimate the toxicity enhancement over the baseline toxicity. The model was calibrated on a training set of 1365 chemicals and provided a quantitative estimation for 757 compounds (55% of the dataset). Details about the performance of ChemProp model are collated in Table 1. In this study, only quantitative predictions of ChemProp were taken into consideration so as to enable a comparison with MICHEM model.

Table 1. Performance of MICHEM, ChemProp, extended MICHEM, fingerprints and *consensus* models in fitting, cross-validation and external validation.

Model	n training	n test	k	Distance threshold	Fitting		Cross-validation		External validation	
					r^2	% out[a]	Q^2	% out[a]	Q^2	% out[a]
MICHEM	436	110	3	1.26	0.78	38	0.78[b]	39	0.72	31
ChemProp	1365	–	≤5	–	0.85	45	0.84[c]	45	–	–
Extended MICHEM	1331	224	5	1.136	0.71	36	0.71[b]	40	0.69	31
Fingerprint-based	1331	224	6	0.664[d]	0.67	29	0.67[b]	33	0.59	24
Consensus 'strict'	1331	224	-	–	0.78	47	0.78[b]	52	0.73	42
Consensus 'loose'	1331	224	-	–	0.70	18	0.70[b]	20	0.67	13

[a]Percentage of unpredicted molecules (outside AD); [b]five-fold cross-validation; [c]leave-one-out cross-validation; [d]complement of Jaccard–Tanimoto similarity (1-S_{jt}).

2.2 *Validation data*

In order to further validate the MICHEM model, an enlarged set of data for LC_{50} 48 hours towards *D. magna* was obtained by merging the data from the ChemProp database [5, 6] and the data provided by the QSAR group at the Technical University of Denmark (DTU) [9,31] (Figure 1). In particular, 1360 compounds were retrieved from the ChemProp software, while DTU provided a set consisting of 388 molecules. DTU provided the SMILES that they used for modelling and additionally the SMILES from the OpenTox database associated with the same CAS-RNs. A comparison between DTU and OpenTox SMILES showed that 12 CAS-RNs were associated with structures with significant differences. These compounds were consequently removed. Therefore, only 376 molecules were retained from the DTU set. These 376 molecules were all external to the MICHEM original dataset, since the MICHEM and DTU sets are not overlapping, as shown in Figure 1.

In order to validate the MICHEM model and to compare it with the ChemProp model, it was first necessary to check the overlaps between the three sets (DTU, MICHEM and Chem-Prop). The three sets of molecules were imported and merged together in KNIME [32] and experimental toxicity values were converted to $-Log((mol/L))$. SMILES and CAS-RNs were then imported in MATLAB [33] to check, for each record, whether there was another record with the same CAS-RN or SMILES. A total of 78 molecules that had duplicates in terms of either CAS-RN or SMILES, thus indicating mismatches between CAS-RN and SMILES in different source sets, were detected and removed. The remaining 2204 records (1547 unique molecules) were used to generate three validation subsets, namely 'External to MICHEM', 'External to ChemProp' and 'External to both MICHEM and ChemProp'. Details of the validation subsets are given in Figure 1. For the 'External to MICHEM' subset, when two experimental LC_{50} values were available (one from ChemProp and one from DTU), the average value was calculated and used for the following validation. For seven molecules of the 'External to MICHEM' subset, it was not possible to calculate one of the descriptors of the MICHEM model, due to the presence of either tin or silicon atoms. These seven molecules were therefore excluded and the validation was carried out using the remaining 1009 compounds.
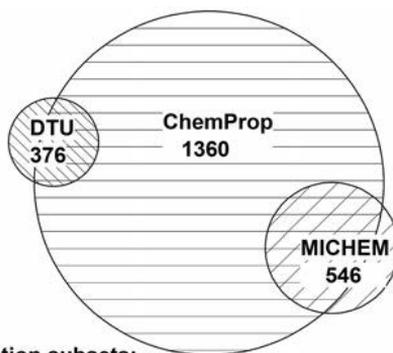
Extracted validation subsets:
1. External to ChemProp: 228 molecules
2. External to MICHEM: 1009 molecules
3. External to both MICHEM and ChemProp: 128 molecules

Figure 1.   Illustration of the retrieved datasets and definition of the three validation subsets.

2.3 *Binary fingerprints*

In the field of similarity analysis, binary fingerprints have recently gained much attention and are currently being employed for similarity searching in databases [34,35]. This is because binary fingerprints are able to provide a holistic view of the molecular structure in terms of all identified fragments. Several different ways to calculate fingerprints were defined [36] and particular interest was paid to hashed fingerprints because these allow a compression of the information in a bit string of defined length. Therefore, it is legitimate to wonder whether fingerprints could provide a broader view of the molecular structure compared with 'classical' descriptors for they consider all the possible fragments in the molecule.

To test this hypothesis, and given that the MICHEM model is based on local similarities, binary fingerprints were used in place of DRAGON descriptors with the same similarity-based approach. In this study, extended connectivity fingerprints [37,38] were generated by means of in-house software and with the following iterative procedure.

(a) Centre on atom 1.
(b) Identify the fragment at *radius* equal to 0, i.e. constituted by the single atom.
(c) Identify the fragment at *radius* equal to one, i.e. consisting in atom 1 and all the atoms bonded to it.
(d) Identify the fragments at *radius* equal to two, i.e. fragments centred on atom 1 and including the bonded atoms and atoms at topological distance equal to two.
(e) Repeat points (a) to (d) centring on each atom of the molecule.
(f) Identify rings.
(g) Generate an empty string of 1024 bits.
(h) Identified fragments 'hit' specific bits (always the same ones). NB One fragment hits only one bit, but more fragments can hit the same bit; thus there is no correspondence between bit and a single fragment.
(i) Repeat for each molecule.

The following features are considered to identify for the fragments: (a) atom type; (b) aromaticity; (c) atomic charge; (d) bond order; (e) connectivity; and (f) attached hydrogen. This means that at least one of these features must be different in order to discriminate between two fragments. It should also be remarked that it is only information about the presence or absence of a particular fragment that is coded by the fingerprint, whereas information about the number of occurrences of a fragment in a molecule is lost. An example of the algorithm is depicted in Figure 2 for 1,2-dichloroethylene. Figure 2(a) shows the identification of the fragments and their codification in SMARTS and Figure 2(b) illustrates the generation of the bit string. Note that in this example, each fragment hits a different bit but that degeneration of more fragments on the same bit is allowed and commonly observed.

2.4 *Consensus modelling*

Consensus modelling consists of combining predictions provided by different models to obtain higher quality predictions and/or broaden the applicability domain (AD) of single models. The assumption is that the strengths of one model will counterbalance the weaknesses of the others and vice versa. Previous studies have already shown the beneficial effects of *consensus* modelling when dealing with ecotoxicological endpoints [39–42].

The generation of a *consensus* analysis can be based on different strategies such as averaging, scoring and probabilities [43–46]. In this study, two approaches ('strict' and 'loose')

Figure 2.   Stepwise example of the generation of binary fingerprints for 1,2-dichloroethylene: (a) identification of the fragments; and (b) generation of the binary fingerprint.

were used to develop *consensus* models. The 'strict' approach provides an averaged prediction only for molecules that are inside the AD of all models. On the other hand, the 'loose' approach also considers molecules that are inside the AD of just one model. If a molecule is inside the AD of only one model, then the only available prediction is considered; if instead the molecule is inside the AD of more models, then an averaged prediction is given in the same fashion as the 'strict' approach.

2.5 *Software*

KNIME [32] was used to merge the three sets of data (MICHEM, ChemProp and DTU), transform the experimental toxicities to –Log(mol/L) and generate the three validation subsets. Molecular descriptors and fingerprints were calculated by means of DRAGON 6 [27] and in-house software, respectively. ChemProp [5] was used to retrieve experimental toxicity data from the dataset of Kühne et al. [6] and to run the corresponding QSAR model. Fitting and validation of MICHEM and fingerprints models were carried out in MATLAB [33] using functions written by the authors. Marvin was used for drawing, displaying and characterizing chemical structures and substructures [47]. The original dataset of the MICHEM model (546 molecules) is freely available on the group's website [48].

3. Results and discussion

3.1 *Model validation and comparison*

The three generated validation subsets were used to test the predictive power of the MICHEM and ChemProp models. In particular, both models were initially tested on the 'External to both MICHEM and ChemProp' subset (128 molecules). The MICHEM and ChemProp models were then tested on the 'External to MICHEM' subset (1009 molecules) and the 'External to ChemProp' subset (228 molecules), respectively. Their performance was evaluated in terms of the coefficient of determination on the external test set (Q^2_{ext}) [49], root mean square error (RMSE) and percentage of unpredicted molecules. The validation results are collated in Table 2.

The 'External to both MICHEM and ChemProp' subset enabled a direct comparison of the performance of the two models: ChemProp had a slightly larger Q^2_{ext}, but also a slightly larger percentage of not predicted molecules. The results on the other two subsets showed that the MICHEM model performed slightly better in terms of both Q^2_{ext} and the percentage of unpredicted molecules. These considerations gained additional value considering the number of molecules in the external validation subsets (1009 and 228 for the MICHEM and ChemProp models, respectively). By comparing these results with those obtained during model calibration (Table 1, MICHEM and ChemProp models), it appeared that the external validation of MICHEM model (Q^2_{ext} equal to 0.72) was a fairly good estimate of the real predictive power of the model, especially considering that the 'External to MICHEM' subset is a large set of data. The performance of the ChemProp model on the two external subsets was instead significantly lower than in cross-validation (Q^2_{cv} equal to 0.84). This indicates that the simple leave-one-out cross-validation could be an optimistic validation procedure in these conditions and external validation gives a more realistic estimate of the predictive power [50].

Table 2. Results of the external validation of the MICHEM and ChemProp models.

Validation subset	No. molecules	*ChemProp*			*MICHEM*		
		% unpredicted	Q^2_{ext}	*RMSE*	*% unpredicted*	Q^2_{ext}	*RMSE*
External to both MICHEM and ChemProp	128	47	0.60	1.073	45	0.56	1.100
External to MICHEM	1009				51	0.66	0.967
External to ChemProp	228	54	0.56	1.134			

Figure 3 collects the experimental versus predicted LC$_{50}$ values for the 'External to both MICHEM and ChemProp' (Figure 3(a) and (b)), 'External to ChemProp' ((Figure 3(c)) and 'External to MICHEM' ((Figure 3(d)) subsets. Figure 3(b) indicates that the MICHEM model had a tendency to overestimate the toxicity (with the exception of one molecule that is largely underestimated), but the results on the larger 'External to MICHEM' subset (Figure 3(d)) do not show the same pattern. The residuals look, in fact, normally distributed. A little bias might be present in the ChemProp model: Figure 3(a) and (c) seem to indicate that the model slightly overestimates toxicity for low LC$_{50}$ values and underestimates it for large LC$_{50}$ values, even though few molecules with medium–high LC$_{50}$ are largely overestimated.
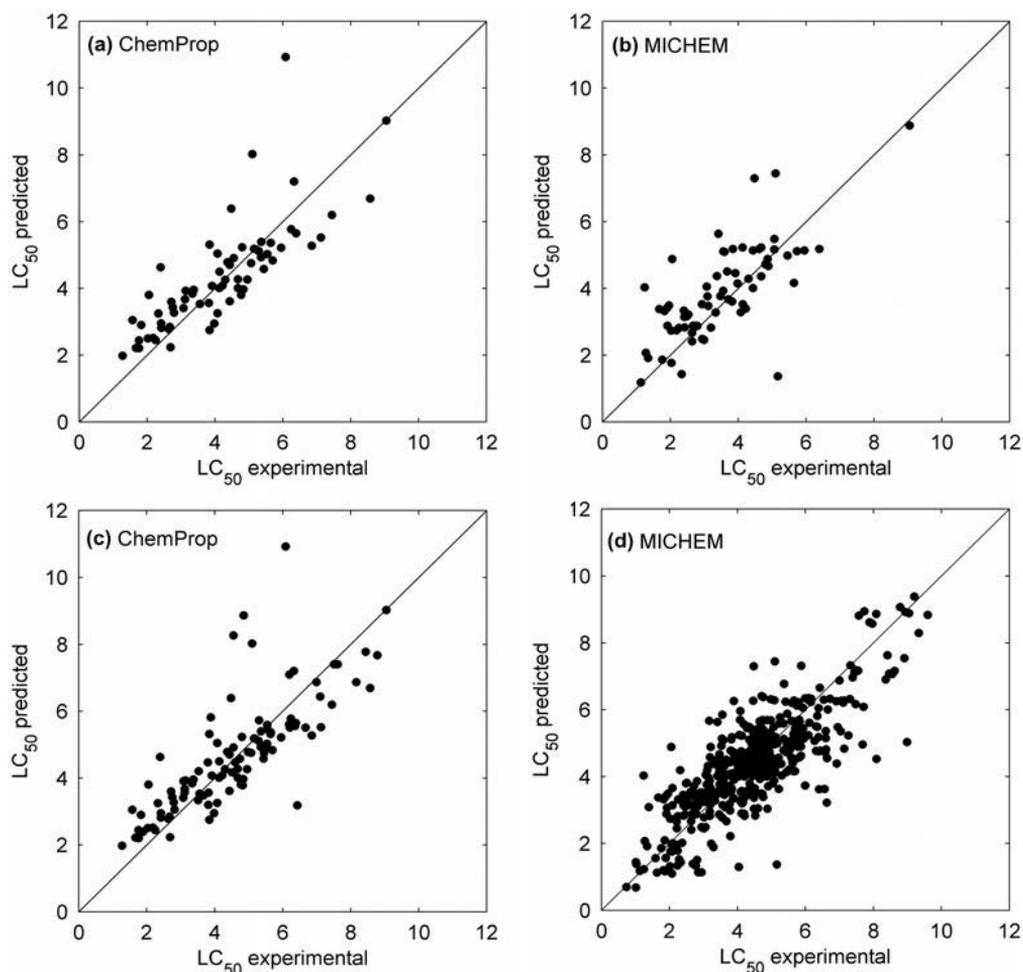
Figure 3.   Experimental versus predicted toxicity values on the three validation subsets: (a) 'External to both MICHEM and ChemProp' by means of ChemProp; (b) 'External to both MICHEM and Chem-Prop' by means of MICHEM; (c) 'External to ChemProp'; and (d) 'External to MICHEM'. LC$_{50}$ values are defined as –log(mol/L).

3.2 *Investigation of the residuals of the MICHEM model*

A central part of the analysis of a QSAR model is the investigation of its residuals and their distribution. This is particularly important for the MICHEM model to verify the assumptions made in the modelling stage regarding the relationship between residuals and average distance from the three nearest neighbours [2]. Figure 4 shows the average Mahalanobis distance of the molecules in the 'External to MICHEM' subset from their three nearest neighbours in the training set versus the standardized residuals, and allows checking of whether the introduced threshold is an effective tool to identify dissimilar molecules whose predictions are likely to be unreliable. Figure 4 confirms the considerations drawn in the modelling phase, that is, there is a trend in increasing the residuals with increasing the average distance from the three nearest neighbours. However, there are clearly four molecules characterized by very large residuals (larger than three), whose average distance is lower than the threshold (highlighted by the black circle in Figure 4). These four molecules were further analysed in order to disclose the reasons for the poor predictions despite their similarities with training molecules.

Table 3 provides detailed information about these four molecules and their corresponding neighbours in terms of experimental and predicted values. It is apparent that propanal (i.e. neighbour one of 2-propen-1-ol) has a very different experimental value with respect to neighbours two (acetone) and three (acetaldehyde). The same applies to the three neighbours of 4-(dimethylamino)-3,5-dimethyl phenylmethyl carbamate (Mexacarbate), namely 3-(3-chloro-4-methoxyphenyl)-1,1-dimethylurea, 2-*sec*-butylphenyl *N*-methylcarbamate and *N*, *N*-dimethylaniline, whose LC_{50} values range from 3.16 to 6.32 in logarithmic units. These
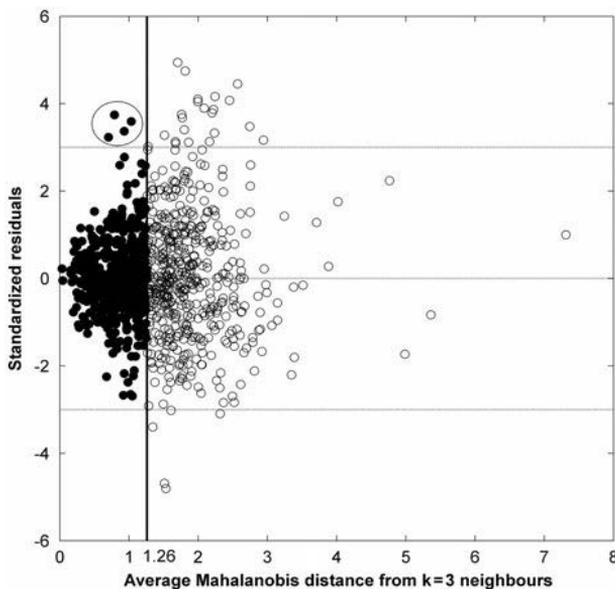
Figure 4. Average Mahalanobis distance from the first three nearest neighbours versus standardized residuals for the 'External to MICHEM' subset. Four molecules with particularly large residuals are highlighted with a black circle. The vertical line represents the threshold value: molecules on the right-hand side (white circles) are considered dissimilar from their nearest neighbours and are therefore not predicted (out of AD).

Table 3. Experimental and predicted LC_{50} values [-Log(mol/L)] for the four poorly predicted test molecules and experimental values of the three neighbours for each test molecule. The SMILES of the analysed chemicals are given in the footnotes.

Test molecule				Neighbour 1			Neighbour 2			Neighbour 3		
Name	LC_{50} exp.[a]	LC_{50} pred.[b]	$LC_{50}ECHA$[c]	Name	LC_{50} exp.[a]	$LC_{50}ECHA$[c]	Name	LC_{50} exp.[a]	$LC_{50}ECHA$[c]	Name	LC_{50} exp.[a]	$LC_{50}ECHA$[c]
2-propen-1-ol[d]	5.16	1.37	4.55	Propanal[e]	2.81	2.97	Acetone[f]	0.67	0.28;0.66	Acetaldehyde[g]	0.55	2.96
4-(Dimethylamino)-3,5-dimethyl phenylmethyl carbamate (Mexacarbate)[h]	8.09	4.53	Aquatic Acute 1	3-(3-Chloro-4-methoxyphenyl)-1,1-dimethylurea[i]	3.16	Aquatic Acute 1	2-sec-Butylphenyl N-methylcarbamate[j]	6.32	Aquatic Acute 1	N,N-Dimethylaniline[k]	4.38	3.86; 4.03;4.59; 4.85
3-Phenoxybenzoic acid[l]	6.63	3.22	Aquatic Acute 1	2-(3-Benzoylphenyl) propanoic acid[m]	3.60	-	4-Chloro-o-tolyloxyacetic acid[n]	3.05	Aquatic Acute 1	4-Hydroxy-3-(3-oxo-1-phenylbutyl) coumarin[o]	2.96	–
Bis(2-ethylhexyl) terephthalate (DEHT)[p]	8.99	5.04	8.45	Bis(2-ethylhexyl) phthalate[q]	4.55	6.58; 6.37	Bis(2-ethylhexyl) adipate[r]	5.75	Aquatic Acute 1	Dibutyl phthalate[s]	4.88	4.97; 4.91

[a]Experimental LC_{50}
[b]predicted LC_{50}
[c]LC_{50} value or classification found in ECHA database
[d]C(=C)CO
[e]CCC=O
[f]CC(=O)C
[g]CC=O
[h]O=C(Oc1cc(c(c(c1)C)N(C)C)C)NC
[i]COc1ccc(cc1Cl)NC(=O)N(C)C
[j]CCC(c1ccccc1OC(=O)NC)C
[k]CN(c1ccccc1)C
[l]OC(=O)c2cccc(Oc1ccccc1)c2
[m]OC(=O)C(c1cccc(c1)C(=O)c1ccccc1)C
[n]OC(=O)COc1ccc(cc1C)Cl
[o]OC(=O)CC(C1=C(O)c2ccccc2OC1=O)c3ccccc3
[p]O=C(OCC(CCCC)CC)c(ccc1)C(=O)OCC(CCCC)CC)c1
[q]CCCCC(COC(=O)c1ccccc1C(=O)OCC(CCCC)CC)CC
[r]CCCCC(COC(=O)CCCCC(=O)OCC(CCCC)CC)CC
[s]CCCCOC(=O)c1ccccc1C(=O)OCCCC.

considerations could indicate that these regions of the chemical space are characterized by activity cliffs, i.e. large variation in the activity in contrast to small structural variation. Keeping in mind the large sensitivity of measured values to different experimental conditions and considering that for all the neighbours only one experimental value was available (with the exception of acetone and dibutyl phthalate), additional experimental values were searched in the ECHA registration database [51]. Furthermore, the calculated baseline toxicity by means of the $LogK_{ow}$ based equation implemented in ChemProp indicated that all of these four molecules exert a large excess toxicity. The results of this evaluation and additional considerations are presented below on a molecule-by-molecule basis and are summarized in Table 3.

3.2.1 *2-Propen-1-ol (CAS-RN 107-18-6)*

The experimental values found in the ECHA database for the test molecule and neighbours one and two are relatively consistent with the values in the validation and training sets, the maximum difference being 0.61 log units. The additional value found for neighbour three is instead almost 2.5 log units larger than the one in the training set, which could indicate a large sensitivity of this compound to different experimental conditions. If the new values were used, the predicted value for 2-propen-1-ol would still be affected by a large standardized residual (2.87). Poor predictions for 2-propen-1-ol (or the related 2-propenal) were obtained also elsewhere, due to its outlying behaviour [52–54].

3.2.2 *4-(Dimethylamino)-3,5-dimethyl phenylmethyl carbamate (Mexacarbate, CAS-RN 315-18-4)*

No experimental value was found in the ECHA database for Mexacarbate and neighbours one and two. However, these three molecules are classified as Aquatic Acute 1 according to the Classification, Labelling and Packaging (CLP) regulation [55]. The classification supports the experimental values for Mexacarbate and 2-*sec*-butylphenyl *N*-methylcarbamate (its second neighbour), but is in contrast with the experimental LC_{50} of neighbour one (3-(3-chloro-4-methoxyphenyl)-1,1-dimethylurea). This is an indirect indication that there must be a large variation in the experimental LC_{50} values for this compound. The additional experimental values found in ECHA database for the third nearest neighbour (*N,N*-dimethylaniline) are consistent with the value used in the training set.

3.2.3 *3-Phenoxybenzoic acid (CAS-RN 3739-38-6)*

No additional experimental value was found in the ECHA database for the test molecule and its neighbours (2-(3-benzoylphenyl) propanoic acid, 4-chloro-*o*-tolyloxyacetic acid and 4-hydroxy-3-(3-oxo-1-phenylbutyl) coumarin). 3-Phenoxybenzoic acid and neighbour two are classified as Aquatic Acute 1: the classification supports the experimental toxicity of the test molecule but is in contrast with the available value for the second neighbour. Despite the classification and the experimental LC_{50} (6.63), Day and Maguire [56] reported 3-phenoxybenzoic acid as not being significantly toxic to *Daphnia magna*. These disagreements for both 3-phenoxybenzoic acid and neighbour two (4-chloro-*o*-tolyloxyacetic acid) could indirectly indicate again a large sensitivity of these compounds to different experimental test conditions.

3.2.4 *Bis(2-ethylhexyl) terephthalate (DEHT, CAS-RN 6422-86-2)*

The additional experimental value found in the ECHA database for DEHT is in good agreement with the one in the validation subset. The situation is different, instead, for the first neighbour [bis(2-ethylhexyl) phthalate] because the two additional values found in the ECHA database are approximately two logarithmic units larger than the training value. As commented earlier for similar cases, such a disagreement in the experimental values could indicate a large sensitivity of the compound to different experimental conditions. The classification of neighbour two [bis(2-ethylhexyl) adipate] as Aquatic Acute 1 is in agreement with the experimental value in the training set. The additional values found for neighbour three (dibutyl phthalate) are in good agreement with the training value. If the new values from ECHA were used for neighbours one and three, the predicted value would still be affected by a large standardized residual (2.93). The LC_{50} values of the neighbours used in the training set should exclude the presence of activity cliffs in this region of the chemical space (because they are relatively close to each other), but the additional values found in the ECHA database show a more incoherent situation.

3.2.5 *Conclusions*

Based on these considerations, it can be stated that the threshold on the average Mahalanobis distance from the three nearest neighbours was in general effective in the identification of dissimilar test molecules, a condition that should lead to poor predictions, and its introduction is therefore justified. Unfortunately, presumed reliable predictions are not always accurate, as for the four analysed cases, because the reliability is evaluated only on the chemical structures but the accuracy depends on the model performance, which is not uniform in the chemical structural domain. The analysis of these four molecules leads to the hypotheses that the poor predictions may be due to outlying behaviours (2-propen-1-ol), potential activity cliffs between the neighbours (2-propen-1-ol and Mexacarbate) or sensitivity to experimental conditions – all aspects that would need expert judgement.

3.3 *Extension of the MICHEM model*

Since the validation results confirmed the efficacy of the similarity-based approach for modelling short-term toxicity towards *D. magna*, the MICHEM model was re-calibrated using also the molecules included in the external validation sets. First, an exploratory analysis based on principal component analysis (PCA) [57] was carried out to check the consistency of the new set of data (1009 molecules) with the original dataset. The score plot (not shown) highlighted that the new set of data and the original dataset covered the same space, thus justifying the combination of the two sets.

Since the original dataset (546 molecules) was divided into training and test sets (436 and 110 compounds, respectively), the new data (1009 molecules) and the original test set (110 molecules) were merged together providing a set of 1119 compounds never used for model calibration. A new extended test set consisting in 224 molecules (20% of the extended external data) was randomly extracted to validate the extended model. The remaining 895 molecules were merged with the original training set (436 compounds) to provide an extended training set of 1331 molecules.

The MICHEM model was re-calibrated and optimized (selection of optimal values of *k* and threshold on the average distance) on the extended training set using five-fold cross-validation.

The optimized values for k and distance threshold were 5 and 1.136, respectively. The extended MICHEM model was then validated on the new extended test set constituted by 224 molecules. The results are summarized in Table 1 (extended MICHEM model).

From Table 1 it can be seen that the performance of the extended MICHEM model is slightly lower than that on the original dataset. However, the parameters in fitting, cross-validation and on the test set are more balanced, indicating a higher stability of the model. Additionally, the results on the extended test set show an improvement of the accuracy of the predictions over the previous validation on the 'External to MICHEM' subset (Q^2_{ext} equal to 0.69 and 0.66, respectively) and also a higher generalization ability of the model as highlighted by the lower percentage of unpredicted molecules (31% versus 51%, respectively). Figure 5 shows that large residuals are still associated with large average distances from the neighbours, thus confirming the validity of the similarity-based approach also for the extended model. Still, there are three compounds in fitting ((Figure 5(c)) and one in the extended test set ((Figure 5(d)) that are largely underestimated (standardized residual greater than three) but have an average distance lower than the threshold. These compounds are 2-propenal, Mexacarbate and bis(2-ethylhexyl) terephthalate (DEHT) in the training set and 2,2-dichlorovinyl dimethyl phosphate (Dichlorvos) in the test set. The three training compounds were already commented in the previous paragraph and Table 3; Dichlorvos is an organophosphate insecticide that acts as acetylcholinesterase inhibitor. Furthermore, Figure 5(a) and (b) seem to indicate a little bias leading to underestimation of large LC_{50} values and vice versa.

3.4 *Fingerprint-based modelling*

The extended binary fingerprints introduced in Section 2.3 were used to calibrate a kNN model based on the same similarity approach as the MICHEM model. This was done to check whether similarity analysis based on fingerprints could also be successfully applied to model acute aquatic toxicity.

The fingerprint-based model was calibrated on the extended training set (1331 molecules) and then validated on the extended test set (224 compounds). Since using the Mahalanobis distance (as for the MICHEM model) on binary data would not have been appropriate, the Jaccard–Tanimoto (JT) similarity coefficient [58] was used to find the nearest neighbours and to obtain the weighted prediction for each test molecule. The optimized values of k (number of nearest neighbours) and threshold on the average similarity from the nearest neighbours were equal to 6 and 0.336, respectively. Regression statistics are reported in Table 1 (fingerprint-based model).

In this case, the similarity threshold seems less effective at identifying molecules associated with large residuals, as shown in Figure 6, where observed versus predicted responses and the average similarity from six neighbours versus the standardized residuals of the fingerprints model are plotted. In fact, there are several molecules associated with large residuals whose average similarity is higher than the threshold. For instance, Figure 6(c) and 6(d) show molecules characterized by average similarity with their nearest neighbours equal to one but relatively large standardized residuals. An average similarity coefficient equal to one means that test molecule and its six nearest neighbours have the exact same binary fingerprint.

Recalling the explanation of the fingerprint generation algorithm of Section 2.3, this situation occurs when the fragments identified in the molecules are the same. The difference in the structure of these compounds can lie in the number of occurrences of the identified fragments because this piece of information is lost. In particular, in the case of the analysed data, the difference lies in the length of carbon chains (number of occurrences of carbon-based
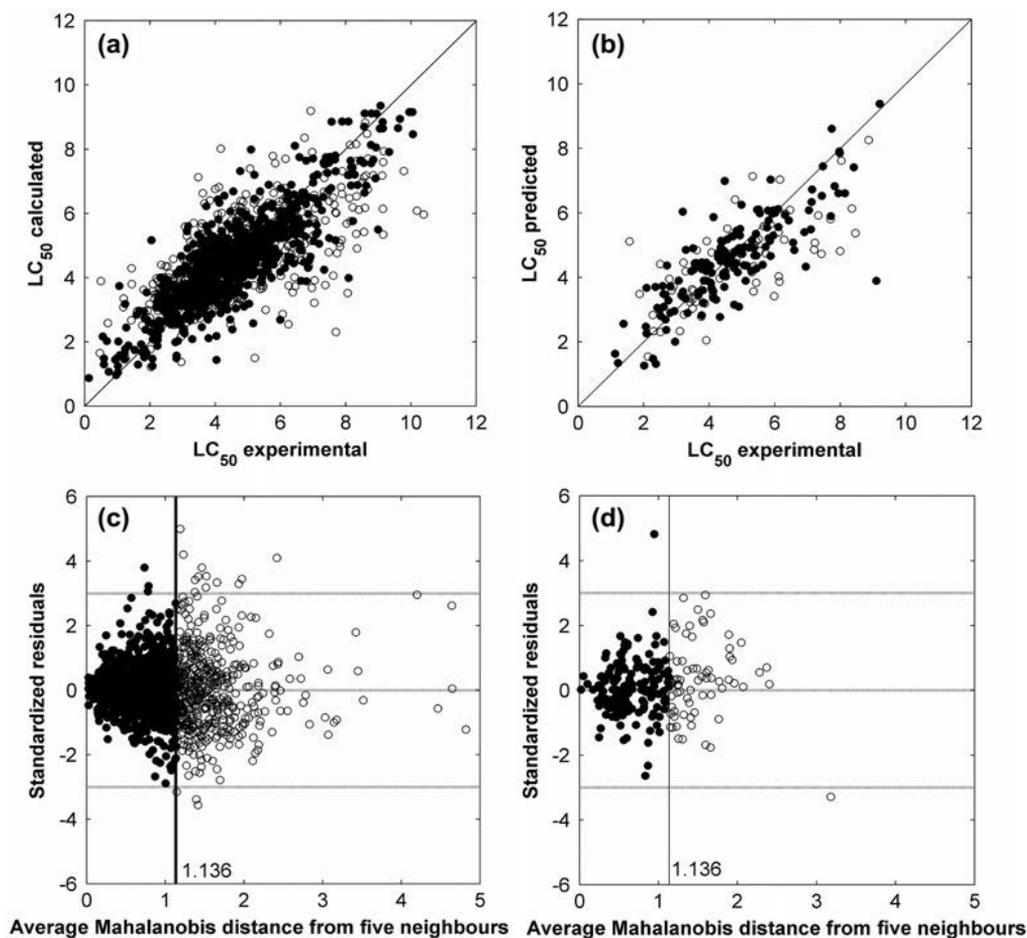
Figure 5. Extended MICHEM model: (a) experimental versus calculated LC_{50} for training set; (b) experimental versus predicted LC_{50} for test set; (c) standardized residuals versus average Mahalanobis distance from five neighbours for training set; and (d) standardized residuals versus average Mahalanobis distance from five neighbours for test set. Black circles: predicted molecules (average distance lower than the distance threshold); white circles: unpredicted molecules (average distance larger than the distance threshold). In subplots (c) and (d) the vertical line represents the distance threshold. The x-axis of subplot (c) was limited in the range [0–5] for visualization purposes: four molecules with average distance larger than five are therefore not shown.

fragments). It is widely known that molecules with longer carbon chains have higher lipophilicity in virtue of the low polarity, so it can be said that this type of binary fingerprints does not encode part of the information related to the lipophilicity of the compounds because, to a certain extent, they do not discriminate the length of carbon chains. Since lipophilicity plays a crucial role in aquatic toxicity, it occurs that molecules with significantly different lipophilicity, and therefore LC_{50} values, happen to be similar and the corresponding predictions suffer from large residuals. Table 4 details the case of hexylamine. The extended fragments (maximum *radius* equal to two) are the same for hexylamine and its six nearest neighbours (heptylamine, nonylamine, decylamine, dodecylamine, tetradecylamine and octadecylamine).
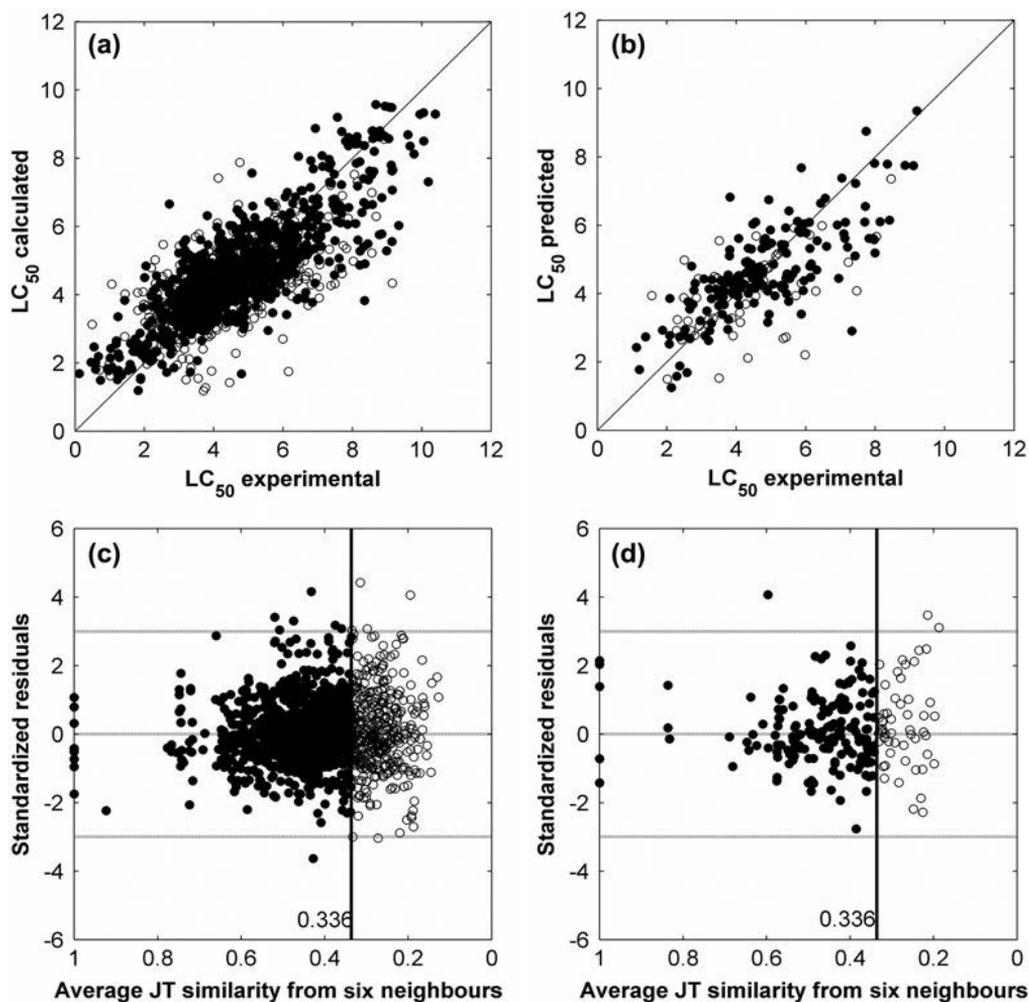
Figure 6.   Fingerprints model: (a) experimental *versus* calculated LC_{50} for training set; (b) experimental versus predicted LC_{50} for test set; (c) standardized residuals versus average similarity from six neighbours for training set; and (d) standardized residuals versus average similarity from six neighbours for test set. Black circles: predicted molecules (average similarity higher than the threshold); white circles: unpredicted molecules (average similarity lower than the threshold). In subplots (c) and (d) the vertical line represents the similarity threshold.

Consequently, the Jaccard–Tanimoto similarity coefficient is equal to one for all the pairs of molecules. However, the length of carbon chains has an effect on the LogP values (calculated in Table 4) and consequently on the experimental LC_{50} values, which span over more than 3 logarithmic units.

   Moreover, the model seems to have a slight bias: overestimation for low LC_{50} and under-estimation for high response values. The comparison of the performance of the two extended models shows that the fingerprint-based model has slightly lower performances (r^2, Q^2_{cv} and Q^2_{ext} equal to 0.67, 0.67 and 0.59, respectively), but the percentage of unpredicted molecules

Table 4. Details of a test molecule with identical fingerprints with its nearest neighbours.

CAS-RN	Structure	$MLOGP^a$	$LC_{50}{}^b$
111-26-2		1.59	4.07
111-68-2		1.94	4.09
112-20-9		3.50	4.95
2016-57-1		3.81	5.43
124-22-1		4.38	6.85
2016-42-4		4.91	7.15
124-30-1		5.90	5.20
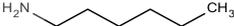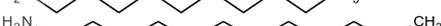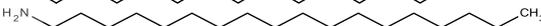
First row is the test molecule (hexylamine); rows 2–7 are the corresponding six nearest neighbours.
[a]Calculated Moriguchi octanol–water partition coefficient; [b]Experimental LC_{50} in –Log(mol/L).

(i.e. outside the AD) is lower at 29%, 33% and 24% in fitting, cross-validation and on the test set, respectively. Also for the fingerprint-based model there are molecules with average similarity greater than the threshold (i.e. inside the AD), but largely underestimated (standardized residuals greater than three). These compounds are: (a) diethyl 2-[(dimethoxyphosphorothioyl)sulfanyl] succinate (Malathion); (b) diethyl *p*-nitrophenyl phosphate (Paraoxon-ethyl); (c) bis(2-ethylhexyl) terephthalate (DEHT); (d) 2,2',3,3',4,4',5,5'-octachlorobiphenyl; (e) 2,4,6-tribromo-*N*-[2,4-dinitro-6-(trifluoromethyl)phenyl]-*N*-methylaniline (Bromethalin); and (f) *O*-ethyl *S,S*-bis(1-methylpropyl) phosphorodithioate (Cadusafos) in the training set ((Figure 6(c)) and docosanoic acid in the extended test set ((Figure 6(d)). DEHT is a phthalate plasticizer analysed earlier in Sections 3.2 and 3.3. Malathion, Paraoxon-ethyl and Cadusafos are organophosphate insecticides. 2,2',3,3',4,4',5,5'-Octachlorobiphenyl is a polychlorinated biphenyl (PCB), a class of toxic and persistent organic pollutants (POPs). Bromethalin is a rodenticide active on the central nervous system and docosanoic acid is a fatty acid with a long alkyl chain (C22). Docosanoic acid does not belong to any class of particularly toxic compounds. Since it features a long alkyl chain, the poor prediction may derive from the partly lacking information about the lipophilicity captured by this type of fingerprints, as discussed earlier in this paragraph. Additionally, there is one molecule in the training set ((Figure 6(c)) whose toxicity is largely overestimated. This molecule is triethyl phosphate, another organophosphate used as an intermediate for pesticides, a flame-retardant and a plasticizer.

Based on these results, it can be said that the modified *k*NN approach based on binary fingerprints is less effective than the extended MICHEM model based on DRAGON molecular descriptors as highlighted by the lower statistics. Still it could be claimed that the cause of the lower statistics lies in the molecular description by means of fingerprints, which partly misses information related to lipophilicity, rather than in the unsuitableness of the similarity-based approach.

3.5 *Consensus modelling*

The extended MICHEM and fingerprint-based models implement the same modified *k*NN approach, but have a different 'perspective' on the molecular structure. In fact, on one hand, the extended MICHEM model uses 'classical' DRAGON molecular descriptors that were selected by means of genetic algorithms [59,60] for being relevant to predict short-term toxicity towards *D. magna*. The selected eight descriptors encode information about the

lipophilicity and reactivity (formation of hydrogen and covalent bonds, electrophilicity and nucleophilicity) of the molecule – aspects that were related to aquatic toxicity. On the other hand, no supervised variable selection was carried out for the fingerprints since they provide a holistic description of the molecular structure that considers all the possible fragments and rings. The differences between the two models grant suitable conditions for the application of *consensus* modelling.

As aforementioned in Section 2.4, two approaches to *consensus* modelling were implemented with the intention to fulfil two different needs: higher quality predictions and broader applicability domain. In both cases, *consensus* analysis was based on the two available models, i.e. the extended MICHEM model and the fingerprint-based model. The 'strict' approach considers only molecules that are inside the AD of both models and provides a prediction that is the mean of the two model estimates. This approach is supposed to increase the accuracy of the predictions over the two single models, but also suffers from a narrow AD (the AD of the 'strict' approach is the intersection of the ADs of the extended MICHEM and fingerprints models because only molecules falling inside the AD of both models are considered). The 'loose' approach gives also a prediction for molecules that are inside the AD of just one model, either the extended MICHEM or the fingerprints model: the only available prediction is taken for molecules inside the AD of just one model. This latter approach is assumed to broaden the AD of the two single models because its AD is the union of the ADs of the extended MICHEM and fingerprints models, or in other words, the AD of the 'loose' approach is obtained by merging the ADs of the two single models.

The results of *consensus* modelling are reported in Table 1 and clearly show the benefits of using the *consensus* strategy. The 'strict' approach yields higher performance (r^2, Q^2_{cv} and Q^2_{ext} equal to 0.78, 0.78 and 0.73, respectively) than both the extended MICHEM and fingerprints models (even higher than the original MICHEM model shown in Table 1). The percentage of unpredicted molecules is large (47%, 52% and 42% in fitting, cross-validation and on the test set, respectively), but this was expected since the applicability domain is the intersection of the ADs of the two single models. The 'loose' approach has approximately the same performance as the extended MICHEM model (r^2, Q^2_{cv} and Q^2_{ext} equal to 0.70, 0.70 and 0.67, respectively), but the percentage of unpredicted molecules has been approximately halved (18%, 20% and 13% in fitting, cross-validation and on the test set, respectively). The AD of the 'loose' *consensus* model is, in fact, the union of the ADs of the extended MICHEM and fingerprints models. The experimental versus calculated/predicted responses for training and test sets obtained by means of both *consensus* approaches are shown in Figure 7. It is apparent that the 'strict' approach leads to high accuracy predictions as indicated by the very low number of predicted molecules (black circles) with high residuals. Figure 7(c) clearly shows that some of the predictions provided by the 'loose' approach are affected by large residuals, both positive and negative. Both approaches seem to have a little bias that implies overestimation for low and underestimation for large LC_{50} values.

Since the 'strict' *consensus* model is the one providing the most accurate toxicity estimates, a further analysis of this model was carried out in order to understand its behaviour for different classes of chemicals. Figure 8 shows a bubble plot for the 'strict' model on the extended test set and provides information about the number of molecules having a specific functional group and the related performance of the model. Only molecules inside the AD of the model are considered. The root mean square error of predictions (RMSEPs) for single functional groups tend to converge to the RMSEP over the entire extended test set along the x-axis, i.e. from the least to the most represented moiety. There is a large variability in the RMSEP values for functional groups that few molecules possess (left-hand side of Figure 8
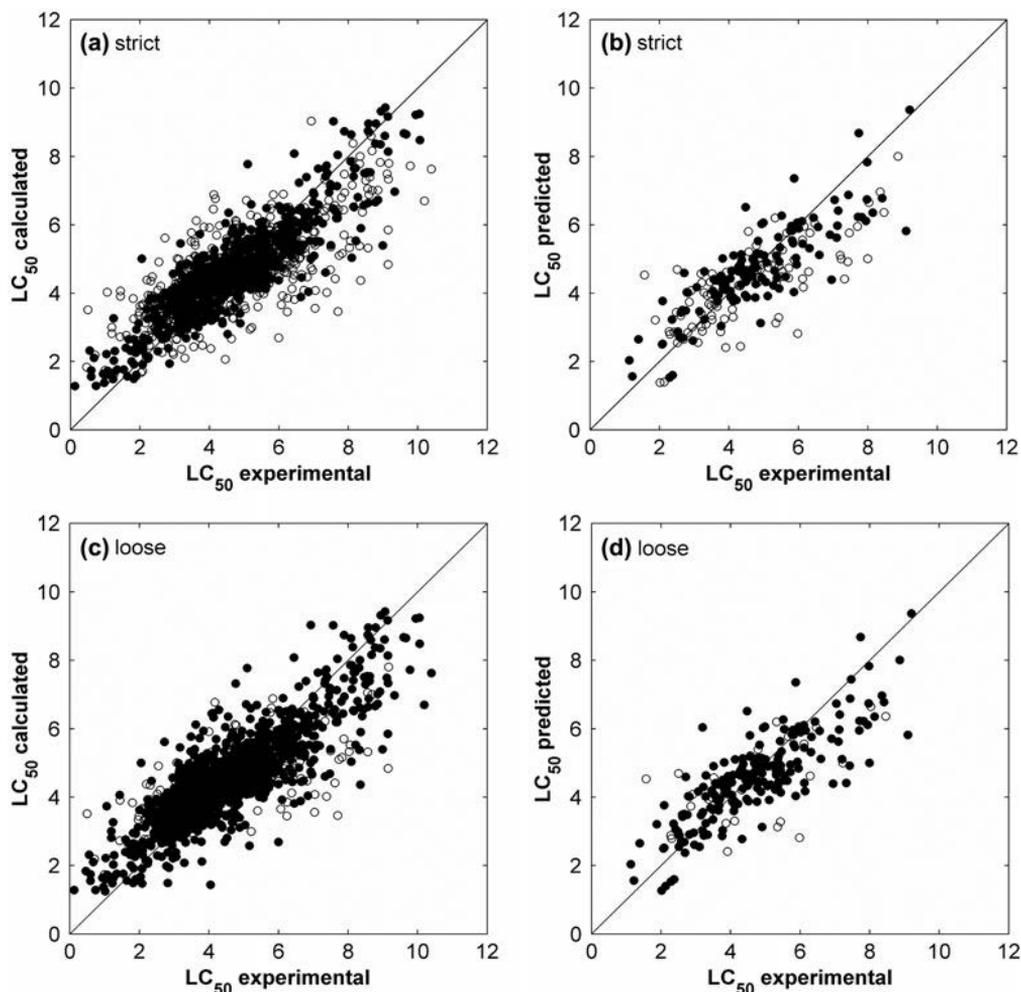
Figure 7. 'Strict' *consensus* model: (a) experimental versus calculated LC_{50} for training set; and (b) experimental versus predicted LC_{50} for test set. 'Loose' *consensus* model: (c) experimental versus calculated LC_{50} for training set; and (d) experimental versus predicted LC_{50} for test set. Black circles: molecules inside the AD; white circles: molecules outside the AD.

plot): for instance, sulfones, ('*nS(=O)2*'), aliphatic (thio)carbamates ('*nROCON*'), oxazoles ('*nOxazoles*') and pyridines ('*nPyridines*') and aliphatic tertiary amines ('*nRNR2*') are affected by very low errors, while pyrimidines ('*nPyrimidines*'), aromatic nitriles ('*nArCN*'), urea(-thio) derivatives ('*nCONN*') and oxiranes ('*nOxiranes*') show larger errors, with oxiranes being seriously badly predicted, probably due to their high reactivity. However, it would not be appropriate to derive general conclusions for these functional groups because the low number of molecules does not make a reliable statistical sample. On average, the model tends to have larger RMSEPs on aromatic compounds: for example, nitriles ('*nArCN*' versus '*nRCN*') and halides ('*nArX*' versus '*nCRX3*', '*nCH2RX*', '*nR=CHX*' and '*nR=CX2*'). Emblematic is the case of amines ('*nArNH2*' versus '*nRNH2*' and '*nArNR2*' versus

Figure 8. Bubble plot of the performances of the 'strict' *consensus* model on the extended test set. Each bubble corresponds to a functional group and its size is proportional to the number of molecules in the test set having that functional group. The vertical axis is the RMSEP of the model for only molecules with a specific functional group. The horizontal axis is a dummy variable used to sort the functional groups from the least to the most represented.

'*nRNR2*'): it is known that some aromatic amines can be activated to more toxic metabolites by enzymes like cytochrome P450 and these biotransformations, not accounted for in the model, might be one of the reasons for the lower performance on aromatic amines. There are also functional groups that show the opposite trend, such as carboxylic acids ('*nRCOOH*' versus '*nArCOOH*') and ketones ('*nRCO*' versus '*nArCO*'), or similar RMSEPs for the aliphatic and aromatic forms, for instance ethers ('*nArOR*' versus '*nROR*') and alcohols ('*nROH*' versus '*nArOH*'). It can be noticed that there is a different RMSEP for conjugated versus aromatic double bonds ('*nCconj*' versus '*nCar*' and '*nCconjX*' versus '*nArX*'), with the first ones having lower RMSEPs. Eventually the model seems to be slightly better at predicting H-bonds acceptors ('*nHAcc*') rather than donors ('*nHdon*'). Full information about RMSEC/RMSEP for individual functional groups for the all the developed models (extended MICHEM, fingerprints, 'strict' and 'loose' *consensus* models) is given in the supplementary material which is available via the multimedia link on the online article webpage, together with the definition of the symbols used in Figure 8.

3.6 *Comparison with published models*

The MICHEM model appeared to have comparable performance with other QSAR models based on large heterogeneous datasets [2]: the ChemProp model, as well as models proposed by Kar and Roy [61] and Kaiser and Niculescu [62]. The MICHEM model benefited from a simple algorithm, self-determination of its own AD and the possibility of carrying out a chemical analysis for each prediction based on the nearest neighbours. In Section 3.1, we further compared the MICHEM and ChemProp models because: (a) most of the data used for the validation and the extension of the approach derived from the ChemProp model; (b) both models are based on a similar approach; and (c) they claim to be applied for regulatory purposes and are therefore subject to the same applicability criteria. The MICHEM and ChemProp models were directly compared on three validation subsets and had similar performance on a common subset (Table 2).

The re-calibration of MICHEM model with the new data (extended MICHEM model) and the development of a novel model based on binary fingerprints, lead to the definition of two *consensus* models. The 'strict' *consensus* model provided significantly more accurate predictions on the extended test set (Q^2_{ext} equal to 0.73) compared to both ChemProp and the original MICHEM models on the 'External to both MICHEM and ChemProp' subset (Q^2_{ext} equal to 0.60 and 0.56, respectively). In addition, the percentage of molecules outside the AD was reduced to 42% in the 'strict' *consensus* model. The 'loose' *consensus* model also showed better accuracy (Q^2_{ext} equal to 0.67) than the ChemProp and MICHEM models, the biggest improvement being the percentage of molecules outside the AD (only 13%). The partial least squares (PLS) model of Kar and Roy [61] was developed on a training set of 222 compounds and gave a Q^2_{ext} equal to 0.74 on an ad hoc selected test set of 75 molecules. The probabilistic neural networks (PNNs) of Kaiser and Niculescu [62] showed similar performance (Q^2_{ext} equal to 0.76) on a test set of 76 compounds, but were developed on a larger training set (700 compounds). Recently, Golbamaki et al. [4] tested eight *in silico* modelling packages [Discovery Studio (DS) TOPKAT [63], ACD/Tox Suite [64], ADMET Predictor™ [65], ECOSAR (Ecological Structure Activity Relationships) [66], TerraQSAR™ [67], T.E.S.T. (Toxicity Estimation Software Tool) [68], and VEGA DEMETRA and VEGA EPA [69]) on a dataset containing 480 industrial and pesticide chemicals. The general performance of these models was not high. Four software packages (TerraQSAR™, T.E.S.T., VEGA DEMETRA and VEGA EPA) provided the training sets and allowed therefore to test the models only on external chemicals. The authors reported poor performance (Q^2_{ext} lower than 0.49) on the external test sets, which only slightly improved when considering only compounds inside the AD of each model (maximum Q^2_{ext} equal to 0.54 for VEGA EPA). For the other four *in silico* packages it was not possible to identify compounds present in the training sets and the test was made on all the data. The maximum performance was reached by DS TOPKAT (Q^2 equal to 0.63) on the 453 compounds inside the AD. The performance of the models developed by Kar and Roy and Kaiser and Niculescu is comparable with that of the 'strict' *consensus* model (Q^2_{ext} equal to 0.73) which was validated on a larger test set of 224 compounds (58% inside the AD). All the models detailed in the present study seem to outperform the *in silico* packages tested by Golbamaki et al. The 'loose' approach could provide more accurate predictions (Q^2_{ext} equal to 0.67) while retaining 87% of test set chemicals inside the AD; even higher accuracy was given by the 'strict' *consensus* model (Q^2_{ext} equal to 0.73) which, however, has a more narrow AD (only 58% of test set chemicals considered inside the AD).

4. Conclusions

The aim of this study was to further validate and optimize a recently proposed similarity-based QSAR model (MICHEM) for the prediction of short-term aquatic toxicity (LC_{50} 48 hours) towards *Daphnia magna*. A new large set of data was used to thoroughly validate the MICHEM model. In order to compare its performances with those of the ChemProp model, three validation subsets were extracted. Results showed that the MICHEM and ChemProp models performed similarly on a common subset of external compounds. The performances on the other two subsets, each external to one model, highlighted MICHEM as being slightly superior in terms of both Q^2_{ext} (0.66 versus 0.56) and percentage of unpredicted molecules (51% versus 54%).

The MICHEM model was re-calibrated on the novel data and showed slightly lower performance than the original MICHEM model, but a more stable behaviour in fitting, cross-validation and on the test set. The decreased percentage of unpredicted test molecules compared to the results on the 'External to MICHEM' subset (31% and 51%, respectively) reflected the wider applicability of the model. The accuracy of the predictions was also improved compared with the results on the 'External to MICHEM' subset (Q^2_{ext} equal to 0.69 and 0.66, respectively).

A new model based on the same kNN approach and defined by extended connectivity binary fingerprints was also calibrated on the same toxicity data. Less molecules were identified as being outside the applicability domain of the model but performances were lower than those of the extended MICHEM model. The likely reason being that this type of fingerprints can miss information about lipophilicity in case of long aliphatic chains.

Finally, two *consensus* models ('loose' and 'strict'), based on the extended MICHEM and the fingerprint-based models, yielded better results both in terms of accuracy of the predictions and width of the applicability domain. In fact, the 'strict' model gave r^2, Q^2_{cv} and Q^2_{ext} equal to 0.78, 0.78 and 0.73, respectively; the 'loose' model gave about the same statistics as the extended MICHEM model but the percentage of unpredicted molecules was approximately halved. *Consensus* models had similar performance on the external test set as published models developed from large heterogeneous datasets and seemed to outperform eight *in silico* packages tested on a set of 480 industrial and pesticide chemicals.

References

[1] European Union, *Regulation (EC) No 1907/2006 of the European Parliament and of the Council*, Off. J. Eur. Union L396 (2006), pp. 1–849.

[2] M. Cassotti, D. Ballabio, V. Consonni, A. Mauri, I.V. Tetko, and R. Todeschini, *Prediction of acute aquatic toxicity toward* Daphnia magna *by using the GA-kNN method*, Altern. Lab. Anim. 42 (2014), pp. 31–41.

[3] Organization for Economic Development and Co-operation (OECD), *Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models*, ENV/JM/MONO (2007)2, OCED Publishing, Paris. 2007.

[4] A. Golbamaki, A. Cassano, A. Lombardo, Y. Moggio, M. Colafranceschi, and E. Benfenati, *Comparison of* in silico *models for prediction of* Daphnia magna *acute toxicity*, SAR QSAR Environ. Res. 25 (2014), pp. 673–694.

[5] *Chemical Properties Estimation Software System (ChemProp)*, UFZ Department of Ecological Chemistry, Leipzig, 2013; software available at www.ufz.de/index.php?en=6738.

[6] R. Kühne and R.-U. Ebert P. C. von der Ohe, N. Ulrich, W. Brack, and G. Schüürmann, *Read-across prediction of the acute toxicity of organic compounds toward the water flea* Daphnia magna', Mol. Inform. 32 (2013), pp. 108–120.

[7] US Environmental Protection Agency (US EPA), *ECOTOX Database, Release 4.0*; available at http://cfpub.epa.gov/ecotox/.

[8] European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC), *TR 091 – ECETOC Aquatic Toxicity (EAT) database*, 2003; available at www.ecetoc.org/technical-reports.

[9] Organisation for Economic Co-operation and Development (OECD), *The OECD QSAR Toolbox for Grouping Chemicals into Categories*, 2013; available at: www.qsartoolbox.org.

[10] R.J. Bernot, M.A. Brueseke, M.A. Evans-White, and G.A. Lamberti, *Acute and chronic toxicity of imidazolium-based ionic liquids on* Daphnia magna, Environ. Toxicol. Chem. 24 (2009), pp. 87–92.

[11] W.F. Randall, W.H. Dennis, and M.C. Warner, *Acute toxicity of dechlorinated ddt, chlordane and lindane to bluegill (*Lepomis macrochirus*) and* Daphnia magna, Bull. Environ. Contam. Tox. 21 (1979), pp. 849–854.

[12] H. Sanderson and M. Thomsen, *Comparative analysis of pharmaceuticals versus industrial chemicals acute aquatic toxicity classification according to the United Nations classification system for chemicals. Assessment of the (Q) SAR predictability of pharmaceuticals acute aquatic toxicity and their predominant acute toxic mode-of-action*, Toxicol. Lett. 187 (2009), pp. 84–93.

[13] A. Jemec, T. Tišler, D. Drobne, K. Sepčić, D. Fournier, and P. Trebše, *Comparative toxicity of imidacloprid, of its commercial liquid formulation and of diazinon to a non-target arthropod, the microcrustacean* Daphnia magna, Chemosphere 68 (2007), pp. 1408–1418.

[14] E. Zou and M. Fingerman, *Effects of estrogenic xenobiotics on molting of the water flea,* Daphnia magna, Ecotox. Environ. Saf. 38 (1997), pp. 281–285.

[15] S.D. Costanzo, A.J. Watkinson, E.J. Murby, D.W. Kolpin, and M.W. Sandstrom, *Is there a risk associated with the insect repellent DEET (N, N-diethyl-m-toluamide) commonly found in aquatic environments?*, Sci. Total Environ. 384 (2007), pp. 214–220.

[16] C.A. Staples and J.W. Davis, *An examination of the physical properties, fate, ecotoxicity and potential environmental risks for a series of propylene glycol ethers*, Chemosphere 49 (2002), pp. 61–73.

[17] J.C. Martins, M.L. Saker, L.F.O. Teles, and V.M. Vasconcelos, *Oxygen consumption by* Daphnia magna Straus *as a marker of chemical stress in the aquatic environment*, Environ. Toxicol. Chem. 26 (2007), pp. 1987–1991.

[18] P.C. von der Ohe, R. Kühne, R.-U. Ebert, R. Altenburger, M. Liess, and G. Schüürmann, *Structural Alerts – A new classification model to discriminate excess toxicity from narcotic effect levels of organic compounds in the acute daphnid assay*, Chem. Res. Toxicol. 18 (2005), pp. 536–555.

[19] E.S. Williams, J.P. Berninger, and B.W. Brooks, *Application of chemical toxicity distributions to ecotoxicology data requirements under REACH*, Environ. Toxicol. Chem. 30 (2011), pp. 1943–1954.

[20] K. Nørgaard and N. Cedergreen, *Pesticide cocktails can interact synergistically on aquatic crustaceans*, Environ. Sci. Pollut. Res. 17 (2010), pp. 957–967.

[21] G.D. di Delupis, A. Macrí, C. Civitareale, and L. Migliore, *Antibiotics of zootechnical use: Effects of acute high and low dose contamination on* Daphnia magna Straus, Aquat. Toxicol. 22 (1992), pp. 53–59.

[22] B. Ferrari, R. Mons, B. Vollat, B. Fraysse, N. Paxēaus, R.L. Giudice, A. Pollio, and J. Garric, *Environmental risk assessment of six human pharmaceuticals: Are the current environmental risk assessment procedures sufficient for the protection of the aquatic environment?*, Environ. Toxicol. Chem. 23 (2004), pp. 1344–1354.

[23] K. Foit, O. Kaske, and M. Liess, *Competition increases toxicant sensitivity and delays the recovery of two interacting populations*, Aquat. Toxicol. 106–107 (2012), pp. 25–31.

[24] H. Ochoa-Acuña, W. Bialkowski, G. Yale, and L. Hahn, *Toxicity of soybean rust fungicides to freshwater algae and* Daphnia magna, Ecotoxicology 18 (2009), pp. 440–446.

[25] O. Horn, S. Nalli, D. Cooper, and J. Nicell, *Plasticizer metabolites in the environment*, Water Res. 38 (2004), pp. 3693–3698.

[26] K. Kyriakopoulou, P. Anastasiadou, and K. Machera, *Comparative toxicities of fungicide and herbicide formulations on freshwater and marine species*, Bull. Environ. Contam. Tox. 82 (2009), pp. 290–295.

[27] *DRAGON 6 (Software for Molecular Descriptor Calculation)*, Talete srl, Milan, 2012; software available at www.talete.mi.it.

[28] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, and R. Todeschini, *Comparison of different approaches to define the applicability domain of QSAR models*, Molecules 17 (2012), pp. 4791–4810.

[29] F. Sahigara, D. Ballabio, R. Todeschini, and V. Consonni, *Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions*, J. Cheminform. 5 (2013), p. 27.

[30] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, and S.K. Kearsley, *Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR*, J. Chem. Inf. Model. 44 (2004), pp. 1912–1928.

[31] J.R. Niemelä, E.B. Wedebye, N.G. Nikolov, G.E. Jensen, and T. Ringsted, *The advisory list for self-classification of dangerous substances*, Danish Environmental Protection Agency, Copenhagen, 2010; available at http://eng.mst.dk/topics/chemicals/assessment-of-chemicals/the-advisory-list-for-selfclassification/.

[32] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, *KNIME: The Konstanz Information Miner, in Studies in Classification, Data Analysis, and Knowledge Organization*, Springer-Verlag, Berlin, 2007, pp. 319–326.

[33] MathWorks Inc., *MATLAB*, Natick, MA, USA; software available at www.mathworks.com.

[34] P. Willett, V. Winterman, and D. Bawden, *Implementation of nearest-neighbor searching in an online chemical structure search system*, J. Chem. Inf. Comput. Sci. 26 (1986), pp. 36–41.

[35] P. Willett, *Similarity-based virtual screening using 2D fingerprints*, Drug Discov. Today 11 (2006), pp. 1046–1053.

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

[37] H.L. Morgan, *The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service*, J. Chem. Doc. 5 (1965), pp. 107–113.

[38] D. Rogers and M. Hahn, *Extended-connectivity fingerprints*, J. Chem. Inf. Model. 50 (2010), pp. 742–754.

[39] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni, *Quantitative structure–activity relationship models for ready biodegradability of chemicals*, J. Chem. Inf. Model. 53 (2013), pp. 867–878.

[40] F. Cheng, Y. Ikenaga, Y. Zhou, Y. Yu, W. Li, J. Shen, Z. Du, L. Chen, C. Xu, G. Liu, P.W. Lee, and Y. Tang, *In silico assessment of chemical biodegradability*, J. Chem. Inf. Model. 52 (2012), pp. 655–669.

[41] S. Lozano, M.-P. Halm-Lemeille, A. Lepailleur, S. Rault, and R. Bureau, *Consensus QSAR related to global or MOA models: Application to acute toxicity for fish*, Mol. Inform. 29 (2010), pp. 803–813.

[42] I.B. Stoyanova-Slavova, S.H. Slavov, B. Pearce, D.A. Buzatu, R.D. Beger, and J.G. Wilkes, *Partial least square and k-nearest neighbor algorithms for improved 3D quantitative spectral data–activity relationship consensus modeling of acute toxicity*, Environ. Toxicol. Chem. 33 (2014), pp. 1271–1282.

[43] N. Baurin, J.-C. Mozziconacci, E. Arnoult, P. Chavatte, C. Marot, and L. Morin-Allory, *2D QSAR consensus prediction for high-throughput virtual screening. An application to COX-2 inhibition modeling and screening of the NCI database*, J. Chem. Inf. Comput. Sci. 44 (2004), pp. 276–285.

[44] J.R. Votano, M. Parham, L.H. Hall, L.B. Kier, S. Oloff, A. Tropsha, Q. Xie, and W. Tong, *Three new consensus QSAR models for the prediction of Ames genotoxicity*, Mutagenesis 19 (2004), pp. 365–377.

[45] M. Ganguly, N. Brown, A. Schuffenhauer, P. Ertl, V.J. Gillet, and P.A. Greenidge, *Introducing the consensus modeling concept in genetic algorithms: Application to interpretable discriminant analysis*, J. Chem. Inf. Model. 46 (2006), pp. 2110–2124.

[46] M. Hewitt, M.T.D. Cronin, J.C. Madden, P.H. Rowe, C. Johnson, A. Obi, and S.J. Enoch, *Consensus QSAR models: Do the benefits outweigh the complexity?*, J. Chem. Inf. Model. 47 (2007), pp. 1460–1468.

[47] ChemAxon Ltd, *Marvin Suite*, 2012; software available at www.chemaxon.com.

[48] Milano Chemometrics and QSAR Research Group, *Acute aquatic toxicity to* Daphnia magna *dataset*; available at http://michem.disat.unimib.it/chm/download/toxicity.htm.

[49] V. Consonni, D. Ballabio, and R. Todeschini, *Comments on the definition of the Q2 parameter for qsar validation*, J. Chem. Inf. Model. 49 (2009), pp. 1669–1678.

[50] A. Golbraikh and A. Tropsha, *Beware of q2!*, J. Mol. Graph. Model. 20 (2002), pp. 269–276.

[51] European CHemical Agency (ECHA), available at http://echa.europa.eu/information-on-chemicals.

[52] A. Colombo, E. Benfenati, M. Karelson, and U. Maran, *The proposal of architecture for chemical splitting to optimize QSAR models for aquatic toxicity*, Chemosphere 72 (2008), pp. 772–780.

[53] M. Moosus and U. Maran, *Quantitative structure–activity relationship analysis of acute toxicity of diverse chemicals to* Daphnia magna *with whole molecule descriptors*, SAR QSAR Environ. Res. 22 (2011), pp. 757–774.

[54] K.S. Akers, G.D. Sinks, and T.W. Schultz, *Structure–toxicity relationships for selected halogenated aliphatic chemicals*, Environ. Toxicol. Phar. 7 (1999), pp. 33–39.

[55] European Union, *Regulation (EC) No 1272/2008 of the European Parliament and of the Council. Classification, labelling and packaging of substances and mixtures, amending and repealing Directives 67/548/EEC and 1999/45/EC, and amending Regulation (EC) No 1907/2006*, Off. J. Eur. Union L353 (2008), pp. 1–1355.

[56] K.E. Day and R.J. Maguire, *Acute toxicity of isomers of the pyrethroid insecticide deltamethrin and its major degradation products to* Daphnia magna, Environ. Toxicol. Chem. 9 (1990), pp. 1297–1300.

[57] I. Jolliffe, *Principal component analysis*, in *Encyclopedia of Statistics in Behavioral Science*, B.S. Everitt and D.C. Howell, eds., John Wiley & Sons, Chichester, UK, 2005. doi: 10.1002/0470013192.bsa501.

[58] P. Jaccard, *Nouvelles recherches sur la distribution florale*, Bull. Société Vaudoise Sci. Nat. 44 (1908), pp. 223–270.

[59] R. Leardi and A.L. González, *Genetic algorithms applied to feature selection in PLS regression: How and when to use them*, Chemometr. Intell. Lab. Syst. 41 (1998), pp. 195–207.

[60] J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, Bradford Books, Cambridge, MA, 1992.

[61] S. Kar and K. Roy, *QSAR modeling of toxicity of diverse organic chemicals to* Daphnia magna *using 2D and 3D descriptors*, J. Hazard. Mater. 177 (2010), pp. 344–351.

[62] K.L.E. Kaiser and S.P. Niculescu, *Modeling acute toxicity of chemicals to* Daphnia magna: *A probabilistic neural network approach*, Environ. Toxicol. Chem. 20 (2001), pp. 420–431.

[63] *Discovery Studio TOPKAT*, Accelrys, Inc.; software available at http://accelrys.com/products/discovery-studio/admet.html.

[64] *ACD/Tox Suite*, ACD/LabsInc.; software available at www.acdlabs.com/products/pc_admet/tox/tox/.

[65] *ADMET Predictor*[TM], Simulations Plus, Inc.; software available at www.simulations-plus.com/Products.aspx?pID=13.

[66] *ECOSAR (Ecological Structure Activity Relationships)*, US Environmental Protection Agency; software available at www.epa.gov/oppt/newchems/tools/21ecosar.htm.

[67] *TerraQSAR*[TM]. TerraBase, Inc.; software available at www.terrabase-inc.com.

[68] T.*E.S.T. (Toxicity Estimation Software Tool)*, US Environmental Protection Agency; software available at www.epa.gov/nrmrl/std/qsar/qsar.html#TEST.

[69] *VEGA Non-Interactive Client*, Istituto di Ricerche Farmacologiche Mario Negri. Milan, Italy; software available at www.vega-qsar.eu.

Appendix III

Cassotti, M., Ballabio, D., Todeschini, R., Consonni, V., 2015. A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*). *SAR QSAR Environ. Res.*, 26 (3), 217-243.

Available at:

Taylor & Francis
Taylor & Francis Group

A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*)

M. Cassotti*, D. Ballabio, R. Todeschini and V. Consonni

Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milano, Italy

REACH regulation demands information about acute toxicity of chemicals towards fish and supports the use of QSAR models, provided compliance with OECD principles. Existing models present some drawbacks that may limit their regulatory application. In this study, a dataset of 908 chemicals was used to develop a QSAR model to predict the LC_{50} 96 hours for the fathead minnow. Genetic algorithms combined with k nearest neighbour method were applied on the training set (726 chemicals) and resulted in a model based on six molecular descriptors. An automated assessment of the applicability domain (AD) was carried out by comparing the average distance of each molecule from the nearest neighbours with a fixed threshold. The model had good and balanced performance in internal and external validation (182 test molecules), at the expense of a percentage of molecules outside the AD. Principal Component Analysis showed apparent correlations between model descriptors and toxicity.

Keywords: QSAR; fathead minnow; aquatic toxicity; REACH; similarity; kNN

1. Introduction

The entrance into force of REACH regulation [1] in June 2007 boosted the interest in the field of *in silico* methodologies. In fact, REACH, by introducing the concept 'no data, no market', obliged manufacturers and importers to prove that their products are safe for both human health and the environment. The avoidance of unnecessary testing (especially animal testing) being an explicitly declared goal, REACH provided registrants with a number of tools to pursue this objective, including the promotion of alternative test methods, such as *in vitro* and *in silico* methodologies.

Among the latter methods, quantitative structure–activity relationships (QSAR) analysis emerged as one of the suggested approaches, because of its low cost and time for application. QSAR analysis comprises a variety of mathematical and statistical methods that aim at finding functional relationships between the structure of chemical compounds, described by means of experimental or theoretical variables called molecular descriptors [2], and their measured properties and activities.

As part of the assessment of toxicity towards aquatic organisms, REACH requires the evaluation of the short-term toxic effects on fish for substances imported or manufactured in quantities greater than 10 tonnes per year (REACH Annex VIII). The benefits deriving from the availability of suitable QSAR models, both from an economical and animal welfare

---

*Corresponding author. Email: m.cassotti@campus.unimib.it

perspective, are evident. The idiomatic expression 'suitable QSAR model' indicates that the model should be scientifically valid, and the scientific validity for regulatory applications within REACH is outlined in the five principles defined by the Organization for Economic Co-operation and Development (OECD) [3]. In summary, the endpoint and the algorithm should be clearly defined, the model should be accompanied by an estimation of its domain of applicability, the goodness-of-fit and predictivity of the model should be evaluated by means of appropriate strategies and, eventually, a mechanistic interpretation of model descriptors should be given, if possible.

A number of QSAR models have been developed to predict acute toxicity towards fish, and two trends can be identified: some researchers have aimed at classifying chemicals for their mode of action (MoA) [4–6], whereas others have tackled the problem of estimating a quantitative parameter, usually the LC_{50} [7–10]. Considering quantitative models for the fathead minnow (*Pimephales promelas*), some studies focused on small homogeneous sets of chemicals belonging to the same chemical class or supposed to act via the same MoA [11–16]. The use of MoA-based QSARs for toxicity screening depends on the ability to associate query chemicals with the correct MoA, which is not an easy task. Consequently, many investigations also aimed to quantitatively model large heterogeneous datasets altogether. Mainly global strategies were employed to this end. A summary of the characteristics of quantitative models for large heterogeneous datasets is given in Table 1.

Regarding linear methods, multiple linear regression (MLR) was used in many investigations [17–30], whereas partial least squares (PLS) and the multi-linear spline regressions were seldom used [17,19,22]. On the other hand, more complex non-linear methods, such as different types of neural networks (NN) and support vector regression (SVR), were often used to model such heterogeneous datasets [20,22,23,26,31–35]. Some investigations divided chemicals into more homogeneous clusters (not necessarily corresponding to chemical classes or known MoAs) in a preliminary step and calibrated local regression models for each cluster [20,21,36–38]. The k nearest neighbours (kNN) method was also used to derive models that can be considered local because just a small neighbourhood of similar chemicals is used to estimate the toxicity of the query compound [21,39]. Read-across based on kNN was also used to assess the excess toxicity from a baseline estimated from the log P [39]. The statistics of these models, in general, were lower compared with those of models developed for specific chemical classes or modes of action.

This study focuses on the development of a new QSAR model to predict the acute toxicity of diverse chemicals, defined as LC_{50} 96 hours, towards the fathead minnow (*Pimephales promelas*). The model was developed keeping in mind the five OECD principles in order to make it applicable for regulatory purposes within REACH. To this end, attention was paid to the curation of the experimental data, which led to the definition of an extended dataset consisting of 908 organic molecules. The model, based on six molecular descriptors, used a similarity-based algorithm (kNN) to predict the toxicity. The applicability domain (AD) was automatically evaluated for each prediction and an additional analysis of the performance was carried out for individual functional groups. The predictive power was estimated by means of thorough and appropriate internal and external validation procedures. Moreover, the chemical information encoded by model descriptors was explained, and we attempted to put it in relation with aquatic toxicity. Eventually, an example of the application of the model was given.

Table 1. Characteristics of literature models for LC_{50} towards the fathead minnow developed from large heterogeneous datasets. In case of multiple models, the range of the statistics is reported between square brackets. The dash symbol is used for lacking information.

Reference	No. models[a]	Method[b]	n train[c]	n test[d]	p[e]	r^2	Q^2_{cv}	Q^2_{ext}
			Heterogeneous datasets					
[17]	2	MLR	[560–568]	–	1	[0.61–0.65]	[0.61–0.65][f]	–
	20	MLR	568	–	2	[0.64–0.67]	[0.64–0.66][f]	–
	20	MLR	568	–	3	[0.67–0.68]	[0.67–0.68][f]	–
	20	MLR	568	–	4	[0.69–0.70]	[0.68–0.69][f]	–
	2	PLS	562	–	13	[0.72–0.73]	[0.71–0.72]	0.72
[18]	1	MLR	408	57	4	0.80	0.80[g]	0.72
[19]	6	MLR	344	115	[6–10]	[0.76–0.79]	[0.75–0.76][f]	[0.75–0.79]
	3	MLR-Spline	344	115	[5–7]	[0.76–0.78]	[0.75–0.77][f]	[0.78–0.78]
[20]	1	MLR	445	110	5	0.71	–	0.55
	1	ANN	334 + 111[h]	110	5	0.80	0.62[i]	0.62
[36]	2	RP-MLR	445	110	[5–5][j]	[0.75–0.76]	–	[0.60–0.63]
	8	*Consensus*	445	110	–	[0.78–0.80]	–	[0.63,0.67]
[31]	1	NN + NN	454	114	156	–	–	0.76
[39]	1	SVR	457	114	8	0.83	–	0.80
	2	LR + *k*NN	692	–	–	[0.73–0.73]	[0.72–0.72][f]	–
	2	LR + *k*NN (0.8)	692 (419)[k]	–	–	[0.78–0.78]	[0.76–0.78][f]	–
	2	LR + *k*NN (0.9)	692 (230)[k]	–	–	[0.87–0.87]	[0.87–0.87][f]	–
[32]	11	Fuzzy NN	392	170	9	[0.20–0.70]	–	[0.00–0.50]
[33]	1	CPNN	275	274	150	0.97	–	0.59
[21]	1	HC + MLR	659	164	$\leq n_k/5$[l]	–	–	0.71
	1	SC + MLR	659	164	–	–	–	0.63
	2	MLR	659	164	–	–	–	[0.69–0.70]
	1	*k*NN	659	164	–	–	–	0.67
	1	*Consensus*	659	164	–	–	–	0.73
[34]	2	MLR + NN	484	85	[10–23]	[0.82–0.75][m]	[0.46–0.54][i]	[0.64–0.66][m]
[22]	3	MLR	607[n]	–	147	[0.83–0.87]	[0.59–0.67][i]	–
	3	PLS	607[n]	–	147	[0.81–0.83]	[0.62–0.70][i]	–
	3	NN	607[n]	–	147	[0.89–0.92]	[0.48–0.60][i]	–
	3	PMM	607[n]	–	147	[0.79–0.82]	–	–
[37]	1	Tree + MLR	560	–	[3–5][j]	[0.83–0.99]	[0.81–0.98][i,o]	–

(Continued)

Table 1.   (*Continued*).

			Heterogeneous datasets					
Reference	No. models[a]	Method[b]	n train[c]	n test[d]	p[e]	r^2	Q^2_{cv}	Q^2_{ext}
[23]	1	MLR	287	88[p]	8	0.83[m]	0.90[i,m]	0.75[m]
	2	NN	287	88[p]	[8–8]	[0.81–0.71][m]	[0.88–0.77][i,m]	[0.84–0.74][m]
[24]	1	MLR	484	121	2	0.71	0.70	0.61
	2	*Consensus*	484	121	–	–	[0.71–0.71]	[0.59–0.60]
[38]	1	MLR + L-MLR	675	–	[2–8][j]	0.88	0.95[q]	–
[25]	10	MLR	557	–	[4–17]	[0.62–0.73]	[0.50–0.70][r]	–
	1	*Consensus*	557	201 + 144[s]	–	0.71	–	0.60 + 0.58[t]
[26]	1	MLR	373	188	6	0.73	0.72[f]	0.66
	1	BPNN	373	188	6	0.78	–	0.73
[27]	1	MLR	532	–	2	0.61	–	–
[35]	2	PNN	800	86	76	[0.89–0.99]	–	[0.78–0.52]
[28]	2	MLR	249	200	[5–6]	[0.79–0.81]	[0.78–0.80][g]	[0.71–0.72]
[29]	3	MLR	[246–271] + [144–164][h]	[148–158]	[1–1]	[0.67–0.68]	[0.79–0.85][i]	[0.77–0.79]
[30]	1	MLR[u]	652	164(39)[k]	21	0.69	–	0.69

[a]Number of developed models; [b]MLR = Multiple linear regression; PLS = Partial least squares; RP-MLR = Recursive partitioning coupled with MLR for each class; NN + NN = Clustering by means of self-organized neural network coupled with local regression by means of feedforward neural networks; SVR = Support vector regression; LR + kNN = Linear regression coupled with k nearest neighbours; LR + kNN (0.8) = Linear regression coupled with k nearest neighbours with similarity threshold = 0.8; LR + kNN (0.9) = Linear regression coupled with k nearest neighbours with similarity threshold = 0.9; Fuzzy NN = fuzzy neural network; CPNN = CounterPropagation neural network; HC + MLR = Hierarchical clustering coupled with multiple linear regression; SC + MLR = Single clustering coupled with multiple linear regression; kNN = k nearest neighbours; MLR + NN = Multiple linear regression for base-line coupled with neural networks to model the residuals; PMM = Powell's minimization method; Tree + MLR = decision tree to partition chemicals into nine clusters coupled with MLR for each cluster; MLR + L-MLR = Multiple linear regression for baseline coupled with local MLR for clusters of molecules sharing a common toxicophore; BPNN = BackPropagation neural network; PNN = Probabilistic neural network; [c]number of compounds in training set; [d]number of compounds in the test set; [e]number of model descriptors; [f]leave-one-out cross-validation; [g]bootstrap with 5000 iterations; [h]training set + validation set; [i]Q² on the validation set; [j]number of descriptors in each MLR; [k]number of compounds inside the applicability domain; [l]maximum allowed number of descriptors in each cluster model (n_k = number of compounds in the cluster); [m]Root mean square residuals; [n]split in training-validation sets (2:1) three times; [o]range of statistics for the nine cluster models. Overall statistics not reported; [p]divided into validation and test sets; [q]leave-10%-out cross-validation, repeated three times; [r]3-fold cross-validation; [s]number of molecules in each of the two test sets; [t]results on each of the two test sets; [u]re-implementation in VEGA of the MLR model of [21].

2. Materials and methods

2.1 *Experimental data*

Experimental acute toxicities of chemicals were retrieved from three databases, namely OASIS, ECOTOX [40] and EAT5 [41]. The OASIS database was downloaded from the OECD QSAR Toolbox [42]. The databases were imported into KNIME [43] and processed by means of *ad hoc* designed workflows in order to extract the concentrations causing death in 50% of test fathead minnows (*Pimephales promelas*) over a test duration of 96 hours (LC_{50} 96 hours). Experimental data were merged together regardless of test conditions (water pH, temperature, etc.) and test designs (flow-through, static, static renewal). In the EAT5 database, LC_{50} data were reported as EC_{50} (effective concentration) with lethality as observed effect. Records in the ECOTOX database indicating ranges or thresholds of experimental values were removed.

2.2 *Data curation and filtering*

In order to guarantee data consistency, data were checked and ambiguous molecular structures and anomalous experimental values were disregarded. Data curation and filtering were carried out in KNIME.

2.2.1 *Checking identity of records*

For almost every record (4626 records in total) both the CAS registry number (CAS-RN) and chemical name were available. In order to check that the CAS-RN and chemical name referred to the same structure, queries were set up to the ChemSpider database [44] and the Chemical Identifier Resolver (CIR) of the CADD Group at NCI/NIH [45]. CAS-RNs and chemical names were used independently as input for the queries. The retrieved structures were compared and if they all matched, the identity was considered correct.

Out of 4626 records (corresponding to 1139 unique CAS-RNs, plus 12 compounds lacking a CAS-RN), more than 50% presented mismatches (2422 records, corresponding to 518 different CAS-RNs and the 12 compounds lacking a CAS-RN). The records showing mismatches were exported and checked manually using PubChem [46], Sigma-Aldrich [47] and again ChemSpider as additional sources. Some records were deleted during this screening for different reasons, such as: (a) non-existent CAS-RN; (b) missing specification of which structural isomer(s) had been used; (c) unavailability of the molecular structure as it was a commercially named chemical; (d) impossibility to resolve a CAS-RN–chemical name mismatch, for example because the original publication was not found or not accessible; and (e) the record pertained to a mixture of several chemical species. At the end of this phase, 2192 records corresponding to 441 different CAS-RNs were retained and merged with the 2204 records (692 different CAS-RNs) with matching structures, giving a set of 4396 records (1060 different CAS-RNs, with 73 of the 441 CAS-RNs from the mismatching set present among the 692 CAS-RNs from the matching set).

A further inspection was carried out in order to check that each CAS-RN was associated with only one structure and vice versa. This investigation lead to the identification of 10 structures associated with two different CAS-RNs. These mismatches, probably due to obsolete CAS-RNs, were resolved by retaining only the CAS-RN indicated on the Sigma-Aldrich database.

2.2.2 *Filtering and dissociation*

Records with units coded as '%', '% v/v' and 'AI ng/L' were removed (13 records). The LC_{50} values of the remaining 1047 molecules were converted to molarity and transformed to logarithmic units ($-Log_{10}(mol/L)$).

Several molecules had multiple experimental values, which could correspond to: (a) different measurements; (b) the same measurement published in a paper that had been included in more source databases; and (c) the same measurement published in different papers. Since the median of the LC_{50} values would have been used for modelling, duplicates of the same measurement (cases b and c) had to be removed because they would have affected the calculation of the median. Duplicates of the same experiment from different databases (case b) were removed. Duplicates of the same experiment published in more papers (case c) often lacked references. Therefore, it was decided to consider all the records with exactly the same LC_{50} value as duplicates of the same experiment and only one record was retained. This decision was taken considering that experimental measurements are implicitly affected by error, thus the probability that two measurements would give the same LC_{50} value is, in principle, extremely low.

Since the goal was to develop a model for acute toxicity limited to discrete organic molecules, only molecules with at least two carbon atoms and comprising only certain elements were retained (H, Li, B, C, N, O, F, Na, Mg, P, S, Cl, K, Ca, Br, and I). Symbols that specify the stereochemical configuration were removed from the *Simplified Molecular Input Line Entry System* (SMILES). Salts and mixtures were submitted to a dissociation algorithm in the OASIS Database Manager [48] that first checked whether the species could be dissociated and then screened the potential dissociation products for non-toxic species. If more than one species was considered the source of toxicity, the record was removed. By doing so, it was possible to convert 50 mixtures and salts to a single organic component, assumed as the only source of the measured toxicity. Ions such as Na^+, Mg^{2+}, Cl^- were therefore not used for modelling. The dissociation products were neutralized, unless they were quaternary ammonium ions for which the charged form was retained. An outcome encountered for three salts was that both the organic ion (acetate, benzoate and 2-hydroxybenzoate) and the inorganic counter-ion (K^+ or Na^+) were considered not toxic by the algorithm and, consequently, removed. For these three cases, the organic component was re-introduced and considered for modelling. In 15 cases, the dissociation product coincided with another molecule in the dataset. Toxicity values of these two species were very close for most instances, thus justifying the validity of the dissociation procedure and allowing us to pool the data. In four cases, mixtures of the type A+A+B were present and the dissociation algorithm returned only one molecule of A as assumed source of toxicity. The LC_{50} (molarity) values were accordingly doubled to correct for the approximation.

At the end of this filtering stage, 929 molecules were retained. Final validation of the structures was made by comparing the SMILES in the dataset with those in the OpenTox database after processing also the latter ones with the dissociation converter. OpenTox database lacked a structure for 58 compounds in the dataset. Large agreement in the structures of the remaining 871 molecules was observed. Only nine mismatches were detected and solved by looking for the correct structure in the Sigma-Aldrich database. Only one case consisted of completely different compounds, whereas the other differences were mainly due to tautomers, valence and charge.

2.2.3 *Curation of experimental values*

For several compounds, multiple experimental values were available, showing differences of up to three logarithmic units. In order to reduce dependence on outlying toxicity data, the median, which is a more robust measure of central tendency than the mean, was calculated together with the corresponding standard deviation on the logarithmically transformed molarities ($-Log_{10}$(mol/L)). The pooled standard deviation over the entire dataset was calculated ($\sigma = 0.229$ Log_{10}(mol/L)) and used to derive an alert for inconsistent data ($2\sigma = 0.458$ Log_{10}(mol/L)). Molecules with a standard deviation larger than 2σ were filtered out and each experimental value was searched in the original scientific publication in order to detect errors in the compilation of the databases. If the scientific publication was not available or not found, the corresponding experimental value was deleted. During this phase, 21 chemicals with large standard deviations were removed because none of the original publication was accessible or found.

The final dataset included 908 organic molecules and is freely available [49].

2.3 *Molecular descriptors*

The SMILES of the 908 chemicals in the dataset were used to calculate molecular descriptors by means of DRAGON 6 software [50]. Only zero-, one- and two-dimensional descriptors were calculated from the SMILES for a total of 3582 descriptors. Descriptors from the Drug-like block were not calculated because they were supposed not to be relevant for modelling aquatic toxicity. Constant, near constant and descriptors with at least one missing value were removed (1361 descriptors). Eventually, a filter on pairwise correlation was applied: if two descriptors had a coefficient of correlation greater than 0.95, only the one with the lowest average correlation with all the remaining descriptors was retained. A final pool consisting of 1218 molecular descriptors was retained and used for the subsequent modelling phase. The distribution of the 1218 retained descriptors in the 18 logical blocks of DRAGON was as follows: constitutional indices (32), ring descriptors (25), topological indices (29), walk and path counts (15), connectivity indices (14), information indices (25), 2D matrix-based descriptors (66), 2D autocorrelations (160), Burden eigenvalues (47), P_VSA-like descriptors (33), ETA indices (13), edge adjacency indices (85), functional group counts (90), atom-centred fragments (66), atom-type E-state indices (35), CATS 2D (98), 2D atom pairs (378), molecular properties (7).

2.4 *Modelling methods*

Since the compounds in the dataset belong to a variety of chemical classes, it is expected that they also possess different MoAs. Literature models calibrated on the largest datasets were based on either non-linear methods, or similarity-based methods, or partitioned chemicals into more homogeneous clusters for which linear models were calibrated (Table 1). This is likely to be due to the differences between MoAs. Therefore, linear modelling methods were expected not to be optimal. Among the several potentially appropriate non-linear methods, it was decided to use the *k*NN method because: (a) it is simple; (b) only the local neighbourhood is used to provide a prediction (presumably chemicals acting via the same MoA); and (c) it allows a chemical analysis for each test molecule and its nearest neighbours.

The distance of a molecule from all the molecules in the training set is computed and the *k* training molecules with the lowest distances are selected. The experimental response values

of the *k* closest training molecules, i.e. the nearest neighbours, are used to calculate the prediction. Obviously, in fitting, the distance of a molecule with itself is neglected. The Jaccard–Tanimoto distance was used to find the nearest neighbours. The Jaccard–Tanimoto distance between two molecules *r* and *t*, d_{rt}, was derived from the corresponding coefficient as [51]:

$$
\begin{aligned}
d_{rt} &= \left(1 - \frac{\sum_{j=1}^{p} x_{rj} \cdot x_{tj}}{\sum_{j=1}^{p} x_{rj}^2 + \sum_{j=1}^{p} x_{tj}^2 - \sum_{j=1}^{p} (x_{rj} \cdot x_{tj})} \right)^{1/2} \\
&= \left(1 - \frac{d_{rt,euclidean}^2}{\sum_{j=1}^{p} x_{rj}^2 + \sum_{j=1}^{p} x_{tj}^2 - \sum_{j=1}^{p} (x_{rj} \cdot x_{tj})} \right)^{1/2} \quad 0 \le d_{rt} \le 1
\end{aligned}
\tag{1}
$$

where *j* runs over the *p* variables. Then, the prediction $\hat{y}_r$ for molecule *r* was taken as the weighted mean over the *k* nearest neighbours, where the weights were calculated as a function of the distance, according to Equation (2):

$$
\hat{y}_r = \sum_{t=1}^{k} y_t \cdot w_t = \sum_{t=1}^{k} y_t \cdot \frac{(1 - d_{rt})}{\sum_{t=1}^{k} (1 - d_{rt})}
\tag{2}
$$

where y_t and w_t are the experimental response and the weight of the *t*th neighbour, respectively, and the sum runs over the *k* neighbours. Molecular descriptors were scaled in the range [0,1] prior to computing the distances.

The *k*NN method was combined with genetic algorithms (GAs) in order to select the relevant molecular descriptors.

2.5 *Applicability domain assessment*

A two-step procedure was implemented in order to have an in-depth assessment of the AD of the model. A preliminary bounding box approach is carried out prior to finding the nearest neighbours. Test compounds with descriptors values outside the range of the training set are therefore considered outside the AD. For all the compounds that have descriptors values within the range of the training set (and are therefore considered inside the AD by the bounding box approach), the further evaluation of the AD is based on the distance with the nearest neighbours as described below.

A number of *k*NN similarity-based approaches have been defined in the scientific literature [52–54] to assess the AD of QSAR models. These methods are based on the calculation of a similarity measure of the molecule to be predicted with respect to molecules in the training set. The similarity can be calculated from the distances of the molecule to be predicted from its nearest neighbours. The obtained distance (or similarity) measure is then compared with a user-defined threshold. If the distance is within the threshold, the query molecule is considered to have enough similar neighbours to assure a reliable prediction: the molecule falls inside the AD of the model. With respect to other basic AD approaches (such as bounding box), *k*NN AD approaches better describe the molecules distribution because they can locally describe the covariance structure of the data. Therefore, *k*NN AD approaches better define the QSAR model space but, obviously, there may be training molecules considered outside the AD, while a bounding box method would include all training molecules.

In this study we used the same approach previously applied to predict acute toxicity towards *D. magna* [55]. In summary, for each molecule the average distance from its *k* nearest neighbours taken from the training set was compared with a fixed distance threshold. If the average distance was greater than the distance threshold, the molecule to be predicted

and its *k* nearest neighbours were regarded to as relatively dissimilar, and therefore the molecule was considered outside the AD. On the contrary, if the average distance was lower than the distance threshold, the molecule was considered to be enough similar to its neighbours (in the training set) to allow a reliable prediction and said to be inside the AD of the model. In this way, the evaluation of the AD is implicitly defined on a *k*NN similarity-based approach and carried out on the fly for each prediction.

Since the model is not parametric and is based on local similarities, compounds from the training set that did not have sufficiently similar neighbours were still retained in the training set because they could be useful for the prediction of future test compounds.

2.6 *Model validation*

In order to properly validate the model, the dataset of 908 compounds was randomly divided into a training set (726 chemicals) and a test set (182 molecules). The training set was used to carry out variable selection by means of GAs, following the strategy proposed by Leardi and González [56], and calibrate the final model. The settings used for GAs are reported in Table 2.

During GA runs, the performance of the models was assessed by means of internal five-fold cross-validation with venetian blinds splitting of the training samples. The coefficient of determination in cross-validation (Q^2_{cv}), defined according to Equation (3), was used as fitness function.

$$Q^2_{cv} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (3)$$

where y_i and $\hat{y}_i$ are the experimental and predicted responses of the *i*th object, respectively; $\bar{y}$ is the average response value. Genetic algorithms were used to select molecular descriptors, optimize the number of nearest neighbours (*k*) and the distance threshold used for the assessment of the AD. Values of *k* (number of nearest neighbours) from 1 to 10 were tried for each model.

In order to carry out a more thorough validation, the final model was also internally validated by means of leave-more-out strategy: 20% of training molecules, randomly selected, were left out. The procedure was reiterated 1000 times and the average coefficient of determination (Q^2_{cv}) was calculated.

Table 2. Settings used for genetic algorithms.

Option	Value	option	Value
number of chromosomes	30	twins allowed	no
average number of variables in the chromosomes of starting population	5	hybridization	yes
mutation probability	0.01	frequency of hybridization	1 every 100 runs
cross-over probability	0.50	number of cross-validation groups	5
number of independent runs	100	cross-validation type	venetian blinds
number of evaluations for each run	100	maximum number of nearest neighbours	10

Finally, the test set was used to assess the predictive power of the model calibrated on the training set. The performance on the test set was assessed by means of the Q^2_{F3} function [57], defined as:

$$Q^2_{ext} = 1 - \frac{PRESS/n_{ext}}{TSS/n_{tr}} = 1 - \frac{\left[\sum_{i=1}^{n_{ext}}(y_i - \hat{y}_i)^2\right]/n_{ext}}{\left[\sum_{i=1}^{n_{tr}}(y_i - \bar{y})^2\right]/n_{tr}} \tag{4}$$

where n_{tr} and n_{ext} are the number of molecules in the training and test sets, respectively.

2.7 Software

KNIME [43] was used to extract the relevant data from the source databases and process them by means of *ad hoc* designed workflows. The OASIS Database Manager [48] was used to retain only organic compounds, apply the dissociation converter and compare the SMILES in the dataset and in the OpenTox database. DRAGON 6 [50] was used to calculate molecular descriptors and apply unsupervised variable reduction. MATLAB [58] was used to carry out variable selection and model validation by means of routines written by the authors. Marvin was used for drawing, displaying and characterizing chemical structures and substructures [59]. ChemProp [60] was used to retrieve the molecules used for the application example (paragraph 3.4).

3. Results and discussion

3.1 Model development and analysis

3.1.1 Variable selection and model calibration

Variable selection was carried out on the training set of 726 chemicals in subsequent steps in order to handle the large number of calculated descriptors, i.e. 1218, and avoid potential over-fitting. First, GAs were run separately on each block of DRAGON descriptors (18 blocks in total). The number of independent runs (from which the selection frequencies were calculated) was set to 100. Then, only the descriptors with the largest frequencies of selection from each block were retained and merged together to form a pool of 208 candidate good descriptors, which was input to GA again. Results based on 100 independent runs showed that only one molecular descriptor, i.e. *MLOGP* [61,62], had a considerably larger frequency of selection than the others. Models based on all the possible combinations of the 15 most frequently selected descriptors were calculated with the constraint that *MLOGP* be always included, since it was demonstrated to be relevant for toxicity modelling. The obtained models were then judged taking into consideration both their predictive power and their complexity, which was defined by the number of included descriptors and their ease of interpretation. This procedure resulted in a *k*NN model (*k* equal to six) based on six molecular descriptors (*MLOGP*, *CIC0*, *NdssC*, *NdsCH*, *SM1_Dz(Z)*, *GATS1i*).

3.1.2 Definition of the applicability domain

After selecting the optimal set of descriptors and the number of nearest neighbours (*k* equal to six), an analysis was carried out in order to define the optimal value of the distance threshold, which is used in the second step of the assessment of the AD, as explained previously.
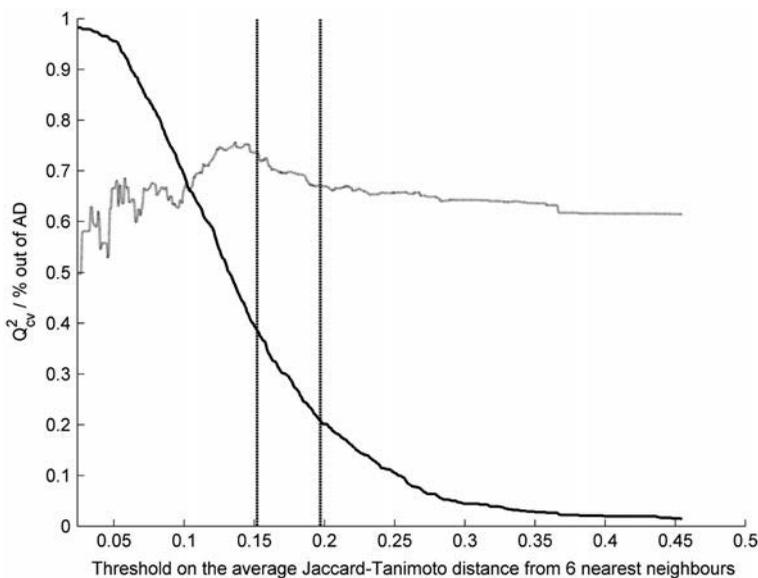
Figure 1. Q^2_{cv} (dotted line) and percentage of molecules out of AD (solid line) as a function of the threshold value on the average Jaccard–Tanimoto distance from six neighbours. The two vertical lines are the selected distance thresholds.

Since only compounds with average distance from the six nearest neighbours lower than the threshold are considered inside the AD, it is evident that a low distance threshold corresponds to a strict AD criterion because it demands all six neighbours to be very similar (low distance) to the molecule to be predicted. Figure 1 shows the Q^2_{cv} values and the percentage of molecules outside the AD of the model, as a function of distance threshold value. From Figure 1 it can be noted that very low distance threshold values correspond to large percentages of molecules out of AD (as expected), but surprisingly the Q^2_{cv} values are not high. The performance of the model (Q^2_{cv}) rises with the distance threshold value up to a maximum and then decreases smoothly, whereas the percentage of chemicals out of AD always decreases more steeply. Two distance thresholds were chosen, corresponding to values of 0.152 and 0.197, which will be referred to as the 'Strict' and 'Soft' distance thresholds, respectively. The 'Strict' distance threshold (0.152) is located nearby the maximum performance of the model.

The chemicals considered inside the AD with the 'Strict' distance threshold are a subset of those inside the AD with the 'Soft' threshold. This subset is characterized by lower distance (higher similarity) between the compound to be predicted and its nearest neighbours and higher prediction accuracy (as shown in the following paragraph). Thus, with the 'Strict' distance threshold, the percentage of compounds out of AD is larger, but the model performs better. It should be stressed that the predictions provided by the model with the 'Soft' and 'Strict' distance thresholds are the same. The 'Soft' and 'Strict' distance thresholds can be used to fulfil different needs. The 'Strict' distance threshold is intended to be used when the risk of using a potentially low-accuracy prediction is high and it is therefore preferable to have no prediction at all. On the other hand, the 'Soft' distance threshold can be applied to situations where it is more desirable to have a toxicity estimate (even if potentially less

accurate) rather than having none, e.g. for high-throughput screening or for regulatory applications in the framework of a weight of evidence approach. In any case, the user can tune the value of the distance threshold to fulfil personal needs.

3.1.3 Model statistics and analysis of the residuals

The statistics of the model in fitting, cross-validation (both five-fold and leave-more-out) and external validation corresponding to both chosen threshold values ('Strict' and 'Soft' distance thresholds) are shown in Table 3. The model reached satisfactory performance both in internal and external validation, especially considering the size of the dataset (to our knowledge the largest analysed so far). Moreover, the performances obtained with both distance thresholds ('Soft' and 'Strict') in fitting, internal and external validation, are balanced, which should indicate absence of overfitting, a common pathology when dealing with high-dimensional data. As expected, the 'Soft' threshold has lower (yet still satisfactory) statistics compared with the 'Strict' threshold, but the percentage of molecules outside the AD is nearly half. If only the bounding box approach is used to evaluate the AD, i.e. no distance threshold, the statistics would be significantly lower (r^2, Q^2_{cv} and Q^2_{ext} equal to 0.62, 0.61 and 0.61, respectively), as highlighted in Table 3 ('No' threshold case), thus indicating that a polyhedral-like representation of the training set space is not appropriate. On the contrary, the similarity-based evaluation of the AD seems effective in identifying unreliable predictions when used in the framework of a kNN model.

Figures 2(a) and 2(b) show the calculated and predicted *versus* experimental toxicity values for the training set (fitting) and the test set, respectively. Compounds inside the AD with the 'Strict' distance threshold are represented by star symbols. The vast majority of these compounds are accurately predicted. The few associated with poor toxicity estimates are affected by both underestimation and overestimation. If the 'Soft' distance threshold is used, also molecules represented by the addition symbol are regarded inside the AD. It is apparent that some of these molecules are not very well predicted (again both overestimated and underestimated). This behaviour seems more evident in the training set (Figure 2(a)). Nevertheless, many of the molecules associated with the worst predictions are regarded as outside the AD with both the 'Soft' and 'Strict' distance threshold. The model seems not to have a noticeable bias.

Figures 2(c) and 2(d) report the standardized residuals *versus* the average Jaccard–Tanimoto distance from the six nearest neighbours for the training and test sets, respectively. It can be seen that the residuals tend to increase when increasing the average distance, i.e. molecules with larger average distances from their neighbours (less similar) are associated with less accurate predictions. This fact is the experimental justification of the ideas behind the introduction of the similarity-based AD approach: the more structurally similar (low distance) a molecule to its nearest neighbours, the more similar their responses and, therefore, the more reliable the prediction. Nevertheless, two opposite behaviours that do not follow these assumptions can be detected. On one side, there are molecules associated with large average distances (out of AD) that are instead well predicted. This could be the case of structural cliffs, i.e. molecules with relatively different structures that possess instead similar activities. On the other side, there are molecules with relatively low average distances, i.e. similar, that are poorly predicted. This could instead be the case of activity cliffs, i.e. molecules with similar structures but different activities. In this regard, there are two chemicals in the training set that are inside the AD of the 'Strict' distance threshold (star symbols) but whose predictions are very poor (absolute standardized residuals larger than three). This situation occurs for an

Table 3. Statistics of the model in fitting, cross-validation and external validation. Training set: 726 molecules; test set: 182 molecules.

Threshold	k	thr	Fitting			5-fold CV			Leave-more-out			Test set		
			r^2	RMSEC	% out[a]	Q^2_{cv}	RMSEC	% out[a]	Q^2_{cv}	RMSEC	% out[a]	Q^2_{ext}	RMSEP	% out[a]
'No'[b]	6	–	0.62	0.879	0	0.61	0.878	1	0.61	0.890	1	0.61	0.888	1
'Soft'	6	0.197	0.69	0.752	17	0.67	0.755	21	0.72	0.756	20	0.73	0.745	15
'Strict'	6	0.152	0.73	0.657	33	0.74	0.641	38	0.79	0.654	39	0.77	0.682	27

[a]Percentage of molecules outside the applicability domain of the model; [b]only bounding box assessment of the applicability domain.
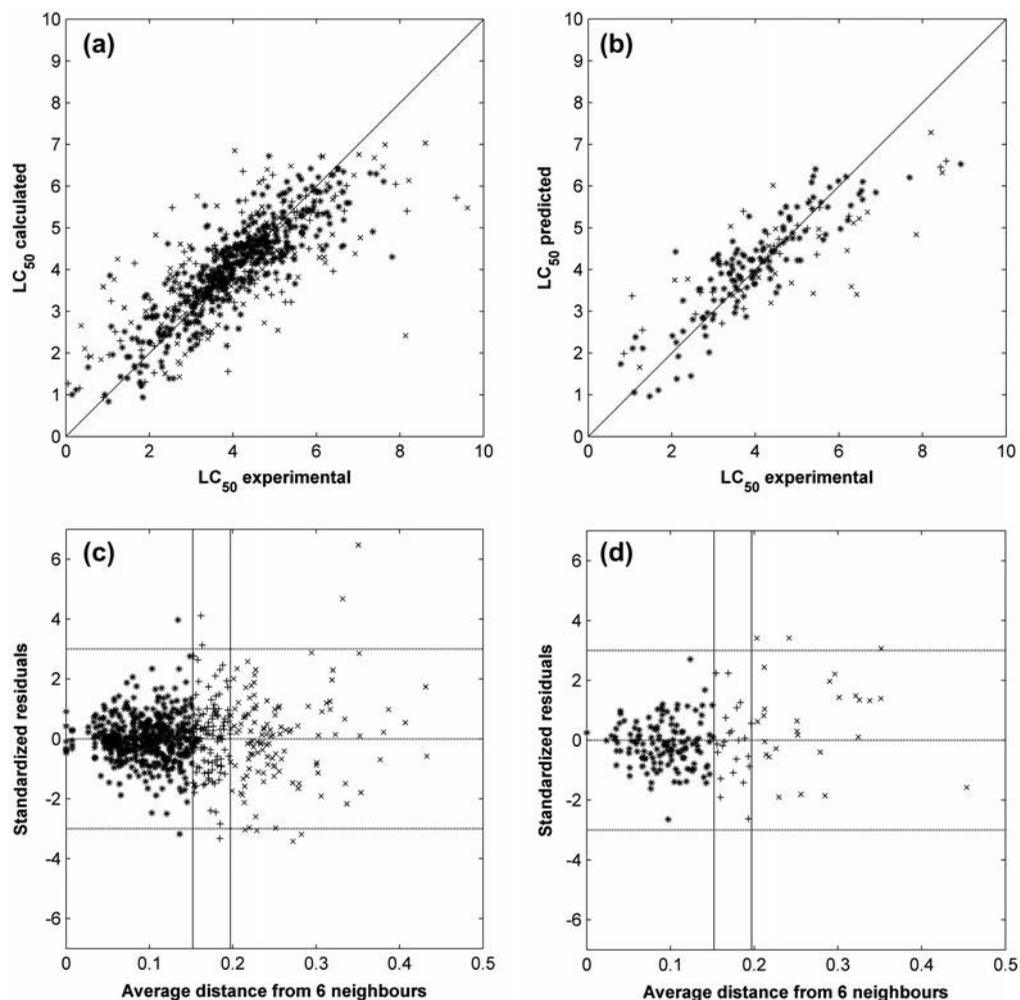
Figure 2.   Results with the 'Strict' and 'Soft' distance thresholds: a) calculated *vs.* experimental LC_{50} for training set; b) predicted *vs.* experimental LC_{50} values for test set; standardized residuals *vs.* average Jaccard-Tanimoto distance from six neighbours for training set (c) and test set (d). Multiplication symbols: molecules out of AD with both distance thresholds; star symbols: molecules in AD with both distance thresholds; addition symbols: molecules in AD of 'Soft' and out of 'Strict' distance thresholds. Vertical lines in Figures 2(c) and 2(d) correspond to the distance threshold values. (LC_{50} values are reported as $-Log_{10}(mol/L)$).

additional three compounds in the training set with the 'Soft' distance threshold, while none of the test set molecules inside the AD with either the 'Soft' or 'Strict' distance threshold has absolute standardized residuals greater than three. These five training chemicals with absolute standardized residuals higher than three are provided in Table 4 together with additional details. Two of these compounds are pyrethroid insecticides and their toxicity is largely underestimated. *N*-vinylcarbazole is predicted less toxic than it is, as well, and had also been detected as an outlier elsewhere [17]. The comparison of the experimental toxicity of these three chemicals with the baseline toxicity calculated by means of the equation reported in

Schüürmann et al. [39] indicates that they exert excess toxicity. The MoA of *N*-vinylcarbazole and Flucythrinate was determined to be electrophile/pro-electrophile and CNS seizure agent, respectively [4]. The LC_{50} ($-Log_{10}$(mol/L)) of the remaining two compounds (3,3-dimethyl-glutaric acid and bis(2-ethylhexyl) phthalate) is instead overestimated. Also, their calculated baseline toxicities are greater than the experimental values. From these results, it can be hypothesized that the model has a tendency to underestimate the toxicity of pyrethroids. An investigation, indeed, highlighted that all pyrethroids in the dataset are underestimated. As aforementioned, the AD assessment approach is based on the idea that similar molecules possess similar toxicities: the predicted toxicity of a test molecule with enough similar neighbours is assumed reliable. Unfortunately, presumed reliable predictions are not always accurate, as for the five analysed molecules, because the reliability is evaluated only on the chemical structures, but the accuracy depends on the model performance, which is not uniform in the chemical structural domain.

3.1.4 *Performance on individual functional groups*

An additional analysis of the performance of the model with the 'Strict' threshold value was carried on individual functional groups. For the list of moieties included in DRAGON software, the root mean square error (RMSE) was calculated from the results in five-fold cross-validation only on the molecules that feature a specific functional group. To this end, the number of occurrences of a functional group within one chemical obtained from DRAGON was transformed into a binary value indicating presence or absence. It should be highlighted that combinations of functional groups were not considered. The results are displayed in Figure 3, which is a bubble plot where the size of each bubble is proportional to the number of molecules that possess each functional group; the *y*-axis is the RMSE between experimental and predicted responses and the *x*-axis ranks the functional groups from the least to the most represented. Figure 3(a) shows that the error of the model on infrequent moieties (left-side) varies significantly from very low, e.g. for (thio/dithio) sulfonates ('*nSO3*'), aliphatic oximes ('*nRCNO*') and aliphatic compounds with secondary or tertiary sp^2 carbon atoms ('*nR=Ct*' and '*nR=Cs*'), to very high, e.g. pyrrolidines ('*nPyrrolidines*'), pyrroles ('*nPyrroles*'), (thio/dithio) sulfonic acids ('*nSO2OH*') and dihalogenated sp^3 carbon atoms ('*nCR2X2*'). However, drawing conclusions from the results on such infrequent functional

Table 4. Details of the five training chemicals with absolute standardized residuals greater than 3 inside the AD of the 'Strict' or 'Soft' distance threshold. All LC_{50} values are expressed as $-Log_{10}$(mol/L).

Name	CAS-RN	pLC_{50} exp^a	pLC_{50} $pred^b$	$Class^c$	pLC_{50} $narc^d$	AD^e
3,3-Dimethylglutaric acid[f]	4839-46-7	1.055	3.859	dicarboxylic acid	2.079[k]	'Strict''Soft'
N-Vinylcarbazole[g]	1484-13-5	7.809	4.305	carbazole	4.228[k]	'Strict''Soft'
Bis(2-ethylhexyl) phthalate[h]	117-81-7	2.548	5.485	phtalate	7.870	'Soft'
Fenpropathrin[i]	39515-41-8	8.169	5.406	pyrethroid	6.255	'Soft'
Flucythrinate[j]	70124-77-5	9.354	5.722	pyrethroid	6.680	'Soft'

[a]experimental LC_{50} value; [b]predicted LC_{50} value; [c]chemical class; [d]baseline toxicity according to [39]; [e]indicates the distance thresholds for which the chemical is inside the applicability domain; [f]OC(=O)CC(CC(=O)O)(C)C; [g]C=Cn1c2ccccc2c2c1cccc2; [h]CCCCC(COC(=O)c1ccccc1C(=O)OCC(CCCC)CC)CC; [i]N#CC(c1cccc(c1)Oc1ccccc1)OC(=O)C1C(C1(C)C)(C)C; [j]N#CC(c1cccc(cc1)Oc1ccccc1)OC(=O)C(c1ccc(cc1)OC(F)F)C(C)C; [k]*MLOGP* used because no experimental log P was found.

Figure 3.   Bubble plot of the performance of the model with the 'Strict' distance threshold in five-fold cross-validation on individual functional groups. Each bubble corresponds to a functional group and its size is proportional to the number of molecules possessing that moiety. Y-axis: RMSE between experimental and predicted responses; x-axis: ranking of functional groups from the least to the most represented. Figure 3(b) is a zoom of Figure 3(a) for the most common functional groups highlighted by the black square.

groups would be misleading because the small number of molecules having these moieties does not make a reliable statistical sample. The RMSE values converge along the *x*-axis to values close to the average RMSE over the entire dataset (0.641) because more represented functional groups are considered. Figure 3(b) allows to make some considerations regarding the most common functional groups. The majority of moieties associated with large RMSE comprise compounds with hydroxyl groups or sp^3 carbon atoms: in particular, secondary and tertiary alcohols ('*nOHs*' and '*nOHt*'), tertiary and quaternary sp^3 carbon atoms ('*nCt*' and '*nCq*'). '*nROH*' accounts in general for hydroxyl groups including secondary and tertiary alcohols. The performance on primary alcohols ('*nOHp*') is instead much better: in fact, a trend of increasing RMSE values for the sequence primary < secondary < tertiary alcohols ('*nOHp*' – '*nOHs*' – '*nOHt*') is evident. The same applies also to the sequence primary < secondary < tertiary < quaternary sp^3 carbon atoms ('*nCp*' – '*nCs*' – '*nCt*' – '*nCq*'). In some cases, the performance of the model on the aromatic form is worse than on the corresponding aliphatic, e.g. primary amines ('*nArNH2*' *versus* '*nRNH2*') and ethers ('*nArOR*' *versus* '*nROR*'). Other functional groups show instead the opposite trend, e.g. alcohols ('*nArOH*' *versus* '*nROH*') and esters ('*nArCOOR*' *versus* '*nRCOOR*'). The error associated with molecules with conjugated π systems ('*nCconj*') is greater than that associated with aromatic compounds ('*nCAr*'). The same applies to thioethers ('*nRSR*') that have a greater RMSE than ethers ('*nROR*'). The performance on aromatic tertiary amines ('*nArNR2*') is slightly worse than on primary amines ('*nArNH2*'). The same consideration applies also to hydrogen bond donors and acceptors ('*nHDon*' and '*nHAcc*', respectively). No difference seems to exist between substituted and unsubstituted benzene C atoms ('*nCb-*' and '*nCbh*', respectively), probably because they are often present in the same compound. It should be highlighted that the RMSEs of 'broad functional groups' that include in their definition more specific ones

(e.g. *nHAcc*, *nHDon*) are an average of the RMSEs of the functional groups converging into them. The fact that the frequent moieties associated with the largest RMSE values correspond to higher degrees of substitution (secondary and tertiary alcohols, tertiary and quaternary sp^3 carbon atoms) could indicate that the model tends to perform less well on chemicals featuring branches.

The aim of this analysis is to provide users with additional information regarding the performance of the model that can be considered to assess the reliability of each prediction. Therefore, the comments outlined above should be taken as indications. Detailed information regarding the performance of the model on individual functional groups in fitting, five-fold cross-validation and external validation with both the 'Soft' and 'Strict' distance threshold is provided in the supplementary material, available via the Supplementary Content tab on the article's online page at http://dx.doi.org/10.1080/1062936X.2015.1018938.

3.2 *Interpretation of model descriptors*

The proposed *k*NN model is based on six molecular descriptors (*MLOGP*, *CIC0*, *NdssC*, *NdsCH*, *SM1_Dz(Z)*, *GATS1i*) selected by means of GAs. Here we give a description of model descriptors, together with a coarse-grained interpretation of their relationship with fish toxicity.

MLOGP is the octanol-water partitioning coefficient (log *P*) calculated by means of the Moriguchi model, which consists in a regression equation based on 13 structural parameters [61,62]. The log P is a widely accepted estimate of the lipophilicity of organic compounds, which is considered the driving force of narcosis.

CIC0 belongs to the set of indices of neighbourhood symmetry [63]. These indices derive from a partitioning of the vertices of the hydrogen-filled molecular graph into equivalency classes. According to this scheme, two vertices (i.e. atoms) are equivalent if they represent the same chemical element and their neighbourhood of order *k*, i.e. the bonded atoms up to topological distance equal to *k*, is identical. In particular, *CIC0* is the complementary information index of order zero, i.e. only graph vertices (i.e. atoms) are considered: it is calculated as deviation of the information content of order zero (*IC0*) from its maximum value. The value of this index decreases with increasing number of different chemical elements present in a molecule: thus, it can be said to encode information regarding heteroatoms, where only the number of different elements is accounted for and not the number of occurrences of each element. Information content indices, including *CIC* indices, have already been proven useful in biological correlations in general [64], and more specifically for modelling acute toxicity to *Pimephales promelas* of alcohols [11] and esters [12].

NdssC and *NdsCH* belong to the atom-type E-state counts, a simplification of the Kier–Hall atom-type E-state indices [65,66] in that they only count the number of occurrences of given atom-types [67]. In particular, *NdssC* and *NdsCH* count the number of unsaturated sp^2 carbon atoms of the type =C< and =CH-, respectively. Hence, these two descriptors account for a variety of functional groups with double bonds, e.g. (thio)ketones and aldehydes, imines, carboxylic and carbamic acids, amides, esters and carbon–carbon double bonds. A common characteristic is an electrophilic carbon atom that can react with nucleophiles, giving substitution or addition reactions. At the borderline we find carbon–carbon double bonds because they can give addition reactions in which the double bond first acts as nucleophile, generating an intermediate electrophilic carbocation, which is then attacked by another nucleophile (e.g. hydrohalogenation). The hypothesis that these descriptors encode information about the electrophilic characteristics of chemicals is corroborated by the presence of several nucleophiles in living organisms and the fact that electrophiles are among the most common

toxicants. A further indication of the appropriateness of *NdsCH* can be found in In et al. [20], where the corresponding index calculated as sum of the E-states, *SdsCH*, had been used in a decision tree to classify reactive chemicals from narcotics. *NdsCH* and *NdssC* are, indeed, related to the electrophiles/pro-electrophiles MoA.

SM1_Dz(Z) belongs to a set of descriptors calculated from 2D matrices derived from the molecular graph (2D matrix-based descriptors) [2]. In particular, *SM1_Dz(Z)* is the spectral moment of order 1 calculated from the Barysz matrix weighted by the atomic number [68]. In other words, this descriptor is the sum of the eigenvalues of the Barysz matrix, whose elements take into account information on both the bond order and the atomic number. Also this descriptor seems to account for heteroatoms. The largest observed correlation is, in fact, with the number of heteroatoms ($\rho = 0.86$): the molecules with the lowest *SM1_Dz(Z)* values are entirely constituted by carbon atoms (both aromatic and not) while the largest values are taken on by highly fluorinated and chlorinated compounds and, more in general, compounds with several heteroatoms.

GATS1i is a 2D Geary autocorrelation descriptor [2]. Geary coefficients vary from zero to infinite and assume low values for positive autocorrelations and vice versa. In particular, *GATS1i* considers the ionization potential of atom pairs at topological distance equal to one, i.e. bonded atoms. *GATS1i* tends to have low values for molecules with pairs of bonded atoms with comparable ionization potentials, such as CC and CBr. Consequently, *GATS1i* tends to have low values for (a) molecules with several carbon–carbon bonds, as highlighted by the relatively large coefficient of correlation with the percentage of carbon atoms ($\rho = -0.79$), and (b) molecules with bromine and iodine. In addition, the distribution of *GATS1i* acquires lower values for aromatic compounds.

Since the model is based on a *k*NN approach, there are no coefficients to quantify the contribution of each descriptor in the calculation of toxicity. The analysis of how descriptors relate to toxicity was carried out by means of principal component analysis [69]. The score and loading plots of the training set are shown in Figure 4. An evident trend in the toxicity values emerges in Figure 4(a): toxicity increases mainly from right to left along PC1; a minor increase is also observed from top to bottom along PC2. Two descriptors have high loadings on PC1, namely *MLOGP* and *GATS1i*, and two on PC2, i.e. *CIC0* and *SM1_Dz(Z)* (Figure 4(b)). The score plot in Figure 4(a) shows that molecules with larger *MLOGP*, i.e.
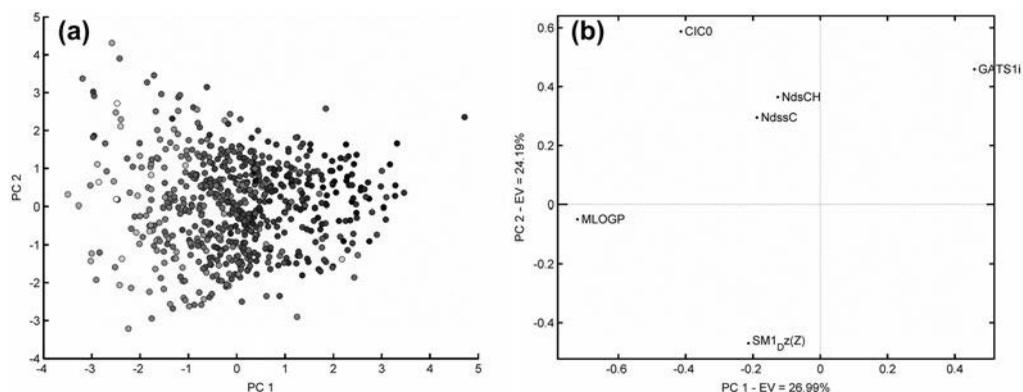
Figure 4.   Principal component analysis of the training set. a) score plot with molecules coloured based on the toxicity values. White: high toxicity; black: low toxicity. b) loading plot of model descriptors.

more lipophilic, tend to have greater toxicity. It is widely known that baseline toxicity is strictly connected with the partitioning of xenobiotics between water and organism, which in turn relies on lipophilicity. The octanol-water partition coefficient has been widely used as an estimate of lipophilicity and is present in most equations to calculate narcosis-level toxicity. The second contribution is provided by *GATS1i*. In this case, toxicity increases with decreasing values of the descriptor. It was previously noticed that low values of *GATS1i* are taken on by molecules with high carbon content, among which especially aromatic compounds are found. This information seems also to be related to lipophilicity. Thus, *GATS1i* seems to account for similar information as log P, at least in regard of the main observed trend along PC1, which is related to lipophilicity. Toxicity also slightly increases moving downwards along PC2. As aforementioned, *CIC0* and *SM1_Dz(Z)* account for the presence of heteroatoms, with which *CIC0* and *SM1_Dz(Z)* have an inverse and direct correlation, respectively. From Figure 4(a) it can be seen that higher toxicity is possessed by molecules with lower *CIC0* and larger *SM1_Dz(Z)* values, i.e. more heteroatoms. This trend might be due to specific reactions, which can vary with the MoA. The remaining two descriptors, *NdssC* and *NdsCH*, have low loadings on PC1 and relatively low ones on PC2. As aforementioned, these descriptors account for the presence of specific electrophilic functional groups: the role of *NdssC* and *NdsCH* is supposed to be limited to finding similar neighbours for compounds having such moieties, i.e. electrophiles/pro-electrophiles.

3.3 *Comparison with existing models*

Published models to predict the LC_{50} towards the fathead minnow developed from large heterogeneous datasets have been based on a variety of modelling methods and descriptors (Table 1). The performance of global models in cross-validation ranged from values of Q^2_{cv} equal to 0.46 [22] up to 0.85 [29] with MLR, while SVR gave the best results in external validation (Q^2_{ext} equal to 0.80 [31]). The largest statistics in cross-validation (Q^2_{cv} equal to 0.87) were obtained by a similarity-based assessment of the AD in a model that combined global and local techniques [39]. 'Local' MLR models calibrated on individual clusters of chemicals gave higher statistics in internal validation [37,38]. However, these models lacked an external validation to test the whole procedure of clustering and toxicity prediction. The following detailed analysis will focus on literature models calibrated on the largest datasets for the sake of comparison with our model.

Niculescu et al. achieved high statistics in fitting and external validation (Q^2_{ext} equal to 0.78) on the second largest dataset by means of NN [35]. The high statistics obtained on all test chemicals indicate high accuracy, but the model lacks an approach to estimate the AD in which this accuracy is granted.

The third largest dataset was modelled in the T.E.S.T. and VEGA software [21,30]. Good predictivity was provided by the T.E.S.T. *consensus* model (Q^2_{ext} equal to 0.73) and the MLR VEGA model with a narrow AD (Q^2_{ext} equal to 0.69).

The model of Schüürmann et al. combined the concept of baseline toxicity with read-across to evaluate the toxicity enhancement [39]. High statistics in fitting and leave-one-out internal validation (Q^2_{cv} equal to 0.87) were obtained by a strict AD criterion, but no external validation was carried out.

The remaining QSAR models were calibrated on smaller datasets, often using the sole MED-Duluth database. Some authors used complex regression methods, such as NN and SVR.

The model proposed in this study was calibrated on a dataset that, to our knowledge, is the largest published so far: 908 compounds. This implied a higher structural diversity and therefore, presumably, presented additional challenges for modelling. Data were collected from different sources and presented high variability for the same chemical. This aspect rendered the calibration of QSAR models more difficult compared with using data measured in the same laboratory. The performance of the model is comparable with those of models in the literature, especially regarding the predictivity on the test set. In fact, the highest accuracy in prediction (Q^2_{ext} equal to 0.80 [31]) is similar to that achieved by our model with the 'Strict' criterion on a larger test set (Q^2_{ext} equal to 0.77). The model is based on a reduced number of descriptors (six), in contrast to some literature models [22,33,35,36]. In addition, the six descriptors are derived from the simple 2D structure. Some published models were instead based on 3D and quantum-chemical descriptors [17,18,23,25–28,31–33,36,37] that required geometry optimization, which can be time-consuming and might also limit the future application of the model due to inconsistencies with the generation of 3D structures. The model is also built with a simple *k*NN algorithm based on local similarities. This aspect has two beneficial effects: from the modelling viewpoint, it can handle non-linearity and is supposed to overcome the issue related to the different modes of action because only the local neighbourhood participates in the prediction; from the regulatory application viewpoint, simple algorithms are more transparent and therefore provide increased confidence in their use. In contrast, some literature models were based on more complex algorithms and strategies, such as SVR [31], NN [20,22,23,26,32–36] or introduction of a preliminary classification/clustering step [20,21,36–38]. In addition, the model implements a systematic AD assessment, which is lacking in several literature models. Eventually, in compliance with OECD principle four, the model was thoroughly validated by means of appropriate techniques (five-fold and leave-more-out cross-validation and external validation). This should assure that the reported statistics are reasonably valid for real applications of the model. On the other hand, the predictive power of some published models was assessed by means of less strict procedures, e.g. lack of an external validation [17,22,27,37–39] or internal validation by means of leave-one-out strategy [17,19,26,37,39], a technique that was reported to give optimistic results on large datasets [70,71]. Considering all these points, the model presented in this study may be considered satisfactory.

3.4 *Application example*

An example of the application of the model to external molecules is provided in this section. The results presented here do not constitute a validation of the model, but only serve the purpose of showing the information provided in the output and the further analyses that can be undertaken by the user in order to evaluate the reliability of each prediction. The dataset of the model published in Russom et al. [4] for predicting the MoA was retrieved from its implementation in the ChemProp software [60]. Nine molecules were unambiguously identified as being external to the dataset of the model and were therefore submitted for prediction. Five molecules, namely tetrabutyltin, chloroform, chloromethyl styrene, dichloromethane and iodoform, were outside the AD of the model with the 'Strict' distance threshold because the average distance from the six nearest neighbours was greater than the distance threshold (0.152). The predictions of the LC_{50} ($-Log_{10}$(mol/L)), for the four molecules inside the AD, namely tetraethyltin, di-*n*-butylisophthalate, 4,9-dithiadodecane and *p*-chlorophenyl-*o*-nitrophenyl ether, were equal to 4.41, 5.28, 4.98 and 5.44, respectively, *versus* experimental values equal to 7.33, 5.49, 4.84, and 5.11, respectively. It is evident that the predictions for

Table 5. Example of the output of the model for di-*n*-isobutylphtalate. All LC$_{50}$ values are expressed as $-$Log$_{10}$(mol/L).

Name	CAS-RN	Yexp[a]	Ypred[b]	Average distance[c]	Distance threshold	Applicability domain
di-*n*-butylisophthalate	3126-90-7	5.49	5.28	0.057	0.152	in AD
Identified functional groups						
Functional Group	RMSEP[d]	Functional Group	RMSEP[d]			
primary C (sp^3)	0.713	secondary C (sp^3)	0.785			
aromatic C	0.660	unsubstituted benzene C	0.659			
substituted benzene C	0.659	non-aromatic conjugated C	1.081			
aromatic ester	0.439	H-bond acceptor	0.714			

NEIGHBOUR 1

Name	di-*n*-butyl terephtalate
CAS-RN	1962-75-0
Yexp[a]	5.67
Std. dev[e]	0.010
No. data[f]	2
Ycalc[g]	5.03
Distance[h]	0.000
Ref.[i]	[40][41]

NEIGHBOUR 2

Name	diisobutyl phthalate
CAS-RN	84-69-5
Yexp[a]	5.49
Std. dev[e]	–
No. data[f]	1
Ycalc[g]	5.03
Distance[h]	0.000
Ref.[i]	[40]

NEIGHBOUR 3

Name	di-*n*-butyl phthalate
CAS-RN	84-74-2
Yexp[a]	5.37
Std. dev[e]	0.23
No. data[f]	8
Ycalc[g]	5.03
Distance[h]	0.000
Ref.[i]	[40][41]

NEIGHBOUR 4

Name	benzyl butyl phthalate
CAS-RN	85-68-7
Yexp[a]	5.22
Std. dev[e]	0.13
No. data[f]	2
Ycalc[g]	5.66
Distance[h]	0.086
Ref.[i]	[40]

(*Continued*)

Table 5. (*Continued*).

NEIGHBOUR 5

Name	dibutyl adipate
CAS-RN	105-99-7
Yexp[a]	4.85
Std. dev[e]	–
No. data[f]	1
Ycalc[g]	5.04
Distance[h]	0.128
Ref.[i]	[40]

NEIGHBOUR 6

Name	diethyl decanedioate
CAS-RN	110-40-7
Yexp[a]	4.98
Std. dev[e]	0.017
No. data[f]	3
Ycalc[g]	5.04
Distance[h]	0.128
Ref.[i]	[40]

[a]experimental LC$_{50}$; [b]predicted LC$_{50}$; [c]average distance from nearest neighbours; [d]RMSEP of the model on molecules in the test set that have a specific functional group; [e,f]standard deviation and number of experimental values from which the median used in the training set was calculated; [g]calculated LC$_{50}$ in fitting; [h]Jaccard–Tanimoto distance between test–training molecule; [i]source of experimental values from which the median used in the training set was calculated.

di-*n*-butylisophthalate, 4,9-dithiadodecane and *p*-chlorophenyl-*o*-nitrophenyl ether are accurate and affected by an error that is lower than the RMSEP obtained on the external validation (0.682, Table 3). The prediction of tetraethyltin is instead affected by a large error, probably due to the lack of compounds with tin atoms in the training set. Therefore, a warning about the potential low accuracy of this prediction could be derived from the composition of the training set. Chloroform, iodoform and dichloromethane are correctly regarded out of AD; in fact, only molecules with at least two carbon atoms were retained in the training set.

In order to give a practical example of the use of the model, the prediction of di-*n*-butylisophthalate is discussed and the output of the model is provided in Table 5. Table 5 gives information regarding the test molecule in terms of predicted value, average distance from the six neighbours (which was compared with the distance threshold) and the outcome of the assessment of the AD. Furthermore, the test molecule was screened against a list of functional groups. The results of the screening report the list of identified moieties and corresponding RMSEP of the model (on the test set) in order to provide insight about the performance of the model on test molecules bearing the same moieties. In addition, detailed information regarding the nearest neighbours is given. Name, CAS-RN and structure allow to specify the identity of the nearest neighbours and visually evaluate the similarity among the nearest neighbours and between each neighbour and test molecule. The median LC_{50} value of the neighbours used in the training set ('Yexp') is given, together with information about the number of experimental values used for its calculation ('No. data') and the corresponding standard deviation ('Std. dev'). Reference to the source of the data is provided in the 'Ref' field. These fields allow two analyses to be carried out: on one side, it is possible to check whether there is a large variability in the experimental values for each individual neighbour molecule (undesirable situation) and, on the other side, allows checking whether all the neighbours have similar toxicity values (desired situation) or not. The calculated LC_{50} for the neighbours ('Ycalc') is also provided in order to gain knowledge about the performance of the model in the analysed area of the chemical space. Accurate model estimates in the neighbourhood should enhance the confidence in the prediction for the test molecule as well. Finally, the Jaccard–Tanimoto distance between test molecule and each neighbour is given.

In the analysed case of di-*n*-butylisophthalate, the similarity with the neighbours is very high. In fact, all the neighbours are esters of dicarboxylic acids (four benzenedioates and two linear aliphatic dioates). The range of experimental LC_{50} values for the neighbours is limited to less than one log unit [4.85–5.67]. The standard deviations of the experimental values of the neighbours are also in general low, which enhances the confidence in their accuracy. The situation depicted in this example shows high homogeneity between test molecule and neighbours, but it can be reasonably expected that structures that are more heterogeneous could be present among the neighbours in other cases.

4. Conclusions

This study addressed the problem of predicting the acute toxicity (LC_{50} after 96 hours) of chemicals towards the fathead minnow (*Pimephales promelas*) by means of a QSAR model that can be used in the REACH regulatory framework.

Toxicity data from different sources were analysed and led to the definition of a large dataset that was modelled altogether. The *k*NN method was used to estimate the toxicity. The similarity among chemicals was evaluated from six molecular descriptors that do not require geometry optimization.

The AD was assessed by a systematic procedure that is applied to each chemical to be predicted by a two-step procedure. The comparison of the average distance from the nearest neighbours with a distance threshold seemed effective in describing the distribution of chemicals in the model space and identifying unreliable predictions. The distance threshold can be changed to tune the strictness of the AD criterion.

The model was thoroughly validated both internally and externally by means of a test set. Considering the size of the dataset, the variability of the experimental data taken from different sources, the simplicity of the algorithm and the low number of molecular descriptors, the model achieved satisfactory performance. Comparison with literature models showed the statistics to be comparable with those of models calibrated on large (yet smaller) datasets.

Evident correlations between model descriptors and toxicity were highlighted. The main trends were associated with the effect of lipophilicity and number of heteroatoms. Three descriptors were related to known modes of action. Since the dataset was modelled altogether, it was expected that descriptors related to more general trends, rather than closely to each MoA, would be selected.

Eventually, an amount of information regarding the nearest neighbours and the performance of the model on individual functional groups can be provided to the users to further assess the reliability of each prediction.

Acknowledgements

References

[1] *Regulation (EC) No 1907/2006*. 2006, pp. 1–849.
[2] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2nd ed., Vol. 41, Wiley-VCH, 2009.
[3] OECD. The Organization for Economic Development and Co-operation, *Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models*, The Organization for Economic Development and Co-operation, ENV/JM/MONO(2007)2, 2007; available at http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?doclanguage=en&cote=env/jm/mono(2007)2.
[4] C.L. Russom, S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond, *Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow* (Pimephales promelas), Environ. Toxicol. Chem. 16 (1997), pp. 948–967.
[5] L. Michielan, L. Pireddu, M. Floris, and S. Moro, *Support vector machine (SVM) as alternative tool to assign acute aquatic toxicity warning labels to chemicals*, Mol. Inform. 29 (2010), pp. 51–64.
[6] M. Nendza, M. Müller, and A. Wenzel, *Discriminating toxicant classes by mode of action: 4. Baseline and excess toxicity*, SAR QSAR Environ. Res. 25 (2014), pp. 393–405.
[7] A. Levet, C. Bordes, Y. Clément, P. Mignon, H. Chermette, P. Marote, C. Cren-Olivé, and P. Lantéri, *Quantitative structure–activity relationship to predict acute fish toxicity of organic solvents*, Chemosphere 93 (2013), pp. 1094–1103.
[8] W.D. Marzio and M.E. Saenz, *Quantitative structure–activity relationship for aromatic hydrocarbons on freshwater fish*, Ecotoxicol. Environ. Saf. 59 (2004), pp. 256–262.
[9] K. Rose and L.H. Hall, *E-state modeling of fish toxicity independent of 3D structure information*, SAR QSAR Environ. Res. 14 (2003), pp. 113–129.

[10] G. Tugcu, M.T. Saçan, M. Vracko, M. Novic, and N. Minovski, *QSTR modelling of the acute toxicity of pharmaceuticals to fish*, SAR QSAR Environ. Res. 23 (2012), pp. 297–310.

[11] S.C. Basak and V.R. Magnuson, *Molecular topology and narcosis. A quantitative structure-activity relationship (QSAR) study of alcohols using complementary information content (CIC)*, Arzneim.-ForschungDrug Res. 33 (1983), pp. 501–503.

[12] S.C. Basak, D.P. Gieschen, and V.R. Magnuson, *A quantitative correlation of the LC50 values of esters in* Pimephales promelas *using physicochemical and topological parameters*, Environ. Toxicol. Chem. 3 (1984), pp. 191–199.

[13] L.D. Newsome, D.E. Johnson, R.L. Lipnick, S.J. Broderius, and C.L. Russom, *A QSAR study of the toxicity of amines to the fathead minnow*, Sci. Total Environ. 109 (1991), pp. 537–551.

[14] G.D. Veith and S.J. Broderius, *Structure-toxicity relationships for industrial chemicals causing type (II) narcosis syndrome*, in *QSAR in Environmental Toxicology*, K.L.E. Kaiser, ed., Springer, The Netherlands, 1987, pp. 385–391.

[15] G.D. Veith, D.J. Call, and L.T. Brooke, *Structure-toxicity relationships for the fathead minnow,* Pimephales promelas*: Narcotic industrial chemicals*, Can. J. Fish. Aquat. Sci. 40 (1983), pp. 743–748.

[16] A.P. Bearden and T.W. Schultz, *Structure-activity relationships for* Pimephales *and* Tetrahymena*: A mechanism of action approach*, Environ. Toxicol. Chem. 16(1997) (1997), pp. 1311–1317.

[17] T.I. Netzeva, A.O. Aptula, E. Benfenati, M.T.D. Cronin, G. Gini, I. Lessigiarska, U. Maran, M. Vračko, and G. Schüürmann, *Description of the electronic structure of organic chemicals using semiempirical and ab initio methods for development of toxicological QSARs*, J. Chem. Inf. Model. 45 (2005), pp. 106–114.

[18] M. Pavan, T.I. Netzeva, and A.P. Worth, *Validation of a QSAR model for acute toxicity*, SAR QSAR Environ. Res. 17 (2006), pp. 147–171.

[19] K. Roy and R.N. Das, *QSTR with extended topochemical atom (ETA) indices. 15. Development of predictive models for toxicity of organic chemicals against fathead minnow using second-generation ETA indices*, SAR QSAR Environ. Res. 23 (2012), pp. 125–140.

[20] Y.-Y. In, S.-K. Lee, P.-J. Kim, and K.-T. No, *Prediction of acute toxicity to fathead minnow by local model based QSAR and global QSAR approaches*, Bull. Korean Chem. Soc. 33 (2012), pp. 613–619.

[21] T. Martin, P. Harten, R. Venkatapathy, and D. Young, *T.E.S.T. (Toxicity Estimation Software Tool)*. U.S. E.P.A., 2012; available at http://www.epa.gov/nrmrl/std/qsar/qsar.html.

[22] M. Casalegno, E. Benfenati, and G. Sello, *An automated group contribution method in predicting aquatic toxicity: The diatomic fragment approach*, Chem. Res. Toxicol. 18 (2005), pp. 740–746.

[23] D.V. Eldred, C.L. Weikel, P.C. Jurs, and K.L.E. Kaiser, *Prediction of fathead minnow acute toxicity of organic compounds from molecular structure*, Chem. Res. Toxicol. 12 (1999), pp. 670–678.

[24] M. Hewitt, M.T.D. Cronin, J.C. Madden, P.H. Rowe, C. Johnson, A. Obi, and S.J. Enoch, *Consensus QSAR Models: Do the benefits outweigh the complexity?*, J. Chem. Inf. Model. 47 (2007), pp. 1460–1468.

[25] S. Lozano, M.-P. Halm-Lemeille, A. Lepailleur, S. Rault, and R. Bureau, *Consensus QSAR related to global or MOA models: Application to acute toxicity for fish*, Mol. Inform. 29 (2010), pp. 803–813.

[26] U. Maran, S. Sild, P. Mazzatorta, M. Casalegno, E. Benfenati, and M. Romberg, *Grid computing for the estimation of toxicity: Acute toxicity on fathead minnow* (Pimephales promelas), in *Distributed, High-Performance and Grid Computing in Computational Biology*, W. Dubitzky, A. Schuster, P. Sloot, M. Schroeder, and M. Romberg, eds., Springer, Berlin, 2007, pp. 60–74

[27] M. Nendza and C.L. Russom, *QSAR modelling of the ERL-D fathead minnow acute toxicity database*, Xenobiotica 21 (1991), pp. 147–170.

[28] E. Papa, F. Villa, and P. Gramatica, *Statistically validated QSARs, based on theoretical descriptors, for modeling aquatic toxicity of organic chemicals in* Pimephales promelas *(fathead minnow)*, J. Chem. Inf. Model. 45 (2005), pp. 1256–1266.

[29] A.P. Toropova, A.A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, and G. Gini, *Coral: QSAR models for acute toxicity in fathead minnow (*Pimephales promelas), J. Comput. Chem. pp. 1218–1223.

[30] *VEGA Non-Interactive Client*. Milano, Italy: Istituto di Ricerche Farmacologiche Mario Negri; available at http://www.vega-qsar.eu/index.php.

[31] Y. Wang, M. Zheng, J. Xiao, Y. Lu, F. Wang, J. Lu, X. Luo, W. Zhu, H. Jiang, and K. Chen, *Using support vector regression coupled with the genetic algorithm for predicting acute toxicity to the fathead minnow*, SAR QSAR Environ. Res. 21 (2010), pp. 559–570.

[32] P. Mazzatorta, E. Benfenati, C.-D. Neagu, and G. Gini, *Tuning neural and fuzzy-neural networks for toxicity modeling*, J. Chem. Inf. Model. 43 (2003), pp. 513–518.

[33] P. Mazzatorta, M. Vracko, A. Jezierska, and E. Benfenati, *Modeling toxicity by using supervised kohonen neural networks*, J. Chem. Inf. Model. 43 (2003), pp. 485–492.

[34] J. Devillers, *A new strategy for using supervised artificial neural networks in QSAR*, SAR QSAR Environ. Res. 16 (2005), pp. 433–442.

[35] S.P. Niculescu, A. Atkinson, G. Hammond, and M. Lewis, *Using fragment chemistry data mining and probabilistic neural networks in screening chemicals for acute toxicity to the fathead minnow*, SAR QSAR Environ. Res. 15 (2004), pp. 293–309.

[36] G. Gini, M.V. Craciun, C. Konig, and E. Benfenati, *Combining unsupervised and supervised artificial neural networks to predict aquatic toxicity*, J. Chem. Inf. Model. 44 (2004), pp. 1897–1902.

[37] A. Colombo, E. Benfenati, M. Karelson, and U. Maran, *The proposal of architecture for chemical splitting to optimize QSAR models for aquatic toxicity*, Chemosphere 72 (2008), pp. 772–780.

[38] G. Klopman, R. Saiakhov, and H.S. Rosenkranz, *Multiple computer-automated structure evaluation study of aquatic toxicity II. Fathead minnow*, Environ. Toxicol. Chem. 19 (2000), pp. 441–447.

[39] G. Schüürmann, R.-U. Ebert, and R. Kühne, *Quantitative read-across for predicting the acute fish toxicity of organic compounds*, Environ. Sci. Technol. 45 (2011), pp. 4616–4622.

[40] US EPA. U.S. Environmental Protection Agency, *ECOTOX Database, Release 4.0*; available at http://cfpub.epa.gov/ecotox/.

[41] ECETOC. European Centre for Ecotoxicology and Toxicology Of Chemicals, *TR 091 - ECETOC Aquatic Toxicity (EAT) database*. 2003; available at http://www.ecetoc.org/technical-reports.

[42] *The OECD QSAR Toolbox for Grouping Chemicals into Categories*. Organisation for Economic Co-operation and Development, 2010; available at http://www.qsartoolbox.org/.

[43] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, *KNIME: The Konstanz Information Miner*, in *Studies in Classification, Data Analysis, and Knowledge Organization*, 2007, C. Preisach, H. Burkhardt, L. Schmidt-Thieme, R. Decker, eds., Springer Berlin Heidelberg, pp. 319–326.

[44] Royal Society of Chemistry, *ChemSpider*; available at http://www.chemspider.com.

[45] NCI/CADD Group, *Chemical Identifier Resolver*; available at http://cactus.nci.nih.gov.

[46] E.E. Bolton, Y. Wang, P.A. Thiessen, and S.H. Bryant, *PubChem: Integrated platform of small molecules and biological activities*, Annu. Rep. Comput. Chem. 4 (2008), pp. 217–241.

[47] Sigma-Aldrich Co.; available at http://www.sigmaaldrich.com.

[48] N. Nikolov, V. Grancharov, G. Stoyanova, T. Pavlov, and O. Mekenyan, *Representation of chemical information in OASIS centralized 3D database for existing chemicals*, J. Chem. Inf. Model. 46 (2006), pp. 2537–2551.

[49] MICHEM. Milano Chemometrics and QSAR Research Group, *Acute toxicity to fish dataset*; available at http://michem.disat.unimib.it/chm/download/toxicityfish.htm.

[50] DRAGON 6 *(Software for Molecular Descriptor Calculation)*. Talete srl, 2012.

[51] R. Todeschini, D. Ballabio, and V. Consonni, *Distances and other dissimilarity measures in chemometrics*, in *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Accepted for publication.

[52] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, and R. Todeschini, *Comparison of different approaches to define the applicability domain of QSAR models*, Molecules 17 (2012), pp. 4791–4810.

[53] F. Sahigara, D. Ballabio, R. Todeschini, and V. Consonni, *Defining a novel k-nearest neighbours approach to assess the applicability domain of a QSAR model for reliable predictions*, J. Cheminformatics 5 (2013), p. 27.

[54] R.P. Sheridan, B.P. Feuston, V.N. Maiorov, and S.K. Kearsley, *Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR*, J. Chem. Inf. Model. 44 (2004), pp. 1912–1928.

[55] M. Cassotti, D. Ballabio, V. Consonni, A. Mauri, I.V. Tetko, and R. Todeschini, *Prediction of acute aquatic toxicity toward* Daphnia magna *by using the GA-kNN Method*, ATLA — Altern, Lab. Anim. 42 (2014), pp. 31–41.

[56] R. Leardi and A.L. González, *Genetic algorithms applied to feature selection in PLS regression: How and when to use them*, Chemom. Intell. Lab. Syst. 41 (1998), pp. 195–207.

[57] V. Consonni, D. Ballabio, and R. Todeschini, *Comments on the definition of the Q2 parameter for QSAR validation*, J. Chem. Inf. Model. 49 (2009), pp. 1669–1678.

[58] *MATLAB*. Natick, MA, USA: MathWorks Inc., 2012.

[59] *Marvin*. ChemAxon Ltd., 2012.

[60] Chemical Properties Estimation Software System, *ChemProp*, UFZ Department of Ecological, Chemistry, Leipzig, 2013.

[61] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, and Y. Matsushita, *Simple method of calculating octanol/water partition coefficient*, Chem. Pharm. Bull. 40 (1992), pp. 127–130.

[62] I. Moriguchi, S. Hirono, I. Nakagome, and H. Hirano, *Comparison of reliability of log P values for drugs calculated by several methods*, Chem. Pharm. Bull. 42 (1994), pp. 976–978.

[63] V.R. Magnuson, D.K. Harriss, and S.C. Basak, *Topological indices based on neighborhood symmetry: Chemical and biological applications*, in *Studies in Physical and Theoretical Chemistry*, R.B. king, ed.,Elsevier, Amsterdam, The Netherlands, 1983, pp. 178–191.

[64] S.C. Basak, D.K. Harriss, and V.R. Magnuson, *Comparative study of lipophilicity versus topological molecular descriptors in biological correlations*, J. Pharm. Sci. 73 (1984), pp. 429–437.

[65] L.H. Hall and L.B. Kier, *Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information*, J. Chem. Inf. Model. 35 (1995), pp. 1039–1045.

[66] L.H. Hall, L.B. Kier, and B.B. Brown, *Molecular similarity based on novel atom-type electrotopological state indices*, J. Chem. Inf. Comput. Sci. 35 (1995), pp. 1074–1080.

[67] D. Butina, *Performance of Kier-Hall E-state descriptors in quantitative structure activity relationship (QSAR) studies of multifunctional molecules*, Molecules 9 (2004), pp. 1004–1009.

[68] M. Barysz, G. Jashari, R.S. Lall, A.K. Srivastava, and N. Trinajstic, *On the distance matrix of molecules containing heteroatoms*, in *Chemical Applications of Topology and Graph Theory*, R.B. king, ed.,Elsevier., Amsterdam, The Netherlands, 1983, pp. 222–230.

[69] I. Jolliffe, *Principal component analysis,* in *Encyclopedia of Statistics in Behavioral Science*, B. S. Everitt and D.C. Howell, eds., John Wiley & Sons, Ltd, 2005.

[70] A. Golbraikh and A. Tropsha, *Beware of q2!*, J. Mol. Graph. Model. 20 (2002), pp. 269–276.

[71] K.H. Esbensen and P. Geladi, *Principles of proper validation: Use and abuse of re-sampling for validation*, J. Chemom. 24 (2010), pp. 168–187.

Erratum

Cassotti, M., Ballabio, D., Todeschini, R., and Consonni, V. (2015) A similarity-based QSAR model for predicting acute toxicity towards the fathead minnow (*Pimephales promelas*), *SAR and QSAR in Environmental Research*, 26(1-3), pp. 217–243. http://dx.doi.org/10.1080/1062936X.2015.1018938

When the above article was published, an error was introduced into Equation (1) on page 224. The correct version is depicted below:

$$
d_{rt} = \sqrt{1 - \frac{\sum_{j=1}^{p} x_{rj}.x_{tj}}{\sum_{j=1}^{p} x_{rj}^2 + \sum_{j=1}^{p} x_{tj}^2 - \sum_{j=1}^{p} x_{rj}.x_{tj}}}
$$

$$
= \sqrt{\frac{d_{rt,euclidean}^2}{\sum_{j=1}^{p} x_{rj}^2 + \sum_{j=1}^{p} x_{tj}^2 - \sum_{j=1}^{p} x_{rj}.x_{tj}}} \quad 0 \leq d_{rt} \leq 1
$$

Taylor & Francis apologises for this error.

Appendix IV

Cassotti, M., Grisoni, F., Todeschini, R., 2014. Reshaped Sequential Replacement algorithm: an efficient approach to variable selection. *Chemometr. Intell. Lab. Syst.* 133, 136-148.

Available at:
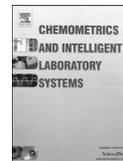
Reshaped Sequential Replacement algorithm: An efficient approach to variable selection ☆

Matteo Cassotti, Francesca Grisoni, Roberto Todeschini *

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

ARTICLE INFO

ABSTRACT

A modified version of the Sequential Replacement (SR) algorithm for variable selection is proposed, featuring modern functionalities aimed to: 1) reduce the computational time; 2) estimate the real predictivity of the model; 3) identify models suffering from pathologies. This redesigned version was called Reshaped Sequential Replacement (RSR) algorithm.

The RSR algorithm was applied to several datasets in regression and classification and was compared with the original SR method by means of a Design of Experiments (DoE). The DoE took into account the functions that affect the outcome of the search in terms of generated combinations of variables and time required for computation. The results were also compared with published models on the same datasets, taken as reference, and obtained by different variable selection methods.

This latter comparison showed that the RSR algorithm managed to find good subsets of variables on all datasets, even though the reference models were not always found. When the reference model was not found the RSR algorithm returned comparable or better subsets of variables, evaluated in cross-validation. The DoE showed that the inclusion of the additional functions allowed to obtain models with equivalent or better performances in a decreased computational time compared to the original SR method.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Current problems in different scientific fields deal with a large number of variables and are investigated by means of multivariate analysis, which involves observation and analysis of many variables at the same time. Often redundant and noisy variables, which can negatively affect the results of the analysis, are present. Problems encountered when such "bad" variables are considered include poor statistical models and difficult interpretation of the results. The presence of redundant and/or noisy variables is often the case in several applications of chemometric methods, such as analysis of spectra and Quantitative Structure–Activity Relationship (QSAR).

The problem is particularly evident in the field of QSAR since the number of molecular descriptors [1], used as **X** variables, hugely increased over time and nowadays thousands of descriptors, able to describe different aspects of a molecule, can be calculated by means of dedicated software [2,5]. However, when modeling a particular property or biological activity, it is reasonable to assume that only a small number of descriptors are actually correlated with the experimental response and are, therefore, relevant for building the mathematical model of interest. This is particularly true when linear approaches are used.

As a consequence, the selection of the optimal subset of **X** variables is a key step for the development of mathematical models. This is precisely the aim of variable selection methods, which allow to:

- improve interpretability, due to a smaller number of **X** variables (simpler models);
- neglect non-significant effects, thus reducing noise;
- increase the model predictive ability;
- speed up modeling time.

The simplest approach is the All Subset Models (ASM) method, which consists in the generation of all the possible combinations of the *p* variables, from size 1 to *p*, *p* being the total number of variables. This method – in principle – guarantees that the best subset of variables is found, but it's very computationally consuming, being the total number of combinations of *p* variables given by:

$$2^p - 1. \tag{1}$$

Often, this approach is not feasible due to the extremely steep increase in the number of models to be generated with a smooth increase in the total number of variables. If one is interested in developing simple models, i.e. models comprising a limited number *k* of variables (e.g. $1 \le k \le 20$), one can calculate all the possible combinations of

the p variables up to a maximum model size V. In this case, the total number of models t, from size 1 to V, is given by:

$$t = \sum_{k=1}^{V} \left(\frac{p!}{k!(p-k)!} \right) \leq 2^p - 1. \qquad (2)$$

The total number of models, t, is still huge when the number of variables p is large, even with a small V value. For example, if we consider a problem where $p = 130$ and $V = 20$, the total number of generated models is 2.04×10^{23} (approximately the Avogadro number!). Assuming that a computer can compute 10,000 models per second, which is a reasonable estimate for current laptops, the time required to compute all the models in this case 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. Fig. 1 depicts the sharp increase in the number of generated models (in logarithmic scale) for a modest increase in the number of variables. An estimate of the computational time is reported along the secondary vertical axis.

In order to overcome this issue throughout the years several variable selection (VS) methods have been proposed, from relatively simple ones (stepwise methods [6,7], sequential replacement algorithm [8,9]) to more recent ones that took inspiration from different scientific fields, like genetics (Genetic Algorithms [10,12]) and ethology (Particle Swarm Optimization [13], Ant Colony Optimization [14]). Furthermore, some methods able to simultaneously perform both regression and variable selection have recently been proposed (LASSO and Elastic Net [15,16]).

In this study the Sequential Replacement (SR) algorithm proposed by Miller in 1984 [8] was applied to some simulated and real datasets. To our knowledge, this method has never been applied in the field of QSAR despite its simplicity. Since the Sequential Replacement method performs an extensive search and shows a tendency to overfit data, some additional modern functionalities aimed to reshape the original algorithm taking into account computational time, model pathologies [17], chance correlation, etc. were implemented. We called this redesigned version Reshaped Sequential Replacement (RSR) in order not to generate confusion with two variants of the Sequential Replacement algorithm, namely the Replacement Method [18] (RM) and the Enhanced Replacement Method [19] (ERM). The RSR algorithm was applied to three and four datasets both in classification and regression, respectively, and it was compared with the SR method.

The models published in the scientific literature together with the datasets (and obtained using different VS methods) were taken as benchmark and used for an external comparison. This comparison allowed to see if both the SR and RSR methods could provide comparable (or better) results compared to other VS methods. The fulfillment of this criterion allowed the analysis the effects of the new modern functions of the RSR algorithm compared with the original SR method, which is the main objective of this study.

In Section 2 the Sequential Replacement algorithm and the new Reshaped Sequential Replacement method are described in detail. In Section 3 the selected datasets are presented and results are then discussed in Section 4. The RSR toolbox is briefly presented in Appendix A and in Appendix B basic notions about the variable selection methods used to obtain the reference models are given.

2. Theory

2.1. Sequential Replacement method

The basic idea of the Sequential Replacement algorithm proposed by Miller is to replace each variable included in a model of size M (with $M < p$) one at a time with each of the remaining variables and see whether a better model is obtained. This procedure differs from the All Subset Models method in that in this case not all the possible combinations of the p variables are tested, the method thus being less time consuming and meta-heuristic. The initial population is usually randomly generated, giving constraints on the number of variables (size) for each model (seed). All the variables in the model are replaced with all the remainders and the new seed is chosen only after all the variables have been replaced and the obtained models have been compared. For example, let's say the initial 4-dimensional seed comprises the variables P, A, S and T (PAST), selected from the 26 letters of the alphabet (Fig. 2). We start replacing P with all the remaining variables (except those already included in the model, i.e. A, S, T) and say model MAST is the one that shows the larger improvement of the performance, measured by a pre-defined fitness function. Similarly, the replacement of variable A, still maintaining the other three variables P, S, and T, leads to the improved model PEST. Replacing variable S, model PART is obtained and the replacement of variable T gives model PASC. Then, only the best model, say PEST, is retained and used as new initial seed to carry out the same procedure (iteration 2). The procedure goes on until no replacement leads to an improvement of the models. The replacement of the variables of model PEST leads to the final model BEST, via the substitution of variable P with variable B. Fig. 2 depicts
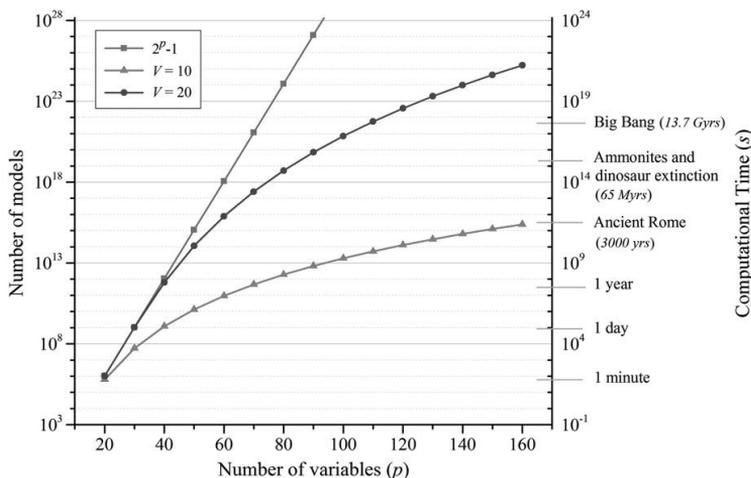
Fig. 1. Number of models vs. number of variables for an All Subset Models method with $V = 20$, $V = 10$ and $2^p - 1$, assuming a computational speed of 10,000 models per second.
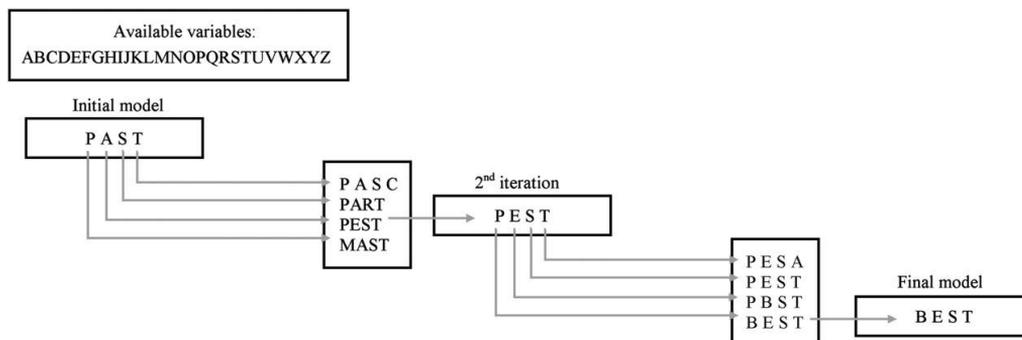
Fig. 2. Basic steps of the Sequential Replacement algorithm proposed by Miller.

the basic steps of the algorithm. Note that the replacement of variable S of model PEST does not lead to any improvement.

2.2. Reshaped Sequential Replacement algorithm

The Reshaped Sequential Replacement (RSR) algorithm adopts the Sequential Replacement method in its inner core, but implements several new functionalities aimed to: a) decrease the calculation time; b) introduce validation tools in order to estimate the prediction ability of the models; c) increase the probability to converge to the optimal model; d) identify models that suffer from pathologies, such as overfitting, chance correlation, variable redundancy and collinearity between predictors.

Fig. 3. Simplified workflow of the RSR algorithm for regression methods. The new features are highlighted red.

In Fig. 3 a scheme of the RSR workflow is provided with the new features highlighted by red boxes and bold italic text.

The introduced new features are discussed below.

2.2.1. Fitness function

The original SR algorithm uses the residuals sum of squares (*RSS*) as fitness function to be minimized in regression problems. However, the *RSS* tells nothing about the predictivity of the model. In our approach the coefficient of determination in cross-validation (Q^2_{cv}) was used as parameter to be optimized. When dealing with classification problems, the Non-Error Rate in cross-validation (NER_{cv}) is used as parameter to be maximized.

2.2.2. Tabu list (TL)

If a very large number of variables are available, it is likely that some of them are not relevant for modeling purposes. A viable option is to carry out a preliminary coarse-grained filter and exclude some variables from the analysis in the initial searching phase. Several parameters can be used to roughly estimate the importance of a variable and define exclusion criteria accordingly. The tabu list contains variables that are estimated not to be potentially relevant for the model. These variables are excluded from the calculation and are recovered only in a final stage to check whether any, in spite of the initial exclusion, can provide a further model improvement.

In this study, the predictive power of univariate models was considered as exclusion criterion for regression problems, that is, variables with a negative value of the Q^2_{cv} are stored in the tabu list, if the tabu option is activated. For classification problems, the Canonical Measure of Correlation (*CMC*) index [20] of each variable was used. The threshold value was set to 0.3, i.e. variables that have a *CMC* index lower than 0.3 are included in the tabu list. For both types of models, when the algorithm reaches convergence (i.e. all the seeds were optimized), tabu variables are recovered and will be selected in each seed only if they provide an improvement to the model higher than a pre-defined threshold (e.g., 0.01 on Q^2_{cv} in regression and on NER_{cv} in classification).

2.2.3. Roulette wheel (RW)

The roulette wheel is a selection algorithm that is biased towards high quality solutions. The roulette wheel has already been applied in Genetic Algorithms to select parent chromosomes [11]. In this way, during the cross-over step, better parents (i.e. models) are more likely to be selected. In the Reshaped Sequential Replacement algorithm, the roulette wheel was used, if the option is activated, in the initialization of the seeds by selecting variables and not models, i.e. giving each variable a probability of entering the initial models according to a pre-defined probability. In regression, the probability associated to each variable is defined on the basis of the value of the Q^2_{cv} of the
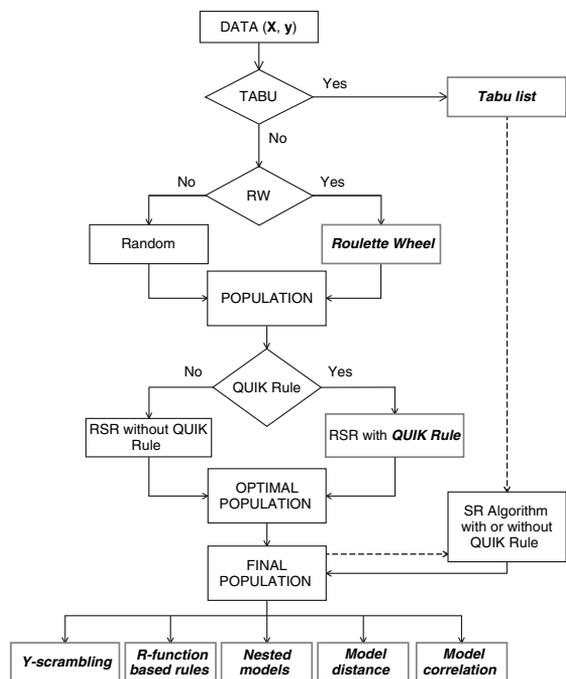
corresponding univariate model. Only variables with positive values of the Q_{cv}^2 can enter the initial population and the higher the Q_{cv}^2 is, the higher is the probability of being selected. In classification, the *CMC* index for each variable is used. Again, the higher the *CMC* index, the higher the selection probability.

2.2.4. QUIK rule (QR)

The *QUIK* rule [21] is a statistical test that allows the rejection of models with high collinearity between predictors. The rule is based on the *K* multivariate correlation index [22], which measures the global correlation of a set of variables. The *K* index is defined as:

$$K = \frac{\sum_{j=1}^{p} \left| \left(\lambda_j / \sum_j \lambda_j \right) - (1/p) \right|}{2 \cdot (p-1)/p} \qquad 0 \leq K \leq 1 \tag{3}$$

where λ_j are the eigenvalues of the correlation matrix and p is the number of variables. The basic assumption is that the total correlation of the **X** variables selected in the model plus the response **y** should be larger than the total correlation of the selected **X** variables only. Thus, the *QUIK* rule rejects all those models whose K_{xy} (correlation of **X** + **y**) is lower than K_x (correlation of only **X**).

In other words:

$$\text{if } K_{xy} - K_x < \delta K \rightarrow \text{reject the model} \tag{4}$$

where δK is a user-defined threshold (e.g. 0.01–0.05). The higher the value, the stricter the criterion and therefore the larger the number of rejected models.

The test is usually carried out before fitting and cross-validating a model, thus saving computational time and rejecting a priori models affected by relatively high collinearity between predictors, which can lead to model instability, especially when using OLS regression. The *QUIK* rule is not applied in classification.

2.2.5. Y-scrambling

Y-scrambling [23] is a statistical test commonly used to identify the potential presence of chance correlation between predictors **X** and response **y**. The test is carried out by scrambling the **y** vector in such a way that each object of the **X** matrix is no longer associated with its correct response in the **y** vector. A model is fitted and validated with this scrambled **y** vector and the statistics are calculated. The procedure is iterated many times (e.g. 100) and the average value of the statistic parameters is computed. Since the response vector has been scrambled, one expects to find poor models; if, instead, one or more models whose quality is comparable with the actual model are obtained, the actual model should be rejected due to the probable presence of chance correlation.

In this study the y-scrambling test was performed only on the final population of models and applied only to regression problems.

2.2.6. Random Error Rate (RER) test

For classification problems, a different statistical test is carried out on the final population of models. The real Error Rates (*ER*) are compared with the random classification error, i.e. the error rate obtained if the objects are randomly assigned to the classes [1]. The Random Error Rate is defined as:

$$RER\% = \frac{1}{n} \cdot \left[\sum_{g=1}^{G} \left(n - n_g \right) \cdot p_g \right] \times 100 \tag{6}$$

where n_g is the number of objects belonging to the g-th class, p_g is the g-th class a priori probability, G is the total number of classes and n is

the total number of objects. The a priori probability (p_g) for each g-th class can be calculated in two different ways:

$$P_g = 1/G \qquad P_g = n_g/n. \tag{7}$$

The first type is a uniform a priori probability: each class is given the same a priori probability; the second probability type is proportional to the number of objects of each class (n_g).

The model is accepted if the estimated error rate is smaller than the Random Error Rate:

$$RER\% - ER\% > thr \tag{8}$$

where *thr* is a predefined threshold (e.g. 0.01).

2.2.7. R function based rules

Regression models may be considered "bad" if two different situations occur: a) a model may show a redundancy in explanatory variables (excess of "good" predictors), or b) can include noisy variables, which can lead to overfitting (excess of "bad" predictors). In order to detect models suffering from these two different pathologies, two rules based on the R^P and R^N functions were introduced [17]. Both indices are defined in terms of the quantity M_j which measures the role of each variable and is calculated as:

$$M_j = \frac{R_{jy}}{R} - \frac{1}{p} \qquad -\frac{1}{p} \leq M_j \leq \frac{p-1}{p} \tag{9}$$

where R_{jy} is the absolute value of the correlation coefficient between the j-th variable and the **y** response; R is the correlation coefficient of the model; p is the number of variables of the model.

The first index, R^P, is calculated as:

$$R^P = \prod_{j=1}^{p^+} \left(1 - M_j \left(\frac{p}{p-1} \right) \right) \quad \forall M_j > 0 \quad \text{and} \quad 0 \leq R^P \leq 1 \tag{10}$$

where the product in R^P runs over the p^+ variables, giving a positive M_j. The term R^P allows to identify models with a redundancy of explanatory variables according to the rule:

$$\text{if } R^P < t^P \rightarrow \text{reject the model} \tag{11}$$

t^P being a user-defined threshold ranging from 0.01 to 0.1 depending on the data. Lower values of the threshold correspond to a stricter test.

In the original paper [17], the test of the negative M_j values was proposed on their sum, i.e.

$$R^N = \sum_{j=1}^{p^-} M_j \qquad \forall M_j < 0 \quad \text{and} \quad -1 \leq R^N \leq 0 \tag{12}$$

where the sum runs over the p^- variables giving negative M_j values; the rule was defined as:

$$\text{if } R^N < t^N \rightarrow \text{reject the model} \tag{13}$$

where t^N is a threshold value for R^N ranging from -0.01 to -0.1. However, this test on the sum seems too strict and we propose here to perform the test on each single negative M_j value, as:

$$\forall M_j < 0, \quad \exists M_j < t^N \rightarrow \text{reject the model} \tag{14}$$

i.e. the model is rejected if, at least for one variable, the negative M_j value is lower than the threshold. The t^N threshold depends on a tunable parameter ε that defines the level of noise. This parameter should be set depending on the knowledge of the **y** response noise. Higher values of ε correspond to higher values of the t^N threshold and therefore to a

stricter rule (more variables will be regarded as noisy and the corresponding models will be rejected).

The *R* function based rules are calculated only on the final population of models in regression analysis (when OLS is used).

2.2.8. Nested models

This is a function that screens the final population of models in order to find potential nested models. A model *F* is considered nested if there is a model *G* of higher size, i.e. including more variables, that comprises all the variables of *F* and has a very similar performance, i.e. their difference in prediction is smaller than a predefined threshold:

$$\left| Q^2(G) - Q^2(F) \right| \leq thr \tag{15}$$

and

$$|NER_{cv}(G) - NER_{cv}(F)| \leq thr \tag{16}$$

for regression and classification, respectively.

If this occurs, model *G* is rejected because its higher model complexity is not balanced by a higher performance. A suggested threshold for nested models is 0.05. The larger the value of the threshold, the stricter the criterion because models will more easily be identified as nested. This check is carried out in the same way both in regression and classification.

2.2.9. Model distance and correlation

The outcome of variable selection methods is often a population of models, which may show comparable predictive powers but differences in the selected variables. In order to allow for an easy comparison of the final models and determine whether models with different variables are really different in their nature, a measure of distance and correlation between models can be applied to the final models.

In this study the Canonical Measure of Distance (*CMD*) and Canonical Measure of Correlation (*CMC*) indices [20] were used to represent distances and correlations between models, respectively.

CMD and *CMC* indices are calculated as:

$$CMD_{AB} = p_A + p_B - 2 \cdot \sum_{j=1}^{M} \sqrt{\lambda_j} \qquad 0 \leq CMD_{AB} \leq (p_A + p_B) \tag{17}$$

$$CMC_{AB} = \frac{\sum_{j=1}^{M} \sqrt{\lambda_j}}{\sqrt{p_A \cdot p_B}} \qquad 0 \leq CMC_{AB} \leq 1 \tag{18}$$

where *A* and *B* are the two sets of variables being compared, p_A and p_B are the number of variables in set *A* and *B*, respectively; λ are the eigenvalues of a symmetrized cross-correlation matrix and *M* is the number of non-zero eigenvalues.

The cross-correlation matrix is an unsymmetrical rectangular matrix of size $(p_A \times p_B)$ and collects the pairwise correlations between the variables of two sets. It is formally defined as:

$$\mathbf{C}_{AB} = \left[\rho \left(\mathbf{x}_i^A, \mathbf{x}_j^B \right) \right], \qquad i = 1, ..., p_A, \ j = 1, ..., p_B \tag{19}$$

where ρ indicates the pairwise correlations between variables of the set A and variables of the set B. The cross-correlation matrix can be alternatively defined exchanging rows and columns, but the two symmetrized cross-correlation matrices $\mathbf{Q}_A$ and $\mathbf{Q}_B$

$$\mathbf{Q}_A = \mathbf{C}_{AB} \cdot \mathbf{C}_{BA} \quad \text{or} \quad \mathbf{Q}_B = \mathbf{C}_{BA} \cdot \mathbf{C}_{AB} \tag{20}$$

have the same non-zero eigenvalues.

The model distance and correlation analysis is carried out in the same fashion both in regression and classification.

3. Datasets

Four and three datasets were used in regression and classification, respectively, to test the performance of the RSR algorithm and the effect of the new added features. To this end, both simulated and real datasets were used.

3.1. Datasets for regression

Two simulated datasets with an a priori known model were built and three different levels of noise were added to the response. The two simulated datasets differ in the number of objects/variables ratio. The features in common to all simulated datasets are:

- the **X** matrix was initially generated with random values uniformly distributed in the interval [0,1];
- the response was generated according to the regression model:

$$y_i = \sum_{j=1}^{p} b_j \cdot x_{ij} + \varepsilon_i \tag{21}$$

where b_j is the coefficient of the *j*-th variable, *p* is the total number of variables ($p = 500$) and ε_i is the percentage of noise extracted from a Gaussian distribution with mean equal to 0 and standard deviation equal to 1

- the developed model is:

$$y_i = 10 \cdot x_{i1} + 9 \cdot_{i2} + 8 \cdot x_{i3} + 7 \cdot x_{i4} + 6 \cdot x_{i5} + \varepsilon_i \text{ and } b_j = 0 \quad \text{for} \quad j = 6, ..., 500. \tag{22}$$

In order to make simulated datasets more similar to real datasets, where correlated variables are often present, variables 6 to 10 correspond to variables 1 to 5 (relevant variables) with an additional 10% of noise extracted from a Gaussian distribution with mean equal to 0 and standard deviation equal to 1. The characteristics of the simulated datasets are shown in Table 1.

Two real datasets for regression, which have been published in the scientific literature together with a QSAR model, were also used. For each dataset, the published model was considered as benchmark. The reference models allowed the comparison of the SR and RSR algorithms with other variable selection methods. The main characteristics of the real datasets are presented in Table 2 together with details about the reference model and the variable selection method used to obtain it.

3.2. Datasets for classification

One real and two pseudo-real datasets were used to test the performance of the algorithm in classification problems. Pseudo-real datasets (Breast and ItaOils) are real datasets in which additional random

Table 1
Simulated datasets used in this study. In all cases the first 5 variables were used to generate the **y** response. The ID number is used in the results Section 4.1 and related tables.

ID	Objects (n)	Variables (p)	Noise (%)
1	500	500	0
2	500	500	10
3	500	500	15
4	100	500	0
5	100	500	10
6	100	500	15

Table 2
Real datasets for the regression case used in this study and corresponding benchmark models.

ID	n training	n test	p	Response	Model size[a]	cv groups	Model Q^2_{cv}	Model Q^2_{ext}	VS[b]	Reference
LC50	408	57	899	LC50 fish	4	408	0.801	0.721	GA[c]	[24]
MP	10,000	2634	150	melting p.	7	–	–	0.600	RF[d]	[25]

[a] Number of variables in benchmark model.
[b] Method of variable selection used to obtain the model.
[c] Genetic algorithms.
[d] Random Forest.

variables were included in order to have a larger pool of variables (in total 50). 41 and 42 random variables were added to Breast [26] and ItaOils [27] datasets, respectively. Since published models on Breast and ItaOils datasets only considered the original variables, Genetic Algorithms as proposed by Leardi and González [28] were run on these pseudo-real datasets in order to allow a direct comparison with another variable selection method. The results obtained with GA were considered as benchmark and are reported in Table 3. As for regression datasets, the real classification dataset (Wines) has been published in the scientific literature together with a mathematical model, which was considered as benchmark. The datasets for classification are presented in Table 3.

4. Results and discussion

In this study all regression and classification models were calibrated using Ordinary Least Squares (OLS) and K-Nearest Neighbors (K-NN) methods, respectively. In both cases, cross-validation was carried out using 5 deletion groups, with the exception of the calculations on Wines dataset where 10 cv groups were used, due to the fact that the reference model had been validated with 10 deletion groups. For all real and pseudo-real datasets a design of experiment (DoE) was carried out. Only the functions that affect the evolution in terms of generated combinations of variables and time required for computation (roulette wheel, QUIK rule and tabu list) were considered as factors for the DoE, while the functions that are applied only on the final population of models (R functions, y-scrambling, nested criterion, RER test, model distance and correlation) have always been enabled. The maximum model size was set to that of the reference model + 1 because the performance of the reference models is simply used to check that the SR and RSR methods can provide comparable (or better) results. The main objective of the study is an internal comparison of the SR and RSR methods by analyzing the effects of the new functions. The results are reported below separately for regression and classification datasets. In the subsequent tables of the manuscript the following acronyms are used: 1) TL = tabu list; 2) RW = roulette wheel; and 3) QR = QUIK rule.

4.1. Simulated datasets for regression

The simulated datasets allowed to evaluate the ability of the new features to: a) speed up modeling time and b) reject models suffering from chance correlation, overfitting and collinearity between predictors,

in conditions where the optimal (or near-optimal) solution is known. For each dataset three calculations were performed with the following settings:

1) QUIK rule, roulette wheel and tabu list disabled (this corresponds to the SR method proposed by Miller, with the exception that Q^2_{cv} is used as fitness function);
2) only QUIK rule enabled;
3) QUIK rule, roulette wheel and tabu list enabled.

In all calculations the minimum model size, maximum model size and number of models for each model size were set to 2, 7 and 10, respectively. All the results were further validated using y-scrambling, R functions and nested criterion. All the thresholds were kept at the default values reported in Table A.2. Table 4 reports a summary of the results on the simulated datasets. Only models accepted by both R functions and nested criterion are considered; only the model with the largest Q^2_{cv} is reported for each dataset and each setting.

From Table 4 it can be seen that the RSR algorithm could find the theoretical model in three out of six cases. In particular, when no noise was added to the **y** response (datasets 1 and 4) the algorithm always converged to the theoretical (and optimal model). When noise was added to the **y** response (datasets 2, 3, 5 and 6), the theoretical model may no longer be the optimal one. In fact, with noisy **y** response (datasets 2, 3, 5 and 6) in one case the RSR converged to the theoretical model (dataset 3) and in the other cases (datasets 2, 5 and 6) the algorithm found models that are equal to or better than the theoretical model in terms of Q^2_{cv}. It can also be noted that on datasets 5 and 6, the algorithm accepts also models with 7 and 6 variables, respectively, comprising random variables not used for the generation of the **y** response. In other words, the R^N function was not able to recognize these additional variables as noisy. However, in such situations where noise was added to the **y** response (10% and 15% on datasets 5 and 6, respectively) and the objects/variables ratio is very low (0.2), random variables can have significant correlations with the **y** response (in our datasets univariate Q^2_{cv} up to 0.154). It is therefore understandable why the R^N function could not detect the presence of these correlated random variables. As a consequence, the inclusion of random variables in some models can be attributed to the particularly difficult conditions of the datasets, rather than to "failure" of the function itself. In fact, in all the datasets with 500 objects, the R^N function could always reject models including noisy variables, as it can be observed by the fact that all the best models have 5 variables selected from the pool of

Table 3
Datasets for classification used in this study and corresponding benchmark models.

ID	n training	n test	p[a]	n. of classes	Model size[b]	cv groups	Model NER_{cv}	Model NER_{ext}	VS[c]	Reference
Breast	524	175	9 + 41	2	6	5	0.976	0.962	GA[d]	This work
ItaOils	572	–	8 + 42	9	7	5	0.966	–	GA[d]	This work
Wines	133	45	13	3	5	10	0.980	0.980	FS[e]	[29]

[a] Original plus random variables.
[b] Number of variables in benchmark model.
[c] Method of variable selection.
[d] Genetic algorithms.
[e] The reference model was developed using Linear Discriminant Analysis by using Forward Stepwise.

Table 4
Results on the simulated datasets. For each setting only the model with the largest Q_{cv}^2 that fulfills the R based and nested model functions is reported. In this table, 0 = disabled, 1 = enabled.

Dataset ID	QR	RW	TL	Time (s)	p best model	Q_{cv}^2 best model	Q_{cv}^2 theor. model	Theor. model[a]
1	0	0	0	2007	5	1.00	1.000	Yes
	1	0	0	1732	5	1.00		Yes
	1	1	1	70	5	1.00		Yes
2	0	0	0	1954	5	0.844	0.843	No
	1	0	0	1782	5	0.844		No
	1	1	1	300	5	0.844		No
3	0	0	0	2199	5	0.731	0.731	Yes
	1	0	0	1875	5	0.731		Yes
	1	1	1	230	5	0.731		Yes
4	0	0	0	1827	5	1.00	1.000	Yes
	1	0	0	1545	5	1.00		Yes
	1	1	1	94	5	1.00		Yes
5	0	0	0	2224	5	0.865	0.861	No
	1	0	0	2011	5	0.865		No
	1	1	1	111	7	0.879		No
6	0	0	0	1936	6	0.814	0.781	No
	1	0	0	1687	6	0.814		No
	1	1	1	155	6	0.814		No

Calculations were carried out on an *Intel i5 M460 @ 2.53 GHz* with 4 GB RAM.
  [a] Indicates whether the theoretical model was found or not.

variables 1–10, i.e. variables used to generate the response (1–5) or their corresponding ones with added noise (6–10).

Due to the way simulated datasets were generated, the rule on the R^P index is always fulfilled, i.e. no model is rejected according to this rule, because no model can possess an excess of explanatory variables.

Two additional considerations regarding the effect of tabu list, roulette wheel and *QUIK* rule can be drawn. First, one can see that in all cases when the *QUIK* rule is activated, the RSR algorithm is more parsimonious in terms of computational time. We believe this can be due to 1) rejection of models with collinear variables that may push the population far from the optimal region; and 2) avoidance of fitting and cross-validating not optimal or not promising models (those that do not fulfill the *QUIK* rule criterion), because the *QUIK* rule is performed prior to fitting and cross-validation. The further activation of tabu list and roulette wheel allows to significantly speed up the calculation, leading to a computational time that is approximately one order of magnitude lower. It should still be kept in mind that this is an optimistic case because a large number of random variables, i.e. variables that are likely to be included in the tabu list, were present. Secondly, the activation of tabu list, roulette wheel and *QUIK* rule does not affect the quality of the final models. In fact, in most of the cases, the same results were obtained with the three different settings used. Therefore it can be concluded that the activation of tabu list, roulette wheel and *QUIK* rule allows to significantly decrease the computational time, still keeping the search ability of the algorithm.

As aforementioned, in all these calculations 10 models (seeds) for each model size were used. The results showed a high degeneration of the final populations. In other words, all the seeds of the same size reached the same final model for each particular setting on all datasets. The only exception is one model of size 7 for the dataset 6 with tabu list, roulette wheel and *QUIK* rule activated, which did not converge to the model found by the other 9 seeds of size 7. Therefore, although simulated datasets represent an optimistic situation, it can be hypothesized that a large number of models (seeds) for each model size is redundant, being the exploration of the space extensive and the degeneration towards optimal solutions good.

As aforementioned, when the algorithm did not identify the theoretical model (datasets 2, 5 and 6), it still managed to provide comparable or better results. These results are analyzed more in detail below.

Dataset 2 The best models have a slightly larger Q_{cv}^2 compared to the theoretical model and the theoretical variable x_5 was

replaced by its corresponding one with additional noise (x_{10}). This replacement is not surprising since variables x_5 and x_{10} have a large correlation coefficient ($\rho = 0.98$) and show almost the same univariate Q_{cv}^2 (0.068 and 0.066, respectively).

Dataset 5 When roulette wheel and tabu list are disabled, the best models have a slightly larger Q_{cv}^2 (0.865) compared to the theoretical model (0.861) and two of the theoretical variables (x_2 and x_5) were replaced by their corresponding ones with added noise (x_7 and x_{10}). Also in these cases the pairs of variables x_2–x_7 and x_5–x_{10} are very correlated ($\rho = 0.98$) and have almost the same univariate Q_{cv}^2 (x_2: 0.078 – x_7: 0.063; x_5: 0.061 – x_{10}: 0.073). The obtained models can therefore be considered as optimal. When the roulette wheel and the tabu list are activated, the RSR algorithm finds a model with 7 variables that fulfills all the criteria. Evidently this model contains 2 variables that were not used to generate the **y** response (2 random variables). As aforementioned, due to the noise added to the **y** response (10%) and the reduced objects/variables ratio (0.2), these 2 random variables have a relatively large Q_{cv}^2, equal to 0.139 and 0.076, and they were not detected as noisy variables with the default settings of the R^N rule.

Dataset 6 The best models found have a slightly larger Q_{cv}^2 compared to the theoretical model and comprise 6 variables. In these models, the theoretical variable x_1 (univariate $Q_{cv}^2 = 0.265$) was replaced by x_6 (univariate $Q_{cv}^2 = 0.280$), i.e. the corresponding of x_1 with 10% added noise, and one noisy variable

Table 5
Results on simulated dataset 2 with *QUIK* rule, roulette wheel and tabu list activated. For R^P, R^N and nested functions 0 = rejected, 1 = accepted.

Model size	R^2	Q_{cv}^2	$\overline{Q_y^2}$	R^N	R^P	Nested	Selected variables						
							1	2	3	4	10	191	362
7	0.851	0.847	−0.023	0	1	0	x	x	x	x	x	x	x
6	0.849	0.846	−0.021	0	1	0	x	x	x	x	x		x
5	0.848	0.844	−0.017	1	1	1	x	x	x	x	x		
4	0.776	0.769	−0.015	1	1	1	x	x	x	x			
3	0.659	0.652	−0.012	1	1	1	x	x	x				
2	0.461	0.453	−0.008	1	1	1	x	x					
1	0.245	0.237	−0.008	1	1	1	x						

(univariate $Q_{cv}^2 = 0.060$) is also included. The difference in univariate Q_{cv}^2 between variables $\mathbf{x}_1$ and $\mathbf{x}_6$ is presumably the reason for the replacement, while the missed detection of the random variable by the R^N function can be explained by 1) its relatively large univariate Q_{cv}^2; and 2) the fact that by definition the R^N function allows the inclusion of one noisy variable, to a certain extent.

As an example more detailed results on dataset 2 with *QUIK* rule, roulette wheel and tabu list activated are reported in Table 5. Only the best model for each model size is reported. Table 5 reports model size, values of R^2, Q_{cv}^2 and average Q^2 of y-scrambling ($\overline{Q}_y^2$), outcome of R functions (R^N and R^P) and nested models check, and selected variables. The usefulness of the R^N index to reject models with an excess of noisy variables is evident. In fact, models with 7 and 6 variables, which include two and one random predictors (variables 191 and 362), respectively, are rejected according to the R^N rule. A confirmation of this point derives also from the nested models check. The addition of variables 191 and 362 does not lead to a significant improvement of the Q_{cv}^2. Therefore, models with 6 and 7 variables are rejected also according to the nested models check.

4.2 Real datasets for regression

4.2.1 LC50 dataset

The number of models for each model size was always set to 10 and the minimum and maximum model sizes were set to 2 and 5, respectively. The other parameters were kept at the default values (Table A.2). The results on the LC50 dataset (899 molecular descriptors) are reported in Table 6. Only the best model, based on Q_{cv}^2, and accepted by both R functions and nested criterion is reported in Table 6 for each setting. Since the reference model had been cross-validated by means of leave-one-out method, the Q_{cv}^2 was recalculated using a 5-fold cv in order to make the statistics comparable with the ones used in this study. Moreover, since the RSR algorithm uses a slightly different formula for the calculation of the Q_{ext}^2 on the external test set [30], which was proved to be more appropriate, also the Q_{ext}^2 of the reference model was recalculated. The recalculated Q_{cv}^2 and Q_{ext}^2 are equal to 0.796 and 0.684, respectively.

From the results in Table 6 it can be seen that in 50% of the cases the RSR algorithm managed to find the benchmark model. In the 4 cases in which the RSR algorithm could not find the reference model, the *QUIK* rule or the roulette wheel (or both) were activated. Two hypotheses to explain this behavior may be that: 1) the reference model is often obtained by replacements from intermediate models that do not meet the *QUIK* rule criterion and, since these intermediate models are discarded, the algorithm cannot reach the reference model; 2) since the variable *nRNH2* of the reference model has a low univariate Q_{cv}^2 (0.016), its inclusion in the initial population, when the roulette wheel is used, is not likely to happen. Another reason can be linked to the type of cross-validation, since in this study a 5-fold cross-validation

was performed, while the reference model was obtained using a leave-one-out procedure. However, it is noteworthy that the reference model does not fulfill the R^N rule with the default settings. As a consequence, even when it is found in the final population, the reference model is not considered as the best model in Table 6. A lower value of the threshold (e.g. 0.01) for the R^N rule would accept the reference model.

The models reported in Table 6 have the same performance in cross-validation (Q_{cv}^2) but a slightly lower performance on the external test set (Q_{ext}^2) compared to the reference model.

It should be remarked that the algorithm is capable to converge to optimal or near-optimal solutions independently from the settings used. In 7 out of 8 cases, in fact, the same best model was found. The most remarkable effect of tabu list, roulette wheel and *QUIK* rule is on the computational time. The activation of the 3 functions allows to obtain the same results but in just 2/3 of the time required when these functions are not activated (1084 versus 1542 s).

4.2.2 MP dataset

The calculations on the MP dataset (150 molecular descriptors) were carried out with 10 models for each model size; minimum and maximum model sizes were set to 2 and 8, respectively. The other parameters were kept at the default values (Table A.2). The results on the MP dataset are reported in Table 7. As for the results on the LC50 dataset, only the best model, based on Q_{cv}^2, and accepted by both R functions and nested criterion is reported in Table 7 for each setting. The statistics of the reference 7-dimensional model were calculated since they were not provided in the reference. The values of R^2, Q_{cv}^2 and Q_{ext}^2 are equal to 0.586, 0.585 and 0.596, respectively.

From Table 7 it can be seen that good subsets of variables were selected by the RSR algorithm, despite the reference model was never found. In fact, the best models with 8 variables show better performance than the reference model in cross-validation and in one case also on the external test set. The larger statistics are likely due to the larger model size compared to the reference model (8 versus 7 variables, respectively). The models with 5 or 6 variables have just a slightly lower performance compared to the reference model. Therefore, they can still be regarded as very good according to the Ockham's principle of parsimony. It should be highlighted that models of size larger than 6 were not accepted by the *QUIK* rule with the default threshold value; this indicates the presence of collinearity between predictors. The calculations with *QUIK* rule activated were therefore performed setting the maximum model size at 6. As a consequence a comparison of the computational time with and without *QUIK* rule would be misleading. Considering only the 4 calculations without *QUIK* rule, it can be seen that the activation of both tabu list and roulette wheel allows to save approximately 250 s compared to the SR algorithm (tabu list and roulette wheel deactivated), while still leading to satisfactory results (the model obtained with both tabu list and roulette wheel enabled shows the largest statistics on the external test set, Q_{ext}^2). The following

Table 6
Results of the DoE on the LC50 dataset. Calculations were carried out on an *Intel i5 M460 @ 2.53 GHz* with 4 GB RAM.

Settings			Time (s)	Best model (based on Q_{cv}^2)				Reference model found
TL	RW	QR		p	R^2	Q_{cv}^2	Q_{ext}^2	
0	0	0	1542	4	0.798	0.796	0.612	Yes
1	0	0	1281	4	0.798	0.796	0.612	Yes
0	1	0	1465	4	0.798	0.796	0.612	No
0	0	1	1387	4	0.798	0.796	0.612	Yes
1	1	0	1242	4	0.799	0.796	0.588	Yes
0	1	1	1206	4	0.798	0.796	0.612	No
1	0	1	1102	4	0.798	0.796	0.612	No
1	1	1	1084	4	0.798	0.796	0.612	No

Table 7
Results of the DoE on the MP dataset. Calculations were carried out on an *Intel Xeon CPU E5-2620 0 @ 2.00 GHz* with 16 GB RAM.

Settings			Time (s)	Best model (based on Q_{cv}^2)			
TL	RW	QR		p	R^2	Q_{cv}^2	Q_{ext}^2
0	0	0	2837	6	0.578	0.578	0.589
1	0	0	2711	8	0.595	0.594	0.592
0	1	0	2588	8	0.595	0.594	0.592
0	0	1	1099	5	0.561	0.560	0.563
1	1	0	2587	8	0.594	0.593	0.601
0	1	1	1018	6	0.569	0.568	0.574
1	0	1	938	5	0.561	0.560	0.563
1	1	1	562	5	0.561	0.560	0.563

Table 8
Results of the DoE on the Breast dataset. Calculations were carried out on an Intel Xeon CPU E5-2620 0 @ 2.00 GHz with 16 GB RAM.

VS method	Settings		Time (s)	Best model (based on NER_{cv})					Selected variables
	TL	RW		p	k	NER	NER_{cv}	NER_{ext}	
RSR	0	0	24,923	10	4	0.972	0.976	0.950	1 2 3 6 8 9 26 32 39 49
RSR	1	0	2040	6	7	0.970	0.976	0.962	1 2 3 6 7 8
RSR	0	1	15,279	10	4	0.965	0.974	0.962	1 2 3 5 6 7 8 25 27 40
RSR	1	1	2171	6	7	0.970	0.976	0.962	1 2 3 6 7 8
GA	–	–	1662	6	7	0.970	0.976	0.962	1 2 3 6 7 8

observations and considerations can be drawn regarding the final models when the *QUIK* rule was deactivated:

- two different models comprising 8 variables were obtained when tabu list, roulette wheel or both were activated (3 out of 4 cases). The two models with 8 variables have practically the same statistics. However, the *QUIK* rule suggests that these two models are not reliable.
- When all the functions were disabled (SR algorithm), the best model comprises only 6 variables and shows slightly poorer statistics. This model was also found in the final population when both tabu list and roulette wheel were activated.
- Two models with 6 variables, which are better (in terms of Q_{cv}^2) than the one obtained with the SR algorithm (tabu list and roulette wheel disabled), were obtained with only roulette wheel and both roulette wheel and tabu list enabled.
- When only tabu list was enabled, all models with 6 variables were rejected by R^N function.

The final population on this dataset shows a higher heterogeneity compared to the previous datasets. This is partly due to the fact that models with more than 6 variables could not fulfill the *QUIK* rule criterion, leading to final best models of different sizes.

It should be noted that the reference model was never found, but even if it was, it would not have been accepted by the R^N rule with default settings. This is due to the fact that two variables of the reference models, *ATSm1* and *nRotB* have a very low univariate Q_{cv}^2 equal to 0.011 and 0.002, respectively, and therefore are regarded as noisy. Moreover, the reference model was obtained by means of subsequent exhaustive search subset selection carried out on the most important variables resulted from a Random Forest model. These differences may also explain the reason why the reference model was never found by the RSR algorithm.

4.3. Pseudo-real datasets for classification

Two pseudo-real datasets were used to test the performance of the RSR algorithm in classification. As for regression, a DoE was carried out considering the functions that affect the outcome of the search. Since the *QUIK* rule is not applicable in classification, only tabu list and roulette wheel were considered as factors for the DoE. Genetic Algorithms were also run on these datasets in order to provide a comparison.

4.3.1. Breast dataset

The calculations on the Breast dataset (9 original variables) were carried out setting the maximum model size to 10 and the number of models for each model size to 3. The results of the RSR and GA algorithms are collected in Table 8. For each point of the DoE with the RSR algorithm, only the model with the largest NER_{cv} fulfilling both *RER* and nested criteria is reported.

Due to the presence of a large number of random variables, the probability of inclusion of noisy variables can be high. The inclusion of random variables occurs, indeed, in the two calculations with tabu list deactivated. The best models found without tabu list (size equal to 10) include 4 and 3 random variables, respectively. It should be recalled that in the case of regression problems, the R^N rule could help the

detection of models suffering from an excess of noisy variables. The R^N rule is not applicable in classification, therefore the detection of such "bad" models is harder. For this dataset, the activation of the tabu list allows to discard random variables. Indeed, when the tabu list is activated, the final population comprises only models with original variables, meaning that random variables could not provide significant improvements to the models obtained by the original variables. Thus, the tabu list seems to be a helpful tool in the detection of noisy variables especially in classification, where the R^N rule cannot be applied.

Moreover, the calculations with tabu list activated are up to more than 10 times faster, leading to a decrease of computational time from 24,923 to 2040 s. This is also partly due to the fact that the maximum model size was automatically decreased to 9, after the inclusion of random variables in the tabu list (only the 9 original variables were not included in the tabu list by the *CMC* index).

The best model found with tabu list activated coincides with the best model provided by GA. It has to be highlighted that our implementation of GA carries out a final Forward Stepwise Selection based on the frequency of selection during the runs of the GA itself. This approach is usually useful in the detection of relevant variables from sets including noisy predictors. The coinciding results seem, therefore, a further proof of the efficiency of the RSR algorithm.

4.3.2. ItaOils dataset

The calculations on the ItaOils dataset were carried out with 3 models for each model size; minimum and maximum model sizes were set to 2 and 9, respectively. The other options were kept at their default values. In order to have a comparison with other VS methods, Genetic Algorithms were also run on this pseudo-real dataset and the obtained model is reported in Table 9 together with the results provided by the RSR algorithm. The last 4 columns of Table 9 indicate by which settings each model was found.

The results from Table 9 clearly show the ability of the RSR algorithm to select optimal subsets of the original 8 variables and none of the 42 added random variables. The results also confirmed the ability of the tabu list to significantly speed up the calculation due to the preliminary exclusion of assumed not relevant variables (calculation time is not reported in Table 9). The roulette wheel alone also provided a sharp decrease in the computational time thanks to the ability of generating initial models that are closer to the optimal solutions. At the same time, the activation of tabu list and/or roulette wheel did not negatively affect the performance of the final models, as it can be seen from Table 9.

The reference model obtained with GA was selected by maximizing the NER_{cv}, i.e. the model showing the highest NER_{cv} was selected. It has to be recalled that this version of GA carries out a final Stepwise Forward Selection based on the frequency of selection of the variables during the runs of the GA itself. The best model comprises 7 variables (only original variables were selected) and has a NER_{cv} equal to 0.966. An analysis of the results in Table 9 clearly suggests that the RSR algorithm is able to find simpler models, i.e. including less variables, showing also a larger NER_{cv} compared to GA. Regarding the computational time, GA took 1/4 of the time required by the RSR algorithm when tabu list and roulette wheel are not activated, but the activation of only tabu list or both tabu list and roulette wheel made the RSR algorithm 12 times faster than GA.

Table 9
Results of the DoE on the ItaOils dataset. Calculations were carried out on an *Intel Core Quad CPU @ 2.50 GHz* with 8 GB RAM.

VS method	Size	NER_{cv}	k	1	2	3	4	5	6	7	8	TL: 0 RW: 0	TL: 1 RW: 0	TL: 0 RW: 1	TL: 1 RW: 1
RSR	5	0.968	4		x	x		x	x		x	•		•	
RSR	5	0.963	3			x	x	x		x	x	•		•	
RSR	4	0.961	4				x	x	x		x	•	•	•	•
RSR	3	0.946	6				x		x		x	•		•	
RSR	3	0.945	4				x	x			x		•		•
RSR	2	0.899	4					x			x	•	•	•	•
RSR	1	0.726	8				x					•	•	•	•
GA	7	0.966	3	x	x		x	x	x	x	x	–	–	–	–

The obtained models are comparable with other published models where different classification methods were used. As an example, Ballabio et al. [29] reported a model obtained with LDA using all the 8 original variables of the ItaOils dataset (NER_{cv} (leave-one-out) = 0.93). The RSR algorithm provided 5 models with a larger NER_{cv} (from 0.945 to 0.968 with 5 deletion groups) and also featuring a smaller number of variables (from 3 to 5, respectively). It has to be noted that Ballabio et al. split the dataset into a training set (428 samples) and a test set (144 samples), this being the likely reason for the lower statistics of the LDA model.

4.4. Wines dataset

For the calculations on the Wines dataset (13 variables), models comprising up to 6 variables were developed; the number of models for each model size was set to 3. Differently from the calculations on the other datasets, here the cross-validation was performed on 10 deletion groups. This choice was taken since the reference model had been validated by a 10-fold cross-validation. It should be highlighted that the reference model was developed using Linear Discriminant Analysis (LDA). The results are reported in Table 10; the last 4 columns indicate with which settings each model was found.

The model comprising all the 13 original variables has a NER_{cv} equal to 0.970 and a NER_{ext} on the external test set equal to 1.000. From Table 10 it can be said that also on the Wines dataset the RSR algorithm is able to find good subsets of variables. In fact, a larger NER_{cv} than that of the model with all the 13 variables is obtained only with 3 variables (x_7, x_{10} and x_{11}). The obtained models with 4 to 6 variables show also a larger NER_{cv} than that of the model with all the variables. Taking into account also the performance on the external test set, it can be said that two 5-dimensional models and two 6-dimensional models can provide a larger NER_{cv} (0.985 and 0.990, respectively) and the same NER_{ext} (1.000) that the model with all the 13 variables; the advantages being the larger NER_{cv} and the lower number of variables.

Also the benchmark model comprises 5 variables, but shows slightly lower performance both in cross-validation and on the test set than the 5-dimensional models obtained by the RSR algorithm. However, these differences can be due to the different method, in fact the reference model is based on a LDA, while the models developed in this study are based on the *K*-NN method.

It is interesting to note that the ranking of the variables on the basis of the *CMC* index provided in [29] is somehow reflected in the variable selection performed with the RSR algorithm. In fact, variables 8, 3 and 5, which have the lowest ranking on the basis of the *CMC* index, were never selected, with the exception of 1 model with 6 variables that comprises variable 3.

Due to the limited number of variables, the differences in computational time between the different settings are not sharp and they are not reported here. However, the same trend observed on the other datasets emerges here as well. Additional considerations can be found in the next paragraph.

4.5. Effects of the optimization options

In this study, for each real and pseudo-real dataset a DoE was carried out taking into account the options that affect the search of the algorithm, i.e. roulette wheel, tabu list and *QUIK* rule (only for regression). A further analysis was carried out in order to investigate whether computational time and statistics of the best found models (in cross-validation) linearly depend on the analyzed factors. To this end, OLS models were developed using the design matrices of the aforementioned functions as **X** variables and computational time and the NER_{cv} (or Q_{cv}^2 for regression) as dependent variables. In all cases very good linear relationships were found. The results on computational time are summarized in Table 11. The reported coefficients were standardized in order to allow comparisons.

It can be noted that all the options, when activated, decrease the computational time, as indicated by the negative regression coefficients, the ranking being: tabu list > *QUIK* rule > roulette wheel. The larger

Table 10
Results of the DoE on the Wines dataset.

VS	Size	NER_{cv}	NER_{ext}	1	2	3	4	7	9	10	11	12	13	TL: 0 RW: 0	TL: 1 RW: 0	TL: 0 RW: 1	TL: 1 RW: 1
RSR	6	0.990	1.000	x		x	x	x		x			x	•		•	
RSR	6	0.990	0.985	x	x			x	x		x		x	•			•
RSR	6	0.990	1.000	x	x			x		x	x		x			•	•
RSR	5	0.985	1.000	x	x						x	x	x			•	
RSR	5	0.985	0.985	x				x	x		x		x			•	
RSR	5	0.985	1.000	x	x			x		x			x			•	
RSR	4	0.985	0.985	x				x	x				x	•	•	•	•
RSR	3	0.975	0.953					x	x	x				•	•	•	•
RSR	2	0.963	0.936					x	x					•	•	•	•
RSR	1	0.865	0.839					x						•	•	•	•
FS	5	0.980	0.980	x				x	x			x	x	–	–	–	–

Table 11
OLS models of the effects of tabu list, roulette wheel and *QUIK* rule on computational time.

Dataset (n, p)	R^2	Variable coefficients		
		TL	RW	QR
LC50 (465, 899)	0.955	−0.7213	−0.2550	−0.6080
MP (12634, 150)	0.991	−0.1030	−0.1149	−0.9836
Breast* (699, 50)	0.935	−0.9351	−0.2472	–
ItaOils* (572 50)	0.959	−0.9581	−0.2021	–
Wines (178, 13)	0.956	−0.9609	−0.1802	–

* Pseudo-real datasets.

coefficients of the tabu list for classification datasets may be due to the fact that, in classification, for each model several values of K (number of neighbors) are tested, leading to an increase in the computational time. The exclusion of some variables (those included in the tabu list), can therefore allow a sharper saving in the computational time compared to regression. The roulette wheel has a comparable effect on all datasets, the standardized coefficients ranging between −0.1149 and −0.2550. For the MP dataset, since when the *QUIK* rule was activated the maximum model size had to be decreased to 6 (no models comprising more than 6 variables could fulfill the *QUIK* rule criterion), the effect of the functions are not meaningful and the coefficients cannot be compared.

The regression models on NER_{cv} or Q^2_{cv} gave null or near-null regression coefficients, indicating that the activation of these functions does not affect the quality of the developed models compared to the SR algorithm.

It is therefore given a mathematical justification for the comments drawn afore regarding the effects of the analyzed functions. However these equations are not meant to be the regarded as the general rule, but simply as the results of the calculations performed in this study on the analyzed datasets.

5 Conclusions

In this study a modified version of the Sequential Replacement algorithm for variable selection is proposed. Some modern functionalities were implemented with the aim to 1) reduce the computational time; 2) use a fitness function that is related to the real predictivity of the model; 3) identify models suffering from pathologies, such as collinearity between variables, chance correlation, excess of noisy or explanatory variables; 4) allow an easier comparison of the different final models. We called this redesigned version Reshaped Sequential Replacement (RSR) algorithm.

The RSR algorithm was applied to four and three datasets in regression and classification, respectively, including simulated and pseudo-real datasets (real datasets with additional random variables). The methods used in this work were Ordinary Least Squares (OLS) in regression and K-Nearest Neighbors (K-NN) in classification. A Design of Experiments (DoE) was carried out for all real and pseudo-real datasets taking into account only the functions that affect the outcome of the search, namely tabu list, roulette wheel and *QUIK* rule. The DOE aimed at studying the effects of the new introduced functions on a) computational time, b) quality of final models and c) their ability to identify models suffering from pathologies, with respect to the SR algorithm. The results were also compared with models obtained by different variable selection methods.

The RSR algorithm managed to find good subsets of variables on all datasets, even though the reference models were not always found. When the reference model was not found the RSR algorithm returned comparable or better subsets of variables, evaluated in cross-validation (Q^2_{cv} or NER_{cv}). On the MP and LC50 datasets the reference model did not fulfill the test based on the R^N index, which indicates the likely presence of noisy variables. For the LC50 dataset, all the models found by the RSR algorithm showed a larger Q^2_{cv} but a smaller

Q^2_{ext} compared to the reference model (the reference model, when found, was not considered because it did not fulfill the R^N function). For the MP dataset, the final models showed both larger and smaller Q^2_{cv} and Q^2_{ext} compared to the reference model. The results on the simulated datasets in regression showed the ability of the additional functions to save computational time and discard noisy variables when the objects/variables ratio is relatively high.

Since the R^N index, which proved to be effective in the identification of noisy variables in regression, cannot be applied in classification, the tabu list can be used instead to this end when dealing with classification problems.

The effect of tabu list, roulette wheel and *QUIK* rule on the computational time and models' performance was evaluated by the development of regression models on the design matrices. The negative regression coefficients indicated that all the functions decrease the computational time. At the same time, null coefficients were obtained when the performance of the models was considered as response, suggesting that the activation of these functions does not lead to models with lower performance.

Finally, *CMC* and *CMD* indices provide useful information about correlation and distance between the models in the final population, especially when the number of potential variables is large and several pairwise correlations are highly probable.

Appendix A

Appendix A briefly presents the software toolbox of the RSR algorithm. The RSR algorithm was implemented as toolbox for Matlab and will soon be available on the website of our research group [31]. The toolbox follows the scheme reported in Fig. 3 and provides several options, a summary of which is explained in Table A.2. The results can be saved in two excel spreadsheets. The command line of the RSR toolbox is the following:

$$[res, res\ calc] = rsr\ model(X, y, options, labels\ x, X\ test, y\ test)$$

with the meanings collected in Table A.1.

The RSR toolbox provides Ordinary Least Squares (OLS) and Principal Component Regression (PCR) as regression methods; the K-Nearest Neighbors (K-NN) as classification method. Different scaling methods and options to carry out the cross-validation are available. In particular two algorithms for the definition of the cross-validation groups were implemented, namely contiguous blocks and venetian blinds. As a consequence, the cross-validation groups are the same for each model to be validated, making the results comparable and consistent. For K-NN method, the user can choose between different distance measures both for real and binary data, the maximum k value and the type of a priori probability.

By interpreting the parameters given in Table A.2, four seeds are generated for each model size. Being the minimum model size equal to 2 and the maximum model size equal to 5, the total number of generated seeds is 16. The tabu list, *QUIK* rule and roulette wheel are activated. The thresholds for *QUIK* rule and R^P index are set to 0.05

Table A.1
Input/output data for the RSR toolbox.

Name	Explanation
X	Data matrix
y	Data response
options	Options
labels_x	Labels of the independent **X** variables
X_test	Data matrix of external test set (optional)
y_test	Response of external test set (optional)
res	RSR output: includes final population of models and statistics
res_calc	RSR output: calculated responses for final models and predictions for the test set (if provided)

Table A.2
An example of some options available in the RSR toolbox.

Option	Default value	Explanation
minvar	2	Minimum model size (number of variables)
maxvar	5	Maximum model size (number of variables)
numseed	4	Number of seeds (models) for each model size
QR	1	Settings to activate/deactivate the *QUIK* rule
TL	1	Settings to activate/deactivate the tabu list
RW	1	Settings to activate/deactivate the roulette wheel
thr_QR	0.05	Threshold value (δK) for the *QUIK* rule
thr_RP	0.01	Threshold value for R^P index
thr_RN	0.02	Threshold value of ε for R^N index

and 0.01 respectively. The value of ε for the calculation of the threshold for the R^N index is set to 0.02.

It is worth to highlight that the toolbox does not provide a single final model nor it suggests a best model. The functions applied on the final population of models (R-based functions, *Y-scrambling*, nested models and *RER* tests, *CMD* and *CMC*) provide valuable information to take a decision and select a final model, but this analysis and the final choice must be taken by the user. However, we suggest the outcome of the R-based functions and the nested models and *RER* tests be taken much into consideration. The outcome of these tests is a 0/1 flag, where 1 stands for test fulfilled and vice versa.

Appendix B

Appendix B gives a brief introduction to the variable selection methods used to derive the reference models that were published together with the datasets used in this study.

Genetic Algorithms (GA) [32,11] are a meta-heuristic method derived from Darwin's Theory of the evolution. In the terminology of GA each variable, which is represented by a bit that can assume a value of 0 (variable excluded) or 1 (variable included), is called a gene. A combination of genes is named chromosome and defines a specific combination of variables, which corresponds to a model. An initial population is usually generated randomly and the performance of each chromosome is evaluated. The chromosomes are then sorted according to their performance and two procedures are used to create new chromosomes, namely cross-over and mutation. Cross-over consists in selecting parent chromosomes from the population and combining them in order to generate offspring. Mutation instead implies the casual inversion of some bits of existing chromosomes producing mutants. The performance of the new generated chromosomes is evaluated and, if any of them is better than any of the chromosomes in the population, the new chromosomes enter the population and the worst ones are discarded. GAs use cross-over and mutation to improve the performance of the population, exactly in the same way as evolution generates fitter individuals. The evolution continues until a predefined stop criterion is met, such as a certain number of iterations.

Forward Stepwise (FS) [7] is a classic approach to variable selection that begins with a model of size 0 and adds variables that meet a predefined criterion. In its original formulation the variable to be added at each step is the one that minimizes the Residuals Sum of Squares (RSS) at the greatest extent. A stop criterion, based on the F-test, is used to define the optimal model size according to:

$$F_j^+ = \max_j \left[\frac{RSS_p - RSS_{p+j}}{s_{p+j}^2} \right] > F_{in} \tag{A.1}$$

where RSS_p and RSS_{p+j} are the residuals sum of squares of the models with p and $p+j$ variables, s_{p+j}^2 is the variance of the model built with variables $p+j$ and F_{in} is used as stop criterion corresponding to the

probability α, with 1 degree of freedom for the numerator and ($n–p–1$) for the denominator.

Random Forests (RF) [35] are en ensemble learning method that can be used both in classification and in regression. It consists in the generation of a large number of regression (or classification) trees, each of them trained by a bootstrap sample of the dataset. Trees are constituted of nodes and for each node a random subset of the $\boldsymbol{X}$ variables is chosen. The outputs of the individual trees are combined together to provide an overall prediction based on the mode. RF can be used to retrieve information regarding the importance of predictor variables. This estimation is based on the permutation of the values of each variable (leaving the values of the other variables unchanged) and considering the increase in the prediction errors. The estimation of variable importance is carried out tree by tree during the construction of the forest.

References

[1] R. Todeschini, V. Consonni, Molecular descriptors for chemoinformatics, second ed.vol. 41 Wiley-VCH, 2009.
[2] DRAGON (Software for Molecular Descriptor Calculation). Talete srl, 2012.
[3] ADRIANA, Code. Molecular Networks Gmbh, http://www.molecular-networks.com/products/adrianacode.
[4] CODESSA, Codessa Pro, http://www.codessa-pro.com/.
[5] ISIDA Fragmentor. France: Laboratoire d'Infochimie, Institut de Chimie, Université de Strasbourg, 2011. http://infochim.u-strasbg.fr/spip.php?rubrique49.
[6] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, second ed Springer-Verlag, 2009.
[7] M.A. Efroymson, Multiple Regression Analysis, in Mathematical Methods for Digital Computers, Wiley, New York, 1960.
[8] A.J. Miller, Selection of subsets of regression variables, J. R. Stat. Soc. Ser. A (General) 147 (1984) 389–425.
[9] A.J. Miller, Subset Selection in Regression, second ed Chapman & Hall/CRC, 2002.
[10] J.H. Holland, Adaptation in Natural and Artificial Systems, MIT Press, 1992.
[11] R. Leardi, Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier, 2003.
[12] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, MobyDigs: software for regression and classification models by genetic algorithms, in: R. Leardi (Ed.), Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks, Elsevier, 2003.
[13] Q. Shen, J.H. Jiang, C.-X. Jiao, G. Shen, R.Q. Yu, Modified particle swarm optimization algorithm for variable selection in MLR and PLS modeling: QSAR studies of antagonism of angiotensin II antagonists, Eur. J. Pharm. Sci. 22 (2004) 145–152.
[14] M. Goodarzi, M.P. Freitas, 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, Chemom. Intell. Lab. Syst. 98 (2009) 123–129.
[15] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2004) 407–499.
[16] R. Tibshirani, Regression shrinkage and selection via the Lasso, J. R. Stat. Soc. Ser. B Methodol. 58 (1996) 267–288.
[17] R. Todeschini, V. Consonni, A. Mauri, M. Pavan, Detecting 'bad' regression models: multicriteria fitness functions in regression analysis, Anal. Chim. Acta 515 (2004) 199–208.
[18] P.R. Duchowicz, E.A. Castro, F.M. Fernandez, Alternative algorithm for the search of an optimal set of descriptors in QSAR–QSPR studies, MATCH Commun. Math. Comput. Chem. 55 (2006) 179–192.
[19] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories, Chemom. Intell. Lab. Syst. 92 (2008) 138–144.
[20] R. Todeschini, D. Ballabio, V. Consonni, A. Manganaro, A. Mauri, Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between sets of data. Part 1. Theory and simple chemometric applications, Anal. Chim. Acta 648 (2009) 45–51.
[21] R. Todeschini, V. Consonni, A. Maiocchi, The K correlation index: theory development and its application in chemometrics, Chemom. Intell. Lab. Syst. 46 (2009) 13–29.
[22] R. Todeschini, Data correlation, number of significant principal components and shape of molecules. The K correlation index, Anal. Chim. Acta 348 (1997) 419–430.
[23] F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, L. Eriksson, Model validation by permutation tests: applications to variable selection, J. Chemom. 10 (1996) 521–532.
[24] M. Pavan, T.I. Netzeva, A.P. Worth, Validation of a QSAR model for acute toxicity, SAR QSAR Environ. Res. 17 (2006) 147–171.
[25] A. Lang, ONS challenge, http://onschallenge.wikispaces.com/MeltingPointModel001.
[26] W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proc. Natl. Acad. Sci. 87 (1990) 9193–9196.
[27] M. Forina, C. Armanino, S. Lanteri, E. Tiscornia, Food Research and Data Analysis, in: Classification of Olive Oils from their Fatty Acid Composition, Applied Science Publishers, London, 1983.

[28] R. Leardi, A.L. González, Genetic algorithms applied to feature selection in PLS regression: how and when to use them, Chemom. Intell. Lab. Syst. 41 (1998) 195–207.

[29] D. Ballabio, V. Consonni, A. Mauri, R. Todeschini, Canonical Measure of Correlation (CMC) and Canonical Measure of Distance (CMD) between sets of data. Part 3. Variable selection in classification, Anal. Chim. Acta. 657 (2010) 116–122.

[30] V. Consonni, D. Ballabio, R. Todeschini, Comments on the definition of the Q2 parameter for QSAR validation, J. Chem. Inf. Model. 49 (2009) 1669–1678.

[31] MICHEM, http://michem.disat.unimib.it/chm/.

[32] J.H. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. USA: University of Michigan, U Michigan Press, Oxford, England, 1975.

[33] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

Appendix V

Grisoni, F., Cassotti, M., Todeschini, R., 2014. Reshaped Sequential Replacement for variable selection in QSPR: comparison with other reference methods. *J. Chemometrics* 28, 249-259.

Available at:

Reshaped Sequential Replacement for variable selection in QSPR: comparison with other reference methods

F. Grisoni, M. Cassotti and R. Todeschini*

The objective of the present work was to compare the Reshaped Sequential Replacement (RSR) algorithm with other well-known variable selection techniques in the field of Quantitative Structure–Property Relationship (QSPR) modelling. RSR algorithm is based on a simple sequential replacement procedure with the addition of several 'reshaping' functions that aimed to (i) ensure a faster convergence upon optimal subsets of variables and (ii) reject models affected by chance correlation, overfitting and other pathologies. In particular, three reference variable selection methods were chosen for the comparison (stepwise forward selection, genetic algorithms and particle swarm optimization), aiming to identify benefits and drawbacks of RSR with respect to these methods. To this end, several QSPR datasets regarding different physical–chemical properties and characterized by different objects/variables ratios were used to build ordinary least squares models; in addition, some well-known (Y-scrambling) and more recent (R-based functions) statistical tools were used to analyse and compare the results. The study highlighted the good capability of RSR to find optimal subsets of variables in QSPR modelling, comparable or better than those found by the other reference variable selection methods. Moreover, RSR resulted to be faster than some of the analysed variable selection techniques, despite its extensive exploration of the variables space. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: variable selection; Reshaped Sequential Replacement; QSPR; QSAR

1. INTRODUCTION

Variable selection (VS) is a key step in multivariate analysis for modelling purposes. It consists in the selection of optimal subsets of variables, in order to obtain parsimonious models, with maximum predictive power and increased interpretability. VS plays a crucial role in scientific fields that deal with a large number of variables, such as Quantitative Structure–Property/Activity Relationship (QSPR/QSAR). QSPR and QSAR are based on the assumption that the structure of a molecule is responsible for its physical, chemical and biological properties. The QSPR (and QSAR) approach can be generally described as an application of statistical and mathematical methods to the issue of finding empirical relationships expressed in the form

$$Y_i = f(x_1, x_2, \ldots, x_p)_i$$

where Y_i is the property of interest of the i-th compound, $x_1, x_2, \ldots, x_p$ are the p predictors of the i-th molecule and f represents the mathematical relationship between independent variables and the property. Molecular descriptors are used as predictors. They 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' [1].

Nowadays it is possible to calculate thousands of different descriptors. However, according to Occam's razor principle [2], it is reasonable to assume that only a small number of them are correlated to the experimental response and are, therefore, relevant for building the mathematical model of interest. Furthermore, one fundamental aspect is to find the good trade-off

between bias and complexity of the model. The increase of the complexity due to a larger number of descriptors included in the model is able to improve the fitness to the training data, but the inclusion of too many variables can often cause a reduction in the predictive ability, leading to 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. The optimal subset of variables is reflected in a good predictive ability, more robustness and stability of the model [3].

In this scenario, VS plays a key role in QSAR/QSPR, allowing to select the optimal subset of molecular descriptors for modelling the activity/property of interest and to obtain robust and predictive models. In this way, also the interpretability of the models increases, and non-significant effects can be neglected.

Throughout the years, many different methods and techniques have been proposed to address the problem of VS (e.g. [4,5]). From the classical approaches (e.g. stepwise Backward Elimination (BE) and Forward Selection (FS) [6]) to more sophisticated VS methods. In recent years, the so-called nature-inspired methods [7–9], such as Genetic Algorithms (GA) [10], Particle

* Correspondence to: R. Todeschini, Milano Chemometrics and QSAR Research Group—Department of Earth and Environmental Sciences, University of Milan–Bicocca, P.za della Scienza 1, 20126 Milan, Italy.
E-mail: roberto.todeschini@unimib.it

F. Grisoni, M. Cassotti, R. Todeschini
Milano Chemometrics and QSAR Research Group—Department of Earth and Environmental Sciences, University of Milan–Bicocca, P.za della Scienza 1, 20126 Milan, Italy

Swarm Optimization (PSO) [11], Ant Colony Optimization [12] and Evolutionary Programming [13], have progressively increased in importance. Moreover, the group of penalization techniques, such as Least Absolute Shrinkage and Selection Operator (LASSO) [14] and elastic net [15], recently gained interest from the scientific community to address the issue of VS: These methods were initially aimed at improving the problems of ordinary least squares (OLS) regression and are able to select variables via the shrinkage of the regression coefficients towards zero.

Recently, we proposed a VS method based on Miller's Sequential Replacement (SR) [16], the Reshaped Sequential Replacement (RSR) [17]. Being based on the same replacement procedure, the two methods share a good exploration capability. Some new reshaping features were included in the RSR algorithm in order to (i) decrease the computational time (ensuring a faster convergence towards optimal solutions) and (ii) identify models suffering from different types of pathologies. Our previous study highlighted the capability of the method to perform a good exploration of the space of the variables and of the new reshaping functions to significantly speed up the modelling time and discard models that suffer from different pathologies, such as overfitting or chance correlation.

In the present study, we compared RSR with other widely used VS methods: (i) stepwise FS; (ii) GA; and (iii) PSO. RSR and the reference methods were applied to four QSPR datasets in regression with different properties and objects/variables ratios. The primary objectives were to compare the performances of these VS methods in the field of QSPR modelling and identify benefits and drawbacks of RSR method with respect to the others.

After a brief introduction about reference VS methods (Section 2), the theory of RSR algorithm (Section 3) and details about the materials and methods (Section 4) are presented. Results and discussion can be found in Section 5.

2. REFERENCE VARIABLE SELECTION METHODS

2.1. Stepwise regression

Stepwise regression (SWR) methods [6] are among the most known feature selection methods. SWR is based on two different strategies, namely Forward Selection (FS) and Backward Elimination (BE). FS starts with a model of size 0 and proceeds by adding variables that fulfil a pre-defined criterion. BE method proceeds in the opposite way with respect to FS: It starts from a model of size p (p being the total number of variables), and non-relevant variables are eliminated in a step-by-step procedure. Typically, the inclusion (or exclusion) criterion is the residual sum of squares (RSS): At each step, the variable to be added (or eliminated) is the one that leads to the maximum decrease (or minimum increase) of the RSS.

2.2. Genetic Algorithms

Genetic Algorithms are a nature-inspired method [10,18,19] that takes inspiration from Darwin's theory of evolution. In analogy with biological systems, each gene represents a variable, and each chromosome (sequence of genes/variables) can be seen as a potential model. The evolution of the population of chromosomes is determined by two processes: (i) crossover, in which pairs of chromosomes generate offspring according to a crossover probability; and (ii) mutation, in which some genes of a chromosome can change according to a mutation probability.

Every time a new chromosome with a better fitness function (e.g. Q_{cv}^2) than 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'.

2.3. Particle Swarm Optimization

Particle Swarm Optimization is an agent-based method inspired by the behaviour of flock of birds [20,21]. Differently from GA, PSO agents do not compete but cooperate in order to find the best solutions. PSO was initially thought of as an optimization method and only later modified in order to address the problem of VS [11]. PSO agents are particles that move in a binary space (in the variant for VS) in which each dimension corresponds to a variable and each position to a model. The particle motion is controlled by a parameter called static probability, which determines the probability of each particle to move to its previous personal best position, to the best global position or to remain in its current position, balancing exploration and exploitation ability of the method.

3. RESHAPED SEQUENTIAL REPLACEMENT ALGORITHM

The RSR method is based on the SR method proposed by Miller in 1984 [16]. The basic idea of Miller's method is to start from a randomly generated model (seed), replace each variable at a time with all the remaining ones and see whether a better model can be obtained. The best model found in the first replacement procedure becomes the new seed for a further replacement. This procedure goes on until no better models can be found. Miller's method has the advantage of performing a good exploration of the variables space, but with the drawback of being extremely time consuming when the number of variables increases.

The RSR method [17] implements new reshaping functionalities over Miller's algorithm that aim to:

(1) decrease the calculation time, retaining the exploration capability of the method;
(2) increase the probability of convergence upon the optimal models;
(3) identify models that suffer from several pathologies, such as overfitting, chance correlation, variable redundancy and collinearity between predictors.

Moreover, the coefficient of determination in cross-validation (Q_{cv}^2, see Section 4.2) is used as a fitness function instead of the RSS used in the original SR algorithm, the latter not necessarily being related with the predictive ability of the model.

The functions able to 'reshape' the original method are as follows:

(1) Tabu list (TL): Preliminary exclusion of variables not correlated with the response according to their univariate Q_{cv}^2 in regression. It aims at decreasing the computational time with respect to SR algorithm. Variables are excluded according to the following criterion:

$$\text{if } Q_{cv}^2(\boldsymbol{y}, \boldsymbol{x}) < 0 \Rightarrow \boldsymbol{x} \in TL \qquad (1)$$

When the algorithm reaches convergence, tabu variables are re-included and used for a last replacement procedure starting

from the optimal population of models. Tabu variables will be selected in each seed only if they provide an improvement to the model larger than a pre-defined threshold (e.g. 0.01 on Q_{cv}^2). TL resulted to be the principal function able to decrease the computational time with respect to SR algorithm (up to 10 times faster).

(2) Roulette wheel (RW): Used for the initialization of the population. Each variable is given a probability of entering the initial population proportional to a chosen fitness function (univariate Q_{cv}^2 in regression). This pre-selection algorithm is supposed to generate models closer to the optimal solution, being biased towards promising variables.

(3) QUIK rule: A statistical test [22] used in regression during the replacement procedure, in order to reject *a priori* models affected by high predictor collinearity. The collinearity among variables is one of the main problems when applying multiple linear regression that can lead to undesirable consequences [23,24]. The QUIK rule is based on the K multivariate correlation index [25] and the comparison between the internal correlation of the X-block (K_x) and the correlation of the X-block plus the y response (K_{xy}):

$$\text{if } K_{xy} - K_x < \delta K \quad \Rightarrow \quad \text{reject the model} \quad (2)$$

The basic assumption is that the total correlation of the independent variables (**X**) selected in the model plus the response (**y**) should be larger than the total correlation calculated on the selected independent variables only. If this criterion is not fulfilled, the model is rejected before being statistically evaluated.

(4) Evaluation functions: Implemented to evaluate the final population of models.

(i) *R*-function-based rules [22] to identify: (a) models with redundancy in explanatory variables (R^P index) and (b) models with noisy variables (R^N index).

(ii) Y-scrambling (in regression): A statistical randomization test [26] commonly used to identify the presence of chance correlation between predictors and response.

(iii) Canonical Model Correlation (CMC) and Canonical Model Distance (CMD) [27]: For the comparison of final models. CMC and CMD allow an easy comparison of the final models in order to determine whether models with different variables are actually different in their nature.

(iv) Nested models screening. A model *F* can be defined as 'nested' if there is a model *G* of higher size (i.e. including more variables) that comprises all the variables of *F* and has a very similar performance (i.e. the difference in their Q_{cv}^2 is smaller than a pre-defined threshold, e.g. 0.005). If this occurs, model *G* is rejected because its higher complexity is not balanced by a better performance.

A simplified flowchart of the algorithm is depicted in Figure 1.

4. MATERIALS AND METHODS

4.1. Variable selection strategies

4.1.1. Stepwise regression

Stepwise regression was performed with FS using the maximization of the coefficient of determination in cross-validation (Q_{cv}^2) as

criterion for the progressive inclusion of each variable up to a maximum model size, chosen as stop criterion.

4.1.2. Genetic Algorithms

In addition to the classic GA approach, based on a single run with a randomly initialized population, the version of GA proposed by Leardi and González was used. This version aims to overcome the principal limitations of GA, i.e. the tendency to overfit data and to model noise if the response is noisy and/or a limited number of objects is present (or the ratio objects/variables is small). The approach is based on (i) execution of a large number of runs with different randomly generated initial populations; (ii) optimization of the number of evaluations for each run; and (iii) final stepwise selection approach, based on the frequency of selection of each variable over all the runs. Two further features characterize the algorithm by Leardi and González: (i) for the principle of parsimony and to prevent overfitting, a chromosome *M* cannot enter the population if another chromosome *F* exists that has a higher fitness and is a subset of the variables of *M*; and (ii) GA can be hybridized with a BE procedure that is carried out during or at the end of each run. In order to distinguish the classic approach from the approach of Leardi and González, in this work, they were identified as GA and GA-SW, respectively.

4.1.3. Particle Swarm Optimization

In the modified PSO of Shen *et al.* [11], the balance between exploration and exploitation ability of the method is intended to change during the motion of the particles; therefore, the static probability starts with a value equal to 0.5 and decreases to a final value equal to 0.33. According to PSO approach for VS, variables that are not included in the initial random population cannot be included during the run, thus being the exploration capability of this method limited. For this reason, the strategy of Leardi and González (execution of a large number of runs, optimized number of evaluations, check for nested models and BE, and final stepwise) was also adopted for PSO. This version was referred to as PSO-SW.

4.2. Model validation

For all VS methods, the coefficient of determination in cross-validation (Q_{cv}^2) was used as fitness function. Q_{cv}^2 is defined as follows:

$$Q_{cv}^2 = 1 - \frac{\sum_{i=1}^{n_{train}} \left(y_i - \hat{y}_{i/i} \right)^2}{TSS} \quad (3)$$

where n_{train} is the number of training objects, y_i is the real response value of the *i*-th object and $\hat{y}_{i/i}$ is the value of the *i*-th object predicted by the model in which the *i*-th object was not taken into consideration; *TSS* (total sum of squares) is the sum of squared deviations from the dataset mean. In this work, a leave-more-out strategy was used with a 'venetian blind' resampling technique that makes the values of Q_{cv}^2 calculated on different models comparable and consistent. Furthermore, the predictive power of the models was assessed also by means of external validation on a test set. This was expressed by the coefficient of determination on the external test set (Q_{ext}^2), calculated as follows [28]:
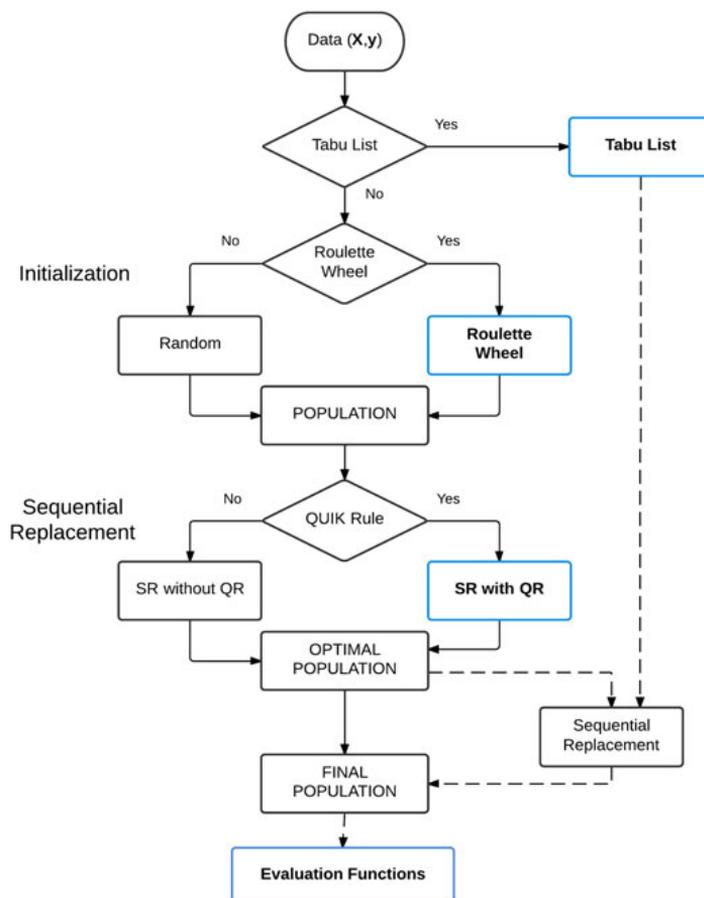
Figure 1. RSR method: simplified flowchart of the algorithm. The new 'reshaping' functions are highlighted in boldface.

$$Q_{ext}^2 = 1 - \frac{\left(\sum_{i=1}^{n_{ext}} (y_i - \hat{y}_i)^2 \right) / n_{ext}}{TSS / n_{train}} \qquad (4)$$

where n_{ext} is the number of objects in the external test set; $\hat{y}_i$ and y_i are the predicted response and the real response of the i-th test object, respectively; n_{train} is the number of objects in the training set; and TSS is the total sum of squares calculated on the training set.

Finally, in order to represent the ability of the model to fit the training data, the coefficient of determination was also calculated:

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{train}} (y_i - \hat{y}_i)^2}{TSS} \qquad (5)$$

where $\hat{y}_i$ and y_i represent the calculated response and the real response of the i-th object, respectively. R^2 represents the percentage of the variance explained by the model.

4.3. Datasets

In the present work, comparisons were made on four QSPR datasets that were retrieved from US EPA website and had been used in T.E.S.T. software to develop models [29]. The chosen properties are (i) boiling point (BP), (ii) vapour pressure (VP), (iii) thermal conductivity (TC) and (iv) flash point (FP).

For each dataset, molecular descriptors from 0D to 2D were calculated by means of Dragon 6 [30]. Constant, near constant and descriptors having a standard deviation lower than 0.1 were deleted. Variables showing a pair correlation larger than 0.95 with other descriptors were deleted. The original random splitting between training and external test set used in T.E.S.T. software was retained. The characteristics of the analysed datasets are reported in Table I.

4.4. Software and codes

In the present work, calculations were performed using MATLAB R2012b [31] on an Intel Xeon CPU E5-2620 0 at 2.00 GHz with 16 GB RAM.

All the MATLAB toolboxes and functions used in this study were written by our research group. The RSR toolbox will be soon available for free download on Milano Chemometrics website [32].

5. RESULTS AND DISCUSSION

The aim of the present work was to compare our method, RSR algorithm, with some reference VS methods, i.e. Stepwise Forward Selection (SWR), Genetic Algorithms (GA and GA-SW) and Particle Swarm Optimization (PSO-SW). To this end, the methods were applied to four QSPR datasets in regression.

For the sake of comparison, for all the VS techniques, OLS was always used as regression method and the internal validation was carried out by means of a fivefold cross-validation (Q_{cv}^2). The maximum dimension of the generated models was arbitrarily set to six for all the methods and all the datasets. This allows (i) an 'internal' comparison of all the VS methods on the same dataset and (ii) a 'global' comparison between the results of the same method on different datasets.

For RSR, three seeds were generated for each model size. RW, TL and the evaluation functions were enabled with the default thresholds; QUIK rule was disabled in order to compare the exploration capability of each method based only on Q_{cv}^2.

For both the GA and GA-SW approaches, crossover and mutation probabilities were set to 0.5 and 0.01, respectively, and the

number of chromosomes was set to 30. For the classic GA approach, a single run was performed with a different number of evaluations depending on the model size: 500 (from two to three variables), 700 (four variables) and 1000 (from five to six variables). For GA-SW and PSO-SW, 100 runs were performed with an optimized number of evaluations for each dataset.

For PSO-SW, the number of particles was set to 10, and the initial static probability to 0.5 (decreasing to a final value equal to 0.33).

The RSR, GA, GA-SW and PSO-SW algorithms were run three times on each dataset, being meta-heuristic methods. Being our implementation of SWR deterministic, it was run only once on each dataset.

Moreover, in order to have a better understanding of the results, R-function based rules and Y-scrambling were also applied *a posteriori* to the final population of models found by each reference method. Only the models that fulfilled Y-scrambling test and R^P and R^N rules were taken into account. Finally, for each dataset, the best model (based on Q_{cv}^2) provided by each method for each model size was used for the comparison.

All the models fulfilled Y-scrambling test, while different percentages of rejection by the R-function-based rules were observed on each dataset.

For BP dataset, 53% of the models were discarded by R-functions-based rules: all the discarded models did not fulfil R^N rule (presence of noisy variables) with the exception PSO-SW and GA-SW models that did not fulfil R^P rule (excess of explanatory variables). No models of GA-SW were accepted for this dataset.

Table II reports the best model (based on Q_{cv}^2) for each dimension found by each VS method. The models with largest Q^2 values, both in cross-validation and external validation, had five variables, thus showing that the increase in model complexity is not always balanced by an increase in predictive power. RSR provided the model with the largest Q_{cv}^2 (81.8%). GA found a model giving very similar Q_{cv}^2 (81.7%) and slightly larger Q_{ext}^2 (81.6%) compared with that of RSR model (81.1%). RSR and SWR found the same model with three variables. This model could be regarded as the best one because of its simplicity (it comprises only three descriptors) and its very good performance (Q_{cv}^2 and Q_{ext}^2 only, respectively, 2.8% and 2.2% lower than those of GA model with five variables).

Table II. Results on BP dataset sorted by $\mathbf{Q_{cv}^2}$: statistics, size and descriptors are reported for each method									
Method	R^2 (%)	Q_{cv}^2 (%)	Q_{ext}^2 (%)	Size			Descriptors		
RSR	82.0	81.8	81.1	5	ATS3p	GATS1v	JGT	CATS2D_05_LL	TPSA(NO)
GA	81.8	81.7	81.6	5	J_D	SM1_B(p)	ATSC1p	B02[F-F]	TPSA(NO)
GA	79.4	79.3	79.4	6	PCR	J_Dt	AVS_B(p)	SM1_B(p)	F-083	TPSA(NO)
RSR	79.2	79.0	79.4	3	SPI	SM1_B(p)	TPSA(NO)		
SWR	79.2	79.0	79.4	3	SPI	SM1_B(p)	TPSA(NO)		
RSR	78.4	78.3	78.0	4	WiA_Dt	SM3_D/Dt	GATS1v	B02[F-F]	
GA	75.2	75.1	71.4	4	nF	SpMaxA_L	HyWi_B(m)	TPSA(NO)	
GA	73.7	73.6	72.5	3	piID	SM1_B(p)	ATSC1i		
RSR	70.0	69.9	68.5	2	SM1_B(p)	TPSA(NO)			
SWR	70.0	69.9	68.5	2	SM1_B(p)	TPSA(NO)			
GA	69.2	69.1	69.6	2	piID	IAC			
PSO-SW	66.8	66.6	67.2	3	piPC07	piID	X3sol		
PSO-SW	61.9	61.8	62.2	2	SCBO	Xt			

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization, Leardi and González approach; SWR, stepwise regression with forward selection.

The most frequently selected descriptor (in eight out of 13 models) is 'TPSA(NO)', which represents the topological polar surface area calculated from polar fragments with nitrogen and oxygen contributions [33]. The 33% of the selected descriptors (14 out of 43) are 2D matrix-based, i.e. topological indices calculated from different graph-theoretical matrices.

Similar results were found for VP dataset. Approximately 43% of the models did not fulfil R-functions-based rules. Once again, most of the models found by GA and PSO with the approach of Leardi and González (GA-SW and PSO-SW) did not fulfil the R^P rules. On this dataset, the model that provided the largest statistics, both in cross-validation and external validation, comprised six variables (RSR algorithm), even though the difference in Q_{cv}^2 and Q_{ext}^2 with its best model with five variables is very small. RSR gave the best model for each dimension, and its best model with three and four variables was also found by SWR (Table III). The most frequently selected descriptors are 'piID' [34] (selected in 12 out of 17 models) and 'TPSA(NO)' [33] (11 out of 17 models), representing the conventional bond-order ID number and the topological surface area, respectively. The majority of the selected descriptors belong to the group of walk and path counts (24%) and molecular properties (22%).

On TC dataset, the percentage of rejected models was significantly lower (28%) than the previous cases. RSR gave the model with the largest Q_{cv}^2 (78.8%) and Q_{ext}^2 (77.5%) (Table IV). RSR model with six variables comprises all the descriptors of RSR model with five variables, plus the descriptor 'O-056'. It is possible to notice that the addition of the descriptor 'O-056' leads to only a modest increase in the Q_{cv}^2 (2.8% larger) but to a significant increase in the Q_{ext}^2 (10.5% larger). The best model provided by GA comprises five variables and had slightly lower Q_{cv}^2 than the RSR model of same size, but larger Q_{ext}^2. Unlike the previous datasets, most of the models provided by GA-SW and PSO-SW were accepted by the R-functions. The most frequently selected descriptors are (i) 'O %' (20 out of 22 models), the percentage of oxygen atoms; (ii) 'JGT' [35] (12 models), a global topological charge index; and (iii)

'X%' (10 models), the percentage of halogen atoms. The descriptor 'O-056' [36,37] (leading to a significant increase in the Q_{ext}^2 of RSR model with six variables) represents the number of alcohol fragments in the molecule. In 47% of the cases, the selected descriptors belong to the class of constitutional indices, i.e. the most simple and commonly used descriptors, which reflect the chemical composition of a compound without any information about its molecular geometry or atom connectivity [1].

Regarding FP dataset, R^N and R^P rules rejected approximately 51% of the models. RSR provided the model with the best performance in both cross-validation and external validation (Table V). The best model of RSR had a CMC of 0.85 with the best of SWR (five variables) and of 0.64 with the best of GA (six variables). In other words, even if RSR and SWR models result to be correlated, RSR provides a model with significantly larger Q_{cv}^2 (3.2% larger) and Q_{ext}^2 (3.5% larger). No models by PSO-SW and three out of six models by GA-SW were accepted by the R-based rules. The most frequently selected descriptors are 'TPSA(NO)' (12 out of 16 models) and 'SCBO' (eight models), the latter representing the sum of conventional bond orders [1]. Molecular properties and constitutional indices are the most frequently selected classes of descriptors, in 20% and 19% of the cases, respectively.

In order to make a global comparison of the results, a ranking of the best models for each dataset (Tables II–5) was made. For each dataset, the final population of models was tailed-ranked according to Q_{cv}^2 and Q_{ext}^2, obtaining two matrices. If for a certain model size the method did not provide accepted models, it was given the last position in the ranking. RSR models occupied high positions in the ranking both regarding the performance on training and test sets with the exception of two models (RSR6 and RSR5 on, respectively, BP and FP datasets) that did not fulfil the R-functions (Figure 2). GA also provided models with high positions in the ranking, and it was the only method that provided models accepted by the R-based rules for each model size on all datasets. On the contrary, GA-SW and PSO-SW, as noticed earlier, often gave models that did not fulfil the R rules and, in

Table III. Results on VP dataset sorted by Q_{cv}^2: statistics, size and descriptors are reported for each method

Method	R^2 (%)	Q_{cv}^2 (%)	Q_{ext}^2 (%)	Size	Descriptors					
RSR	87.1	86.8	88.3	6	S3K	piID	X3sol	GGI3	NssO	TPSA(NO)
RSR	86.5	86.3	88.1	5	nHM	ICR	WiA_Dt	NssO	TPSA(NO)	
RSR	83.9	83.7	85.3	4	piID	Eta_F_A	TPSA(NO)	MLOGP2		
SWR	83.9	83.7	85.3	4	piID	Eta_F_A	TPSA(NO)	MLOGP2		
GA	83.8	83.6	84.6	6	WiA_Dt	SpPosA_B(p)	ATS8m	F03[C-C]	F08[C-O]	TPSA(NO)
GA	83.6	83.4	84.1	5	WiA_Dt	SpPosA_B(p)	ATS8m	F08[C-O]	TPSA(NO)	
RSR	81.6	81.4	82.6	3	piID	Eta_F_A	TPSA(NO)			
SWR	81.6	81.4	82.6	3	piID	Eta_F_A	TPSA(NO)			
GA	81.1	80.7	80.3	4	ICR	piID	CATS2D_02_DA	TPSA(NO)		
GA	76.2	76.1	73.6	3	piID	ATS8m	B04[C-O]			
RSR	76.1	75.9	78.2	2	WiA_Dt	TPSA(NO)				
SWR	75.6	75.4	75.4	2	piID	TPSA(NO)				
GA-SW	74.7	74.5	73.0	3	ICR	piID	X5			
PSO-SW	73.3	73.2	71.7	3	ECC	piPC03	piID			
GA-SW	72.9	72.8	71.4	2	ICR	piID				
GA	71.7	71.4	70.4	2	piID	CATS2D_02_DA				
PSO-SW	68.6	68.5	67.5	2	ECC	piPC03				

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization—Leardi and González approach; SWR, stepwise regression with forward selection.

Table IV. Results on TC dataset, sorted by Q_{cv}^2: statistics, size and descriptors are reported for each method

Method	R^2 (%)	Q_{cv}^2 (%)	Q_{ext}^2 (%)	Size	Descriptors				
RSR	81.5	78.8	77.5	6	N%	O%	J_Dz(p)	EE_B(m)	JGT	O-056
RSR	78.7	76.0	67.0	5	N%	O%	J_Dz(p)	EE_B(m)	JGT	
GA	77.8	74.8	73.2	5	N%	O%	MAXDN	JGT	nROH	
RSR	75.9	73.4	63.0	4	N%	O%	EE_B(m)	JGT		
GA	72.6	70.9	63.1	4	nN	O%	SM1_Dz(Z)	JGT		
RSR	72.0	69.5	56.6	3	N%	O%	JGT			
SWR	72.0	69.5	56.6	3	N%	O%	JGT			
GA-SW	67.2	66.2	62.3	4	O%	X%	JGT	B02[F-F]		
GA-SW	67.6	65.8	63.4	6	O%	X%	JGT	B01[C-O]	B02[F-F]	MLOGP
GA-SW	67.5	65.8	62.5	5	nO	O%	X%	JGT	B02[F-F]	
GA	68.0	65.2	73.2	6	nN	nO	ATS2m	SpMax4_Bh(s)	Eta_sh_p	O-056
PSO-SW	65.0	62.9	61.6	6	O%	X%	SIC1	SpMax_L	B01[C-O]	B02[F-F]
GA	63.5	62.6	72.1	3	O%	X%	nOHp			
PSO-SW	64.3	62.5	60.2	5	O%	IC1	SpMax_L	JGI1	B01[C-O]	
SWR	63.0	62.2	55.2	2	O%	JGT				
RSR	63.0	62.2	55.2	2	O%	JGT				
PSO-SW	62.5	61.5	62.5	3	O%	X%	SpMax_L			
GA-SW	62.1	61.1	58.7	3	O%	X%	B02[F-F]			
PSO-SW	62.5	60.9	62.6	4	O%	X%	SpMax_L	MLOGP		
GA-SW	59.5	58.7	60.0	2	O%	X%				
PSO-SW	59.5	58.7	60.0	2	O%	X%				
GA	44.4	43.2	38.2	2	B01[C-O]	B02[C-F]				

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization—Leardi and González approach; SWR, stepwise regression with forward selection.

Table V. Results on FP dataset, sorted by Q_{cv}^2: statistics, size and descriptors are reported for each method

Method	R^2 (%)	Q_{cv}^2 (%)	Q_{ext}^2 (%)	Size	Descriptors				
RSR	81.3	81.1	81.5	6	MW	ZM1Kup	piID	CATS2D_02_DA	T(O..O)	TPSA(NO)
SWR	78.2	77.9	78.0	5	SCBO	MWC02	SM1_Dz(p)	T(O..O)	TPSA(NO)	
GA	77.9	77.7	79.0	6	J_D	AVS_B(m)	SM1_B(p)	nImidazoles	F04[C-O]	TPSA(NO)
RSR	77.7	77.5	78.6	4	J_D	SM1_B(p)	F02[C-O]	TPSA(NO)		
SWR	77.0	76.4	77.2	4	SCBO	SM1_Dz(p)	T(O..O)	TPSA(NO)		
GA	76.5	76.3	78.5	5	piID	X3sol	Eig13_AEA(ri)	F02[C-O]	TPSA(NO)	
RSR	75.1	74.9	77.2	3	piID	SM1_B(p)	TPSA(NO)			
GA	74.3	74.1	75.7	4	SCBO	CATS2D_02_DA	F06[C-O]	TPSA(NO)		
SWR	74.0	73.7	74.8	3	SCBO	T(O..O)	TPSA(NO)			
GA	73.0	72.7	74.4	3	SCBO	F08[C-O]	TPSA(NO)			
RSR	71.4	71.2	73.3	2	SM1_B(p)	TPSA(NO)				
SWR	70.4	70.2	73.7	2	SCBO	TPSA(NO)				
GA-SW	68.5	68.5	71.8	3	SCBO	nN	MWC02			
GA-SW	67.8	67.7	71.2	2	SCBO	nN				
GA-SW	64.1	63.9	67.8	5	MPC07	Eig04_AEA(dm)	Eig11_AEA(ri)	CATS2D_08_DL	B07[C-N]	
GA	62.0	62.0	66.9	2	MWC02	X3sol				

RSR, Reshaped Sequential Replacement; GA, genetic algorithms; GA-SW, genetic algorithms of Leardi and González; PSO-SW, particle swarm optimization—Leardi and González approach; SWR, stepwise regression with forward selection.

particular, the R^P rule (99% of the cases). This could be related to the final stepwise based on the selection frequency of each variable over all the runs. In fact, if two or more relevant variables are correlated, as it is often the case of molecular descriptors, their frequency of selection over all the runs is likely to be similar.

This reflects on the inclusion of both variables in the final stepwise model even if they carry the same information, thus causing redundancy in explanatory predictors. These limitations could be connected to the fact that the method was originally proposed for PLS modelling of spectral and

255

wileyonlinelibrary.com/journal/cem
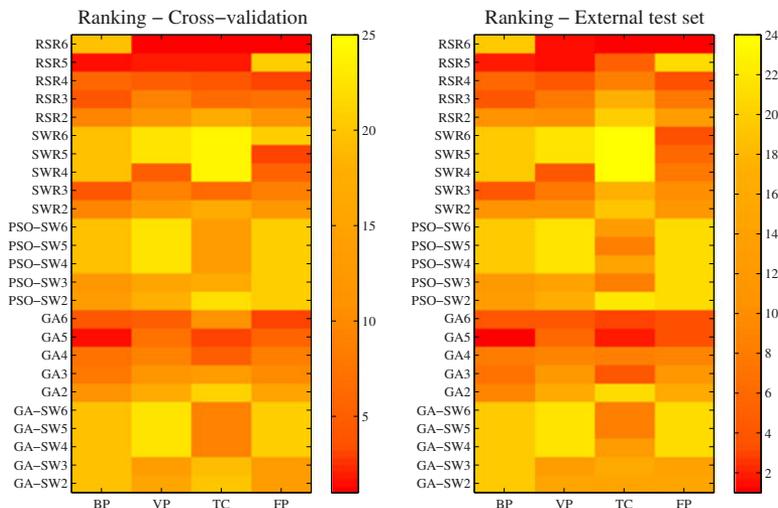
Figure 2. Heat map of the ranking of the best models for each method on training (Q^2_{cv}) and test (Q^2_{ext}) sets. The darker the colour, the higher the ranking position.
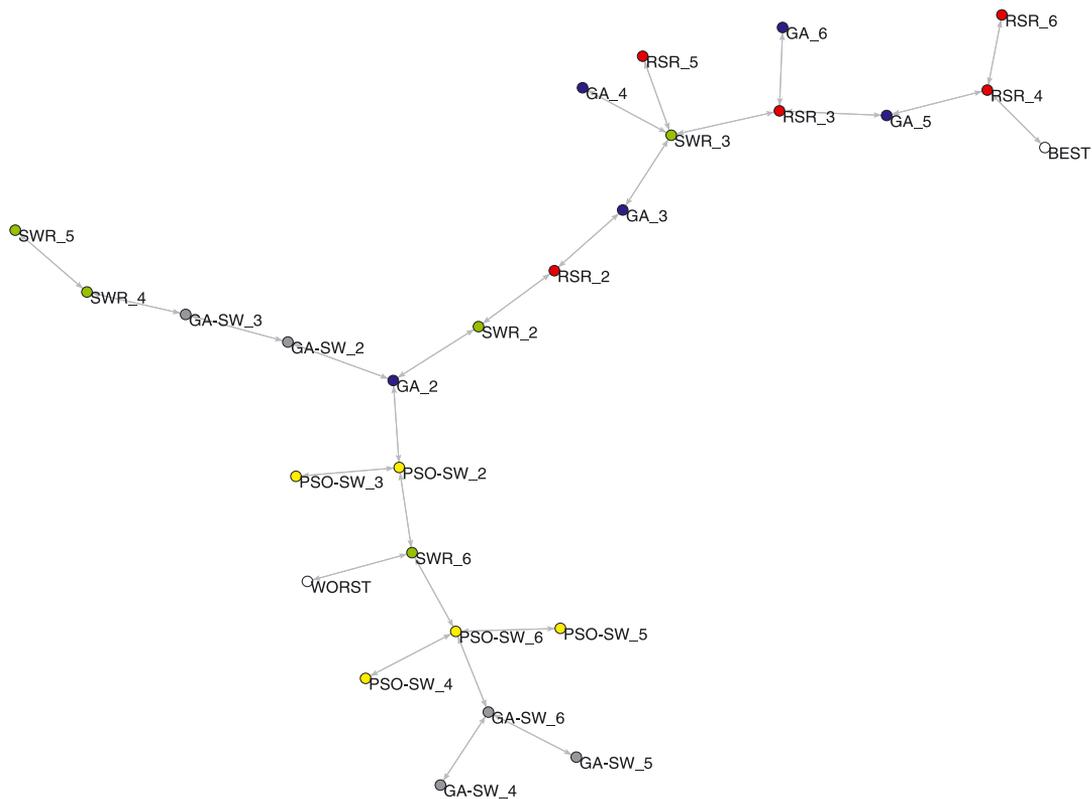

Figure 3. Minimum spanning tree of the ranking of the best model on the basis of $\mathbf{Q^2_{cv}}$. Labels correspond to the method and number of variables.

chromatographic datasets and not thought for OLS regression applied to molecular descriptors.

In order to further compare the results of the methods, the matrices were then range scaled between 1 and 0, and minimum spanning trees (MST) were built (Figures 3 and 4). Two dummy models were also added, BEST (always first position in the ranking) and WORST (always last position), in order to visually locate the optimal and non-optimal regions.

In Figure 3 (MST on the Q_{cv}^2-based ranking), a clear separation occurs between RSR and GA models that lay in the proximity of BEST, in respect to those of GA-SW and PSO-SW, which are close to WORST; SWR models show intermediate behaviour. RSR models with six, four, three and two variables are closer to BEST than those of the same size of the other methods. Finally, PSO-SW and GA-SW lay in the region of WORST because of both their poor performance in cross-validation and their high rejection rate by the R rules.

By observing the MST made on the Q_{ext}^2 ranking (Figure 4), the situation appears similar. Still a clear separation between GA/RSR and other methods can be seen. In this case, however, GA models are closer to BEST with respect to RSR models. This behaviour is connected to the tendency of GA to give models with lower Q_{cv}^2 but slightly larger Q_{ext}^2 than RSR for the same size, even if in three out of four cases RSR provided the model with the largest Q_{ext}^2, regardless of model size. As already noticed for the Q_{cv}^2-based MST, PSO-SW and GA-SW are in the region of WORST for the poor performance in prediction with respect to the other VS methods

and for the large number of models rejected by the R rules. Finally, the simplest model in the proximity of BEST is RSR-4.

The same information can be obtained by ranking the models according to the sum of ranking differences [38], using the maximum as reference value. RSR and GA models always occupy high positions in the ranking for all the datasets, while all the methods based on the final stepwise approach (i.e. GA-SW and PSO-SW) occupy the lowest positions in the ranking; SWR shows an intermediate behaviour, depending on the dataset.

In general, RSR appears to perform similar to SWR for what concerns low-dimensional models: in most of the cases, in fact, the two methods find the same solutions with two and three variables. On the other hand, for higher dimensional models, results of RSR diverge from those of SWR and are more similar to those of GA. SWR is a widely used method but known to have several drawbacks [39,40], such as the bias related to the inclusion of one variable at a time without considering other subsets of variables [41] and the tendency to model noise [39]. These problems increase in their relevance with the increase of the number of variables included. In this perspective, the divergence of RSR from SWR when the model dimension increases can be seen as representative of the ability of RSR to extensively explore the possible subsets and combinations of variables. Moreover, the activation of TL allows to temporarily exclude variables not correlated with the response, thus preventing overfitting and noise modelling, which are often related to the extensive exploration of the combinations
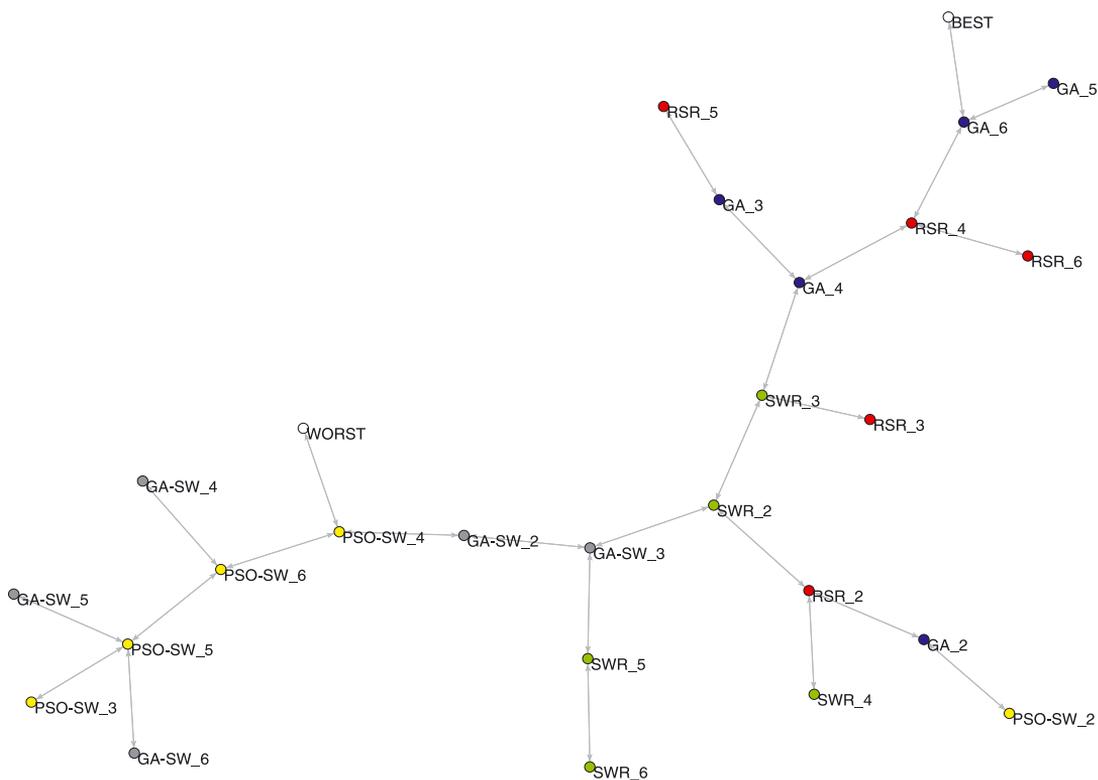

Figure 4. Minimum spanning tree of the ranking of the best model on the basis of $\mathbf{Q_{ext}^2}$. Labels correspond to the method and number of variables.

of variables. The drawbacks of SWR are confirmed by R^N rule, which rejected 100%, 75% and 25% of its models with respectively six, five and four variables.

Finally, for what concerns computational time, the general ranking of the methods is as follows: GA-SW < RSR (about 1.6 times slower than GA-SW) < PSO-SW (about 2.2 times slower than GA-SW) < GA (about 7.1 times slower than GA-SW) < SWR (7.7 times slower than GA-SW). In other words, RSR, despite its extensive exploration of the variables space, has a computational time comparable with that of the other meta-heuristic methods, thanks to the addition of TL and RW functions.

6. CONCLUSIONS

In the present work, the RSR algorithm for VS was compared with other reference methods: Genetic Algorithms (GA and GA-SW), Particle Swarm Optimization (PSO-SW) and Stepwise Forward Selection (SWR).

The methods were applied to four QSPR datasets that differ in their objects/variables ratios and the physical-chemical property to be modelled.

In order to analyse the final populations of models by means of a common procedure, Y-scrambling test and R-function-based rules were applied, and only the models that fulfilled these tests were retained.

In three out of four cases, RSR algorithm found the best models, in terms of Q^2_{cv} and Q^2_{ext}, while in the other case, the best model was found by GA.

The GA-SW and PSO-SW often provided models that did not fulfil R-based rules, and the accepted models had, in most of the cases, a poor performance in cross-validation and external validation. Moreover, RSR found low-dimensional (less than four variables) models similar to those of SWR, while for higher dimensions, the performance of models found was more similar to that of genetic algorithms.

Computational time of RSR resulted to be comparable with that of GA-SW and PSO-SW and lower than that of GA and SWR.

REFERENCES

1. Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics* (2nd Revised and Enlarged Edition). Wiley-VCH: Weinheim, Germany; 2009.
2. Myung IJ, Pitt MA. Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon. B. Rev.* 1997; **4**:79–95.
3. Yu L, Lai KK, Wang S, Huang W. A bias–variance–complexity trade-off framework for complex system modeling. In *Computational Science and Its Applications—ICCSA 2006*, Gavrilova M, Gervasi O, Kumar V, Tan CJK, Taniar D, Laganá A, *et al.* (eds.). Springer: Berlin Heidelberg, 2006; 518–27.
4. Gonzalez MP, Teran C, Saiz-Urra L, Teijeira M. Variable selection methods in QSAR: an overview. *Curr. Top. Med. Chem.* 2008; **8**: 1606–27.
5. Guyon I, Elisseeff A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 2003; **3**: 1157–82.
6. Efroymson M. Multiple regression analysis. *Math. Methods Digital Comput.* 1960; **1**: 191–203.
7. Kim K-J, Cho S-B. A comprehensive overview of the applications of artificial life. *Artif. Life* 2006; **12**(1): 153–82.
8. Leardi R *Nature-inspired Methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*. Elsevier: Amsterdam, The Netherlands; 2003.
9. Yang X-S. *Nature-Inspired Metaheuristic Algorithms: Second Edition*. Luniver Press: Frome, United Kingdom; 2010.
10. Goldberg DE, Holland JH. Genetic algorithms and machine learning. *Mach. Learn.* 1988; **3**(2-3): 95–9.
11. Shen Q, Jiang J-H, Jiao C-X, Shen G, Yu R-Q. Modified particle swarm optimization algorithm for variable selection in MLR and PLS
12. Dorigo M, Di Caro G. Ant colony optimization: a new meta-heuristic. *Proceedings of the 1999 Congress on Evolutionary Computation* 1999. p. -1477 Vol. 2.
13. Fogel DB. *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. John Wiley & Sons: New Jersey, USA; 2006.
14. Tibshirani R Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B (Methodological)* 1996; **58**(1): 267–88.
15. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodology)* 2005; **67**(2): 301–20.
16. Miller AJ. Selection of subsets of regression variables. *J. R. Stat. Soc. Ser. A (General)*. 1984; **147**(3): 89–425.
17. Cassotti M, Grisoni F, Todeschini R. Reshaped Sequential Replacement algorithm: an efficient approach to variable selection. *Chemom. Intell. Lab.* (in press). 2014. DOI: 10.1016/j.chemolab.2014.01.011
18. Bledsoe W. The use of biological concepts in the analytical study of systems. *Proceedings of the ORSA-TIMS National Meeting* 1961.
19. Holland JH. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. University of Michigan: USA, 1975. Oxford, England: U Michigan Press.
20. Eberhart R, Kennedy J. A new optimizer using particle swarm theory. *Proceedings of the Sixth International Symposium on Micro Machine and Human Science* 1995; p. 39–43.
21. Kennedy J, Eberhart RC. A discrete binary version of the particle swarm algorithm. *IEEE International Conference on Systems, Man, and Cybernetics*, 1997; p. 4104–4108 vol.5.
22. Todeschini R, Consonni V, Mauri A, Pavan, M. Detecting 'bad' regression models: multicriteria fitness functions in regression analysis. *Anal. Chim. Acta* 2004; **515**(1): 199–208.
23. Mason CH, Perreault WD. Collinearity, power, and interpretation of multiple regression analysis. *J. Marketing Res.* 1991; **28**(3): 268–80.
24. Stewart GW. Collinearity and least squares regression. *Stat. Sci.* 1987; **2**(1): 68–84.
25. Todeschini R, Consonni V, Maiocchi A. The K correlation index: theory development and its application in chemometrics. *Chemom. Intell. Lab. Syst.* 1999; **46**(1): 13–29.
26. Lindgren F, Hansen B, Karcher W, Sjöström M, Eriksson L. Model validation by permutation tests: applications to variable selection. *J. Chemom.* 1996; **10**(5-6): 521–32.
27. Todeschini R, Ballabio D, Consonni V, Manganaro A, Mauri A. Canonical measure of correlation (CMC) and canonical measure of distance (CMD) between sets of data. Part 1. Theory and simple chemometric applications. *Anal. Chim. Acta* 2009; **648**(1): 45–51.
28. Consonni V, Ballabio D, Todeschini R. Evaluation of model predictive ability by external validation techniques. *J. Chemom.* 2010; **24**(3-4): 194–201.
29. US EPA 2013. Toxicity Estimation Software Tool (T.E.S.T.)—http://www.epa.gov/nrmrl/std/qsar/qsar.html [9 September 2013]
30. Talete srl. Dragon (Software for Molecular Descriptor Calculation) Version 6.0—2012—http://www.talete.mi.it/. 2012.
31. MATLAB. R2012b. *Natick.* The MathWorks Inc.: Massachusetts, 2012.
32. MICHEM website. http://michem.disat.unimib.it/chm/.
33. Ertl P, Rohde B, Selzer P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*. 2000; **43**(20): 3714–7.
34. Randić M, Jurs PC. On a fragment approach to structure–activity correlations. *Quant. Struct.-Act. Relat.* 1989; **8**(1): 39–48.
35. Galvez J, Garcia R, Salabert MT, Soler R. Charge indexes. New topological descriptors. *J. Chem. Inf. Comput. Sci.* 1994; **34**(3): 520–5.
36. Ghose AK, Viswanadhan VN, Wendoloski JJ. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* 1998; **102**(21): 3762–72.
37. Viswanadhan VN, Ghose AK, Revankar GR, Robins RK. 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.* 1989; **29**(3): 163–72.

38. Héberger K, Kollár-Hunek K. Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *J. Chemom.* 2011; **25**(4): 151–8.

39. Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *Br. J. Math. Stat. Psychol.* 1992; **45**(2): 265–82.

40. Steyerberg EW, Eijkemans MJC, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J. Clin. Epidemiol.* 1999; **52**(10): 935–42.

41. Burnham KP, Anderson DR. *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*. Springer: New York, USA; 2002.

Copyright © 2014 John Wiley & Sons, Ltd.