

 *BOOK*

editor: Emilio Benfenati

Theory, guidance and applications
ON QSAR AND REACH

ISBN 978-88-902405-4-6

 ORCHESTRA



Editor

Emilio Benfenati

Istituto di Ricerche Farmacologiche “Mario Negri”

Via Privata Giuseppe La Masa 19, 20156, Milano (Italy)

First edition 2012

© Copyright 2012

Istituto di Ricerche Farmacologiche “Mario Negri”, Milan, Italy

Revisers

Claire Mays (*Institute SYMLOG de France, Paris, France*)

Simon Pardoe (*PublicSpace Ltd. Research Communication, Lancaster, UK*)

Art Direction

AGDesigner - Adriana Gomez - www.agdesigner.it

 *BOOK*

editor: Emilio Benfenati

Theory, guidance and applications
ON QSAR AND REACH

Foreword

R. L. Stevenson, in his masterpiece Strange Case of Dr Jekyll and Mr Hyde, imagined that it was possible to separate the good and evil in man. A chemical potion could reveal the evil side of humankind. At the basis of this work of fantasy there is an ethical issue, merged with the scientific fiction of the revelatory technical device.

Modern science is currently working towards the more modest goal of identifying the good and bad nature of chemical substances, with the aim of eventually pinpointing the components which make the chemical toxic. This can be done through so-called QSAR models.

The basic hypothesis of a quantitative structure-activity relationship (QSAR) model is that a given property or effect can be put into relationship with the structure of a chemical, which is described by reference to certain parameters. In order to achieve this, we need a good mathematical algorithm as well as suitable ways to describe the chemical.

With this eBook you will learn about the state of the QSAR art. This book composed of three parts: The first will address the scientific aspect of modelling. Then we will examine some practical cases. Finally, we will discuss the possible applications of QSAR with particular attention to the REACH legislation.

Other than the scientific issues, we will refer to the debate on the suitable and appropriate use of these models for specific applications, and thus we will also deal with regulatory, political, and cultural matters. The actors interested in the use of QSAR models are diverse, with different backgrounds, opinions, and interests. Thus, the acceptance and broad application of QSAR models are not solely related to the statistical power of a model. This eBook aims to give you tools to evaluate when QSAR models can help further the goal of protecting health and the environment.



The EC project ORCHESTRA

We acknowledge the EC project ORCHESTRA.

The EC funded ORCHESTRA project (2009-12) to promote a wider understanding, awareness and appropriate use of in silico methods. The project team included chemists and social scientists. We interacted directly and online with a range of organisations and individuals.

The ORCHESTRA project made a contribution to good practice and regulation by bringing together EU research on in silico methods and practical experience of their use.

Our web portal is a central information resource. It is a clearinghouse of knowledge and experience for professionals who are developing and using in silico models. For industry users, regulators and academics, it provides downloadable software for the in silico models reviewed by the project.

To build up ORCHESTRA's information offer, we consulted with regulators, chemical manufactures, importers and their associations, academic and other researchers, small business users, NGOs and citizen groups in the EU, to find out their needs for knowledge about in silico methods.

This e-book is just one of many documents developed to generate scientific discussion and understanding of in silico methods. Find them, along with proceedings of ORCHESTRA seminars and workshops, online at the project's official website using one of the following links:

www.orchestra-qsar.eu

www.in-silico-methods.eu

www.insilicomethods.eu

ORCHESTRA was coordinated by Emilio Benfenati

Instituto di Ricerche Farmacologiche 'Mario Negri', Milan.

The project partners

[PublicSpace Ltd. UK.](#)

Institut SYMLOG, Paris.

[Politecnico di Milano, Milan.](#)

[Universitaet Stuttgart.](#)

[University of Patras, Greece.](#)

[Centro Reach S.r.l., Milan.](#)

Table of Contents

| | |
|---|------|
| <i>Foreword</i> | V |
| <i>The EC project ORCHESTRA.</i> | VII |
| The project partners | VIII |
| PART A Theory | 3 |
| <i>Emilio Benfenati, Rodolfo Gonella Diaza, Giuseppina Gini, Luigi Cardamone, Magdalena Gocieva, Marina Mancusic, Rima Padovanic, Lorenzo Tamellini</i> | |
| Introduction to in vivo, in vitro and in silico methods | 3 |
| In silico for predicting properties: the QSAR approaches | 6 |
| The components of QSAR models | 9 |
| <i>Experimental values, their quality and uncertainty</i> | 9 |
| <i>The chemical information: descriptors and fragments</i> | 12 |
| <i>2D and 3D descriptors</i> | 14 |
| <i>The way to represent the chemical structure</i> | 16 |

| | |
|--|----|
| <i>The choice of the suitable complexity in the chemical descriptors and chemical format</i> | 16 |
| <i>Software for chemical descriptors calculation</i> | 18 |
| <i>More on chemical descriptors</i> | 19 |
| <i>The modelling algorithms: classifiers and regression models</i> | 20 |
| <i>Different algorithms</i> | 21 |
| <i>Free tools & algorithms for QSAR modelling</i> | 23 |
| The way that a QSAR model is built up | 24 |
| The validation of the model | 26 |
| <i>Evaluation of a classifier</i> | 27 |
| <i>Evaluation of a regression model</i> | 29 |
| <i>The validation of the QSAR models and non-testing methods and regulatory needs</i> | 29 |
| Hybrid models | 31 |
| Further perspectives on <u>in silico</u> models | 34 |
| Chapter references | 35 |

PART B Practical case 39

Emilio Benfenati, Rodolfo Gonella Diaza, Andrea Gissi

| | |
|--|----|
| Many tools, many purposes, many QSAR models | 39 |
| A series of existing models | 40 |
| Models for bioconcentration factor (BCF). | 42 |
| <i>BCF & REACH</i> | 43 |
| <i>QSAR and BCF for REACH</i> | 44 |
| <i>The CAESAR Model</i> | 45 |
| <i>QSAR Model Reporting Format</i> | 48 |
| <i>Scientific validity</i> | 48 |
| <i>Classification approaches for BCF</i> | 49 |
| <i>The read-across model</i> | 53 |
| <i>The integration and interpretation of the results</i> | 53 |
| <i>Case studies</i> | 54 |

| | |
|---|----|
| The models for mutagenicity | 55 |
| <i>The endpoint</i> | 55 |
| <i>Mutagenicity & REACH</i> | 55 |
| <i>The model</i> | 56 |
| The models for carcinogenicity | 60 |
| <i>The endpoint</i> | 60 |
| <i>Carcinogenicity & REACH</i> | 61 |
| <i>QSAR and Carcinogenicity for REACH</i> | 61 |
| <i>The CAESAR model</i> | 62 |
| <i>The results of the CAESAR classification model for Carcinogenicity</i> | 63 |
| The models for Developmental toxicity | 64 |
| <i>The endpoint</i> | 64 |
| <i>Developmental toxicity & REACH</i> | 64 |
| <i>QSAR and Developmental toxicity for REACH</i> | 65 |
| <i>The CAESAR models</i> | 66 |
| The models for skin sensitization | 68 |
| <i>The endpoint</i> | 68 |
| <i>Skin sensitization & REACH</i> | 69 |
| <i>QSAR and Skin sensitization for REACH</i> | 70 |
| <i>The CAESAR Models</i> | 70 |
| <i>Global QSAR model</i> | 70 |
| <i>Local QSAR model</i> | 71 |
| Chapter References | 75 |

PART C The applications of QSAR models: industrial, pharmaceutical and regulatory applications 81

Emilio Benfenati, Rodolfo Gonella Diaza, Andrea Gissi

| | |
|--|----|
| Regulatory Context | 82 |
| The REACH legislation | 85 |
| Use of QSAR models within a REACH perspective | 85 |

| | |
|---|-----|
| The REACH requirements for QSAR | 88 |
| Model validity | 89 |
| <i>How to evaluate the scientific validity of the QSAR <u>model</u>?</i> | 89 |
| The applicability domain | 90 |
| <i>How to evaluate the applicability domain?</i> | 91 |
| The model adequacy of QSAR prediction | 92 |
| The documentation and the model transparency | 93 |
| QSAR models in the regulatory perspective | 94 |
| <i>Different perspectives for a broader QSAR scenario</i> | 96 |
| <i>The QSAR model and the role of the expert</i> | 97 |
| An ideal framework for the development of QSAR for regulatory purposes | 100 |
| The debate and the open issues | 101 |
| Chapter References | 103 |

ANNEXS

| | |
|--|-----|
| ANNEX 1 QSAR and SAR models: Basic definitions | 105 |
| ANNEX 2 An introduction to toxicology | 107 |
| <i>Giuseppina Gini, Luigi Cardamone, Magdalena Gocieva, Marina Mancusi, Rima Padovani, Lorenzo Tamellini</i> | |
| ANNEX 3 Rigid and flexible topological indices: variety of the representations of the molecular structure | 113 |
| <i>Andrey A. Toropova, Alla P. Toropovaa, Emilio Benfenati, Giuseppina Gini</i> | |
| Introduction | 113 |
| Construction of optimal descriptors | 117 |
| Optimal descriptors based on HSG and HFG | 117 |
| OCWLGI-descriptors based on GAO | 119 |
| Collection of QSPR/QSAR based on OCWLGI-descriptors (optimal descriptors) | 121 |

| | |
|--|-----|
| Discussion | 124 |
| Conclusions | 124 |
| Abbreviations | 125 |
| References | 125 |
| | |
| ANNEX 4 A free and open source informatics library for chemistry: Chemistry Development Kit (CDK) | 133 |
| | |
| References | 139 |
| | |
| ANNEX 5 REACH | 141 |
| <i>Giuseppina Gini, Luigi Cardamone, Magdalena Gocieva, Marina Mancusi, Rima Padovani, Lorenzo Tamellini</i> | |
| | |
| Novelties introduced by REACH | 142 |
| <i>Registration</i> | 143 |
| <i>Evaluation</i> | 144 |
| <i>Authorization</i> | 144 |
| <i>Restrictions</i> | 145 |
| | |
| Overview of interesting websites on REACH | 147 |
| <i>Websites totally dedicated to REACH</i> | 147 |
| <i>General information websites</i> | 147 |
| <i>Consulting services' websites</i> | 148 |
| <i>Websites partly dedicated to REACH</i> | 149 |
| | |
| References | 150 |
| | |
| ANNEX 6 Genetic Algorithms in QSAR for REACH | 153 |
| <i>Orazio Nicolotti, Andrea Gissi, Antonellina Introcaso, Angelo Carotti</i> | |
| | |
| Introduction | 153 |
| What is a GA? | 154 |
| Why use GAs in QSAR? | 157 |
| Stay fit stay well: learn by doing! | 158 |
| | |
| Conclusions | 161 |

| | |
|---|-----|
| Acknowledgements | 161 |
| References | 162 |
| ANNEX 7 The CORAL software: principles, results, perspectives | 167 |
| <i>Andrey A. Toropov, Alla P. Toropova, Emilio Benfenati, Giuseppina Gini</i> | |
| Introduction | 167 |
| Principles | 169 |
| Results | 170 |
| Perspectives | 172 |
| Conclusions | 172 |
| References | 173 |

Emilio Benfenati^a, Rodolfo Gonella Diaza^a,
Giuseppina Gini^b, Luigi Cardamone^b,
Magdalena Gocieva^b, Marina Mancusi^c,
Rima Padovani^c, Lorenzo Tamellini^b

a. Istituto di Ricerche Farmacologiche Mario Negri,
Via La Masa 19, 20156, Milano, Italy

b. Politecnico di Milano, Piazza L. da Vinci 32, 20133,
Milano, Italy

c. Politecnico di Torino, Corso Duca degli Abruzzi 24,
10129, Torino, Italy

Theory

Introduction to in vivo, in vitro and in silico methods

In our everyday lives we have to deal with an exponentially increasing number of different chemical compounds. About 28 million substances have been registered so far, including food colouring and preservatives, drugs, varnishes, paints, pesticides and many others. It is well recognised that chemicals may pose a high risk to the environment and to humans, and therefore their toxic activity has to be assessed.

Biological active substances interact with bio-molecules, triggering specific mechanisms, like the activation of an enzyme cascade or the opening of an ion channel, which finally lead to a biological response. These mechanisms, determined by the chemical composition of the relevant substances, are unfortunately largely unknown; thus toxicity must be studied experimentally.

It is possible use three types of approaches to assess the biological activity of a molecule: *in vivo* experiments, i.e. animal testing; *in vitro* experiments, which involves using tissue culture cells; and *in silico* simulations, which refers to computer-based screening.



Figure 1: The three experimental ways to evaluate chemical substances: *in vivo*, *in vitro* and *in silico*.

Both animal testing and *in vitro* experiments are time consuming and expensive. Additionally, animal testing is now considered ethically unacceptable by a growing majority of people. For these reasons, and also due to improvements in computational power, the scientific community and the industrial world have started to use *in silico* approaches, or at least view them as a possible viable alternative, and have developed a number of models and strategies able to predict the properties of compounds.

Computational chemistry has changed the classical way to engage in experimental science. We are increasingly moving from experiments to simulations. We are able to model molecules according to different views, from the basic valence model to graph representation, from electronic clouds to 3D structure. Algorithms are available to compute the so-called “molecular descriptors”, which are variables (*either continuous or discrete*) representing structural properties. These descriptors range from simple properties to complex molecular fingerprints. Descriptors can help in transforming the study of interactions between molecules and living organisms in a data mining problem. Data mining could reveal relevant correlations between descriptors and the response of interest.

The algorithms look for correlations between the properties of the chemical structure of a compound and a measure of its activity/toxicity in a specific area, such as mutagenicity, carcinogenicity, or skin sensitization. This is called the “endpoint of interest”.

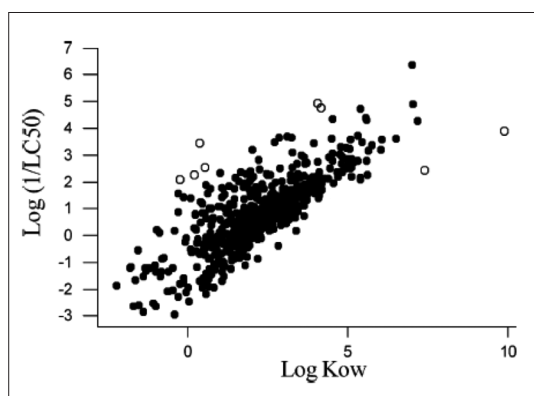


Figure 2: An example of QSAR analysis. The descriptor is LogKow; the endpoint is death and it is measured as a lethal concentration (LC 50).

In other words, once the structure of a compound is quantified in a set of molecular descriptors, these algorithms may be able to establish a mathematical relationship between the compound and, for example, its toxicity. In order to obtain the most reliable relation possible, a large dataset of compounds with known structure and experimentally determined property of interest is necessary to “train” the model.

The underlying idea of these models is that chemicals with similar structures, i.e. with similar values for the considered descriptors, must behave in a similar way. Thus, once the model is built, it can be used as a predictive tool in drug design, environmental protection and hazard analysis for all those compounds whose structure is similar to the structure of the ones used to tune the model.

The use of these models is growing, since they aim to provide fast, reliable and quite accurate estimates of the chemicals’ activity. These features also make them suitable for legislative purposes, and that is why they have been included as an alternative tool for risk assessment in the new European legislation on chemical production, called **R.E.A.C.H.** (*Registration, Evaluation, Authorisation and Restriction of Chemicals*). This legislation sets the rules for chemical production in the E.U., and one of its key points is the requirement of a risk analysis for each chemical placed in the European market in an amount greater than 1 ton/year. To further underline the breadth of this law, suffice it to say that the document is 849 pages long, and international mass media defined REACH as “*the most important legislation in European Union in 20 years*” (*BBC News, 28 November 2005*) and “*the strictest law to date regulating chemical substances*” (*San Francisco Chronicle, 14 December 2006*).

In silico for predicting properties: the QSAR approaches

Quantitative structure-activity relationship (QSAR) models are models linking a property or effect, such as boiling point or toxicity, to parameters associated with chemical structure, such as certain molecular descriptors. They can be used to assess chemical substances within the so-called in silico approach (as an analogy to the in vitro and in vivo approach). (Ref to Annexes “[annex 1. Qsar and sar models: basic definitions](#)”) Thus, the three main components of the QSAR models are:

1. The property to be modelled;
2. The chemical information;
3. The algorithm linking the property and the chemical.

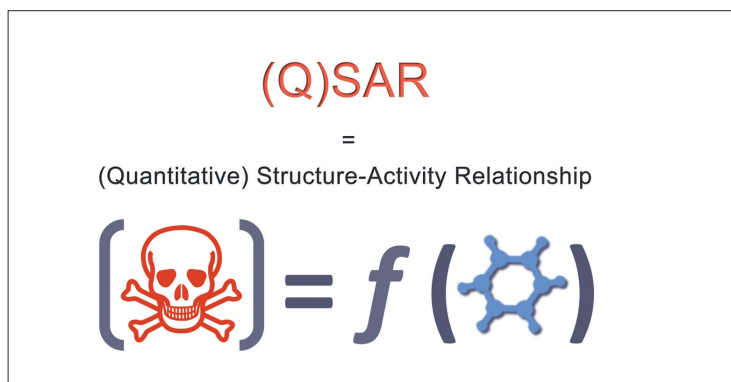


Figure 3: QSAR methods aims at finding the correlation between structural properties of chemicals and their activity.

The QSAR world is very complex, and it would be wrong to think that QSAR is one single method. There are thousands of chemical descriptors and thousands of chemical fragments, many diverse algorithms, studies addressing different endpoints, and for the same endpoint different sets of substances that may interest us. Thus, the number of possible models is stupendously high, and indeed thousands of models have been developed.

The communities working with the QSAR are also diverse, and there are typically conferences dedicated to applications to toxicological (*ref to annexes: [Annex 2. An introduction to toxicology](#)*)

or environmental matters or to the development of new chemicals, in particular pharmaceuticals. Recently there has been an exchange of opinions and methods between these two major communities. However, their goals and methodologies are different. Here we will principally address ecotoxicological applications, while discussing the major differences with pharmaceutical uses.

Even in the case of the models for environmental and toxicological endpoints the field is broad and complex. Historically, some studies started from the interest in the identification of the physico-chemical properties associated with a certain effect. For instance, in the 60's Corwin Hansch studied ecotoxicity, and put it in relationship with LogP, the partition coefficient between octanol and water, expressed as a logarithm. The idea behind this was that the partitioning between water and the organic solvent represents a model for the partitioning between water and the fish body, and the uptake of the toxic compound into the fish body is a good indicator of the toxic effect.

The driving force to conduct these studies was the identification of the key physico-chemical phenomena which could underpin the observed toxicity phenomena.

Under another approach, some studies sought to identify if chemical toxicity was related to the occurrence of a certain chemical moiety, such as a specific fragment for mutagenicity. In both cases there was an attempt to identify the cause of the toxicity, as explained by a descriptor or a fragment. The idea behind it all was that once the cause is known, we can govern the phenomenon, and thus predict the effect. Unfortunately, the situation is more complex, and these approaches were only partially successful in prediction. As we will see, other approaches have been used to analyse equally complex situations using probabilistic tools. Indeed, it is possible, even if the phenomenon is complex, to summarize the general behaviour within a certain probability.

There are some differences between the ecotoxicity and mutagenicity studies. The fish toxicity case introduced above refers to a toxic effect which has a modulation of the value, and is measured with a continuous value, such as fish acute toxicity. Another common endpoint, LD50, which is the dose that kills 50% of the animals (*rat or mouse, most typically*), is conceptually addressed in a similar way to fish toxicity, with toxicity levels increasing depending on the chemicals.

Conversely, toxic effects such as carcinogenicity or mutagenicity are often expressed as a binary conceptual system: toxic or not. The idea behind this is that a chemical can be carcinogenic even in a minute amount, starting a process which magnifies its effect with time. In such a case the idea is to identify the molecular component which provokes the phenomenon. Therefore, a SAR model, without the quantitative assessment, is appropriate.

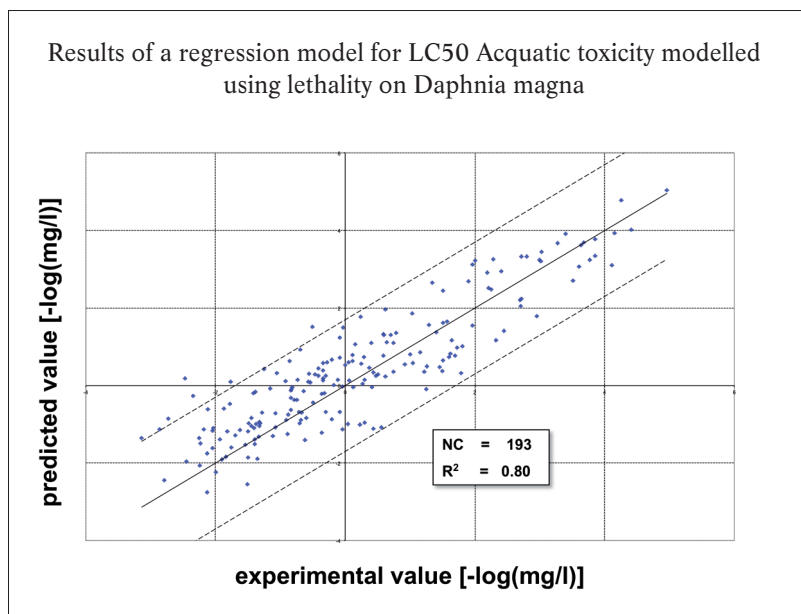


Figure 4: Results of a regression model for aquatic toxicity. The *Daphnia magna* LC50 has been used to build a predictive model for aquatic toxicity of chemicals; NC is the total number of compounds predicted whereas R² is the correlation coefficient calculated between experimental and predicted values.

Examples of molecular fragments related to carcinogenicity

| | | |
|--|--------------------------------|---|
| <p>First Level Structure: 1-Naphtho azocompounds</p> | <p>- 1st Level:</p> | <p>it identifies the structure of the nitrogen fragment characterizing the class and the aromatics structures bonded to that group</p> |
| <p>First Level Inhibition</p> | <p>- 1st Level Inhibition:</p> | <p>it solves the problem of compounds that, even if related to the structure of the subclass, are not carcinogens or have been ascribed to another subclass</p> |
| <p>Second Level Structure: Bensub-1NA residue</p> | <p>- 2nd Level:</p> | <p>it permits the identification of a specific compound or small groups of compounds that refer to the same subclass but differ for some specific elements bound to the nitrogen group and/or to the aromatic structure, and suspected to be involved in the carcinogenicity process.</p> |
| <p>Second Level Inhibition</p> | <p>- 2nd Level Inhibition:</p> | <p>this second inhibition level is useful to exclude a specific compound or a small group of compounds</p> |

Figure 5: Molecular fragments utilised for carcinogenicity in SAR models. In qualitative models (classifiers) the classification between toxic and non-toxic is commonly linked to the presence of particular fragments within the molecular structure; these fragments are usually called “structural alerts”. In this figure two types of fragments are shown: fragments related to carcinogenic effect and fragments which seems to lower the carcinogenicity of a molecule.

In the first case, quantitative tools are used involving quantitative descriptors and algorithms such as regressions (*figure 4* shows an example of results from a regression model). In the second case, fragments are used, and the algorithms are classifiers (*figure 5* shows examples of chemical fragments).

In practice, there are many instances of QSAR models which use combinations of the different chemical descriptors and fragments, and algorithms which can address quantitative or qualitative outputs.

The components of QSAR models

As previously mentioned, the basic hypothesis of a QSAR model is that the activity (*or effect or property*) can be put in relationship with the chemical, using some parameters to describe the chemical compound. Below we will analyse in more detail these three components of QSAR models.

Experimental values, their quality and uncertainty

Even before considering QSAR models, experimental data are necessary for many purposes such as chemical risk assessment. Animal models are used to assess the effects on more complex targets. For instance, when evaluating the effects on the ecosystem, models using fish can be used. Obviously, the model which involves putting a certain number of fish in a tank is much simpler than any ecosystem, where many organisms are present, in conditions which are more complex than those adopted within the experimental model.

Similarly, animal models are frequently used for human toxicity. In this case there is the issue of extrapolation from one species (*a rat, for instance*) to humans. Furthermore, the issue of the effects on the whole human population includes many varying situations, because children, pregnant women, sick people, and other sensitive parts of the population have to be protected.

For practical reasons, experiments on animals are done using a limited number of species and situations. Furthermore, it is necessary to get comparable results when using the same test, and thus the experimental parameters are typically fixed within defined protocols.

A fundamental concept is that any experimental value is associated with an uncertainty value. This is true for physical measurements, and even more so for biological data. Toxicity values are always affected by high levels of uncertainty. For instance, when one biotest is compared to a second for equivalence, it is accepted that the value changes by a factor of five [1]. In the case of BCF the reported experimental uncertainty can be up to 0.75 in Log unit [2].

As another example, the reproducibility of the Ames mutagenicity test, which is a quite simple model using bacteria, rather than complex organisms, is about 85%. In other words, if we give 6 substances to two laboratories, we can expect a contradictory output for one of these 6 chemicals. *Figure 6* shows the reproducibility of experimental toxicity data within three different high quality databases.

Reproducibility of experimental ecotoxicity data *LC50 (Lethal concentration) values from different databases*

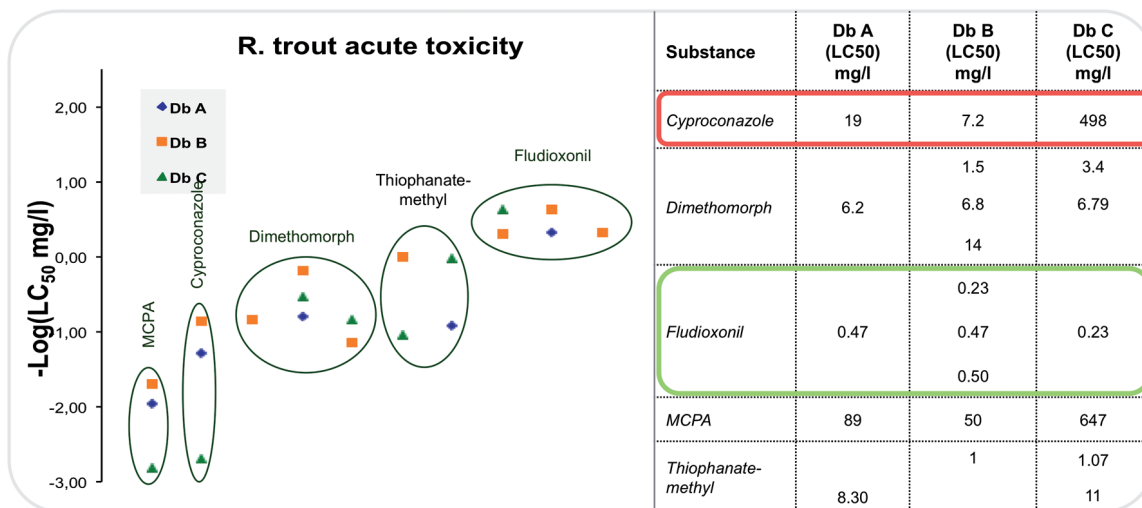


Figure 6: Reproducibility of the experimental results for toxicity studies. Multiple experimentally-determined values for acute toxicity on Rainbow trout (LC50) have been taken for the same chemicals from different databases (in some cases even the same database contains several different values). The graphic shows that differences exist in these values even though they have been obtained using well known and accepted protocols.

Unfortunately, the information regarding the uncertainty of the experimental models is not always available, and often users ignore the fact that this assessment is fundamental. People may think that the lethal dose that kills rats is sharply defined, an absolute threshold which separates a bad or good effect for a chemical, but there is always an uncertainty factor with *in vivo* values.

Furthermore, these values have a probabilistic meaning. Let's consider for instance the lethal dose, which is in fact defined as the dose that kills 50% of the animals. The obvious meaning is that half of the animals die, and 50% don't. Why this happens and what mechanism it is that impacts half but not all of them are unknown. The meaning of the toxicological test is probabilistic for this reason. The content of this experiment, in its statistical nature, is perfectly useful within the risk assessment framework we mentioned above. We know that a certain effect is expected, and the risk assessment procedure will elaborate the value of the possible risk for other situations. The nature of the input is probabilistic, as is the nature of the output.

Of course the uncertainty typical of a given endpoint, when assessed using a single protocol, should be characterised because this affects the uncertainty of the QSAR model. The uncertainty of the final model cannot be inferior to the uncertainty of the input data, and it is suspicious to see values predicted with a precision superior to that of the experimental laboratory model.

The input values should be checked in order to avoid noise. Indeed, it is well known for any model that we cannot extract correct information if we feed the model with poor value: garbage in, garbage out, as it is said. The availability of data from different sources can provide a way to compare and integrate data. It is important to have access to multiple values for the same chemical, and also to know the uncertainty related to a given endpoint.

In the case of toxicity values, some databases are good, other less so. We compared different official databases on pesticides and found differences among reported values [3]. Worse may be the case of data taken from the literature. But this matter is not limited to the property values. We also found many mistakes in the chemical structures reported in journals [4]. All these checks require time and effort, and are not typically done when a QSAR model is applied for academic research. However, in the case of a model to be proposed for regulatory purposes, efforts should be made to quality check the data. For instance, within the CAESAR [5] and DEMETRA [6] projects we spent about one year checking the data, before starting the modelling activities. Some researchers have clearly identified this problem and dedicated efforts to increasing the quality of the toxicity data available. One such researcher is Ann Richard with the DSSTox database [7].

Property data are crucial for further development of QSAR. DSSTox, ECOTOX, and AMBIT are examples of databases. More recently the OECD Toolbox has gathered data on many properties.

An interesting feature of these databases is the availability of toxicity and chemical data/structures together. XML has been identified as the preferred standard within the information technology community for this feature; see for instance the previously mentioned DSSTox database and the EC projects [CHEMOMENTUM](#) [8], [DEMETERA](#), [OpenMolGRID](#) [9] and [OpenTox](#) [10]. The possibility of exchanging data between different databases is also important.

The chemical information: descriptors and fragments

There are two main ways to describe a chemical compound: using global descriptors, or using specific fragments.



Figure 7: Identification of important structural properties and/or fragments. a) QSAR approaches are able to extract, from a training set, the most significant structural properties linked to the specific property to model (in this case: is it a pumpkin or not). b) This element has some characteristics in common with the others but also has some peculiarities.

If we imagine that the pumpkins in *figure 7a* are chemical compounds, we can distinguish them on the basis of their size, shape, colour, etc. However, looking at the pumpkin in *figure 7b* we immediately see that there are peculiar features which make this pumpkin different from any other pumpkin.

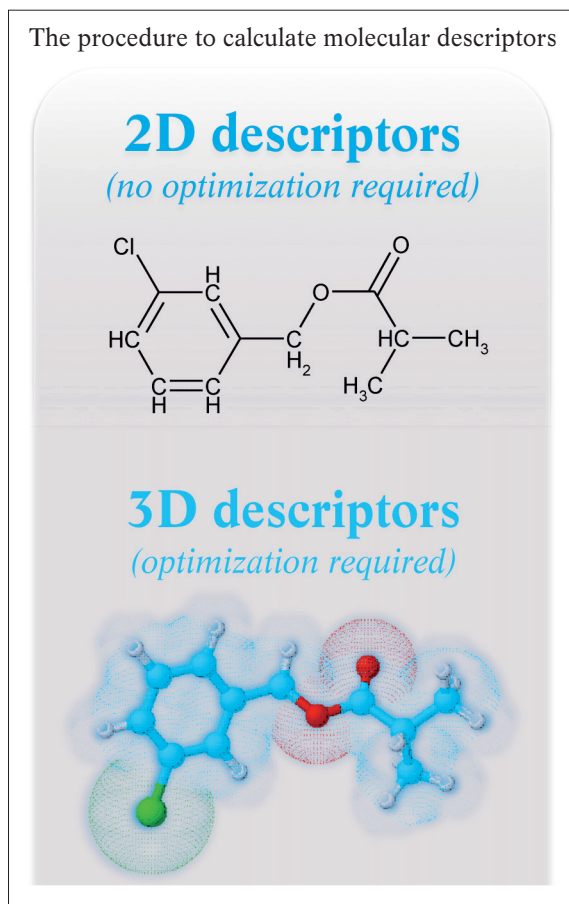
In case of the chemical compounds, some descriptors are global, or general, such as molecular weight or molecular size, while the presence of a certain fragment may also be used to describe a chemical compound. There are many QSAR models using global descriptors, but also a certain number of them using fragments.

The molecular descriptors can be classified as:

- Constitutional descriptors are quite simple; they include molecular weight, number of atoms present in a molecule (*for instance number of chlorine atoms*), number of double bonds, etc.
- Topological descriptors indicate the bonds between atoms, and can be used to represent the ramification of the molecule. Indeed a molecule can be represented as a graph (*ref to annexes: Annex 3. Rigid and flexible topological indices: variety of the representations of the molecular structure*).
- Certain descriptors take into consideration the electronic charge of a certain atom, or its polarity.
- Some descriptors refer to the molecular orbitals of the molecule. Some descriptors calculate the energy of the molecular orbitals, for instance HOMO refers to the energy of the highest occupied molecular orbital, and LUMO refers to the energy of the lowest unoccupied molecular orbital.
- Another kind of descriptor is the so called physico-chemical. They include LogP, lipophilicity, etc. LogP is the logarithm of the partition coefficient between octanol and water. This descriptor has been used since the first QSAR models, and originally it was measured. Nowadays it is much more common to calculate it.
- There are programs which have a list of pre-codified fragments, even thousands of them, and the software checks whether or not they are present in the molecule of interest. There are programs which check for the presence of fragments in a molecule with reference to a list of fragments; these programs are quite fast, and are often used to process huge databases, for similarity purposes. The pharmaceutical industry uses models based on fragments quite often.

A few decades ago, the use of chemical descriptors was very limited. For instance, Corwin Hansch studied ecotoxicity, and put it in relationship with LogP, the partition coefficient between octanol and water, expressed as logarithm. The idea behind this was that the partitioning between water and the organic solvent represents a model for the partitioning between water and the fish body, and the uptake of the toxic compound into the fish body is a good indicator of the toxic effect.

This physicochemical parameter has been used in most of the QSAR models of aquatic toxicity. Slowly, other descriptors have been investigated, in an attempt to better explain certain effects. In particular, further descriptors were introduced to better explain chemical reactivity, molecular size, etc. Nowadays thousands of chemical descriptors can be calculated. Quite often the molecular descriptors are combined. Therefore, dividing one molecular descriptor with a second one, for example, can easily produce a new one.



2D and 3D descriptors

We can also distinguish the descriptors on the basis of the kind of detail needed to represent the molecule. A major difference is between descriptors which need a tri-dimensional (3D) representation and other descriptors. Indeed, some descriptors, such as the number of certain atoms, or topological descriptors, do not need a 3D representation of the molecule. Conversely, descriptors like molecular volume of quantum-mechanical molecules require a 3D representation of the molecule.

In the case of 2D descriptors, the molecule can be represented flat. In case of 3D descriptors we need a 3D representation (*figure 8*). Most typically the 3D representation has to be optimized, and this is done manually. Indeed, there are many different conformations that the molecule may exhibit, depending on the rotations of the bond and the angle between bonds. On the basis of the different conformation, many 3D descriptors can vary, such as molecular volume, length, etc.

Thus, the values of the 3D descriptors typically change depending on the user, the software, the

Figure 8: Molecular descriptors can be of different levels of complexity. The so-called 2D descriptors (e.g. topological information) do not need any conformational information regarding the molecule. The 3D descriptors (e.g. charge distribution) need the molecule to be optimized, creating a series of problems, from increased time for calculation, to more difficult reproducibility.

approximations, etc. Furthermore, the calculation of the 3D descriptors takes more computer time for the necessary optimization of the values.

For these reasons, the reproducibility of the 3D descriptors is lower, compared to that of the 2D descriptors. Thus, if we want fast and reproducible QSAR models, 3D descriptors may be counterproductive.

However, 3D descriptors have some advantages, and can provide better results for specific cases, such as models on restricted chemical classes.

Most typically, the variability and the uncertainty of the experimental values of the property to be predicted is so large that improvements related to the use of 3D descriptors are negligible. For instance, in the case of global models for toxicity developed within CAESAR, we found that the statistical performance of the models based on 2D descriptors was the same as the models based on 3D alone, or 2D and 3D descriptors combined.

We also notice that 2D descriptors may suffer from poor reproducibility. Certain parameters, which are related to the representation of the molecular bonds and the tautomers, may vary the output of the 2D descriptor. This is the case for the number of double bonds. Different numbers may be obtained if we calculate certain bonds as aromatic bonds, or as double bonds. For instance, the benzene ring may be represented as a ring with three double bonds and three single ones, or as a ring with six aromatic rings. While in the case of benzene rings this can be easily solved, it may be more critical for heteroaromatic rings, because there it may be more difficult to distinguish if a bond is aromatic or not and this often depends on the formalism and the conventional approach which is adopted.

Also in the case of tautomers, certain 2D descriptors may change depending on which tautomer is used.

Thus, if we want to have reproducible results in QSAR models it is necessary to use the same software to calculate the chemical descriptors, and to use the same format to represent the chemical structure. For instance, an error which may occur if different formalisms are used is related to the representation of the nitro group.

The way to represent the chemical structure

Before calculating the chemical descriptors or fragments, the chemical formula has to be represented in a suitable way.

There are several ways to do this. Typical ways are the InChI [11], SMILES [12] or sdf¹ format.

InChI is univocal, but so far less used than SMILES, which is simpler. However, care should be taken with SMILES because there may be more than one SMILES for the same chemical. Canonical SMILES are recommended for this reason. Nevertheless, there are different formalisms to write the structure within SMILES; for instance for the nitro group and the kind of bonds between the oxygen and nitrogen. Indeed, the bond between N and O can be written as a double bond or with a separation of charges: N=O or N+O-. The chemical may be read in different ways, generating different results. Thus, the user should not mix SMILES which have been taken from different sources. Other formats, such as sdf, codify the graph of the molecule, and may contain additional information besides the chemical structure. Examples of different formats are in *figure 9*.

The choice of the suitable complexity in the chemical descriptors and chemical format

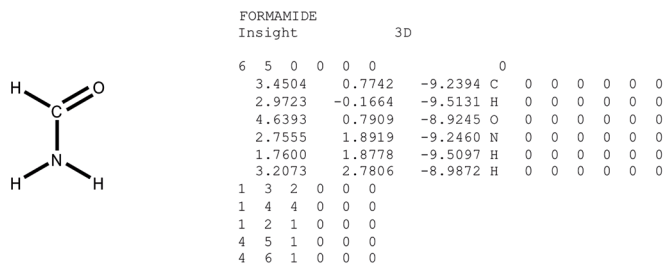
The choice of how to represent the chemical through an appropriate format and structure has to be related to the purpose. Indeed, it depends not only on the output information, but also on the information of the chemicals which necessarily is at the basis of the model.

In most models useful for REACH, related to typical industrial chemicals, the substances are not pure enantiomers. It is quite difficult to find two experimental values specific for the two enantiomers. Thus, in order to verify the appropriateness of the substance for the specific case, and whether the predictive model has been developed on the basis of sufficiently specific data, the detail of the chirality of the substance should be checked.

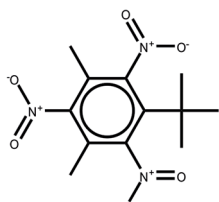
¹ The Structure Data Format (SDF) file is a sort of text-based database which can contain (in a standardized way) information about structure and properties of several chemical substances.

Different type of computer-readable ways to represent molecules

.mol Format (Macromodel)



SMILES (Simplified Molecular Input Line Entry Specification)



Using ASCII strings for depicting chemical information!!

O=[N+](=[O-])c1c(c(c(c1C)[N+](=O)[O-])C(C)C)[N+](=O)[O-]C

Figure 9: Examples of representation of molecules for software input. Different methods have been developed throughout the years to represent and archive structural and conformational information of molecules for computer utilisation purposes. The mol format saves the type and position of each atom as well as the topological information explicitly; this format is complex but defines the molecule clearly. The software has only to place the atoms in the defined position and link them with the defined bond. The simplest way to represent the molecules is the Simplified Molecular Input Line Entry Specification (SMILES); SMILES are text strings formatted in a way which allows the software to know how the atoms are linked to each other; the position is then calculated on the basis of the topology.

The user may wish to get the maximum detail and information on the chemical structure, chirality, etc. However, the real information should refer to the chemical which has been used for the experiment, and thus if it is a mixture of enantiomers, this should be used. Furthermore, the analysis of the results may be limited by the lack of other, related chemicals with the same detail of information. Indeed, if we want to build up a model for enantiomers, we need a series of cases, not just a few. However, as we have explained this may be very difficult.

Software for chemical descriptors calculation

There are several commercial programs for chemical descriptors, but for a few years, some free programs have been available on the internet.

- T.E.S.T. (<http://www.epa.gov/nrmrl/std/qsar/qsar.html#TEST>)
- Dragon (used to be a freeware, but this version has not been updated. Dragon is now a commercial software)
- VEGA (<http://www.vega-qsar.eu/>)
- OCHEM (<http://ochem.eu/>)
- CDK (ref to annexes: Annex 4. [A free and open source informatics library for chemistry: Chemistry Development Kit \(CDK\)](#))

Furthermore, there are many commercial tools:

- Accord for Excel uses Accord Chemistry Engine to handle chemical structures and incorporates a number of add-ins to perform chemical calculations based on the structures of a compound in a record.
- ADAPT is a QSAR toolkit with descriptor generation (*topological, geometrical, electronic and physico-chemical descriptors*), variable selection, regression and artificial neural network modelling.
- CODESSA has been developed for the calculation of several topological, geometrical, constitutional, thermodynamic, electrostatic and quantum-chemical descriptors. It also includes tools for regression modelling and variable selection.
- DRAGON has been developed for the calculation of several sets of molecular descriptors from molecular geometries (*topological, geometrical, WHIM, 3D-MoRSE, molecular profiles*).
- GRIN/GRID calculates the GRID empirical force field at grid point.
- HYBOT-PLUS has been developed for the calculation of hydrogen bond and free energy factors.

- MOLCONN-Z, successor to MOLCONN-X, MOLCONN-Z, calculates the most known topological descriptors, including electrotopological and orthogonalised indices.
- POLLY has been developed for the calculation of topological connectivity indices.
- SYBYL/QSAR has been developed for the calculation of EVA descriptors, CoMFA and CoMSIA fields. It also includes several QSAR tools.
- TSAR is characterized by statistical and database functions with molecular and substituent property calculations.

More on chemical descriptors

See these for more info on descriptors:

- <http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml> for a list and classes of molecular descriptors;
- http://www.chem.ucsb.edu/~kalju/chem162/public/molecules_qsar.html for a brief intro on molecular descriptors.

Moreover, there are some books or chapters on chemical descriptors:

- Todeschini R. & Consonni V., Handbook of Molecular Descriptors, J. Wiley & Sons, New York, 2008, 688 pp.
- Karelson M., Molecular Descriptors in QSAR/QSPR, J. Wiley & Sons, New York, 2000, 430 pp.
- Basak S. C., Grunwald G. D. & Niemi G. J., Use of Graph Theoretical and Geometrical Molecular Descriptors in Structure-Activity Relationships, in From Chemical Topology to Three-Dimensional Geometry, Plenum Press, New York, 1997, pp. 73-116
- Benfenati E., Casalegno M., Cotterill J., Price N., Spreafico M., and Toropov A., Characterization of chemical structures, in: Benfenati E. (Ed.), Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier Science Ltd, Amsterdam, The Netherlands (2007), 83-109.

The modelling algorithms: classifiers and regression models

In the last decades in addition to the thousands of chemicals descriptors that have been made available, many advanced, powerful modelling algorithms have also been developed. The older QSAR models were linear equations with a few parameters. Then, other tools were introduced, such as artificial neural network, fuzzy logic, and data mining algorithms, making possible non-linear models and automatic generation of mathematical solutions [13-14].

We can distinguish two kinds of algorithms: those for regressions and those for classification. Regression methods get a continuous value. Classifiers find the category, e.g. the toxicity class. *Figure 10* lists some common classifiers and regression models used in QSAR.

In fact, in the case of classifiers the appropriate definition would be SAR, since the purpose of classifiers is not to obtain a quantitative evaluation, but a category. In some cases the categories refer to different thresholds, such as the toxicity classes for acute toxicity for mammals. We may

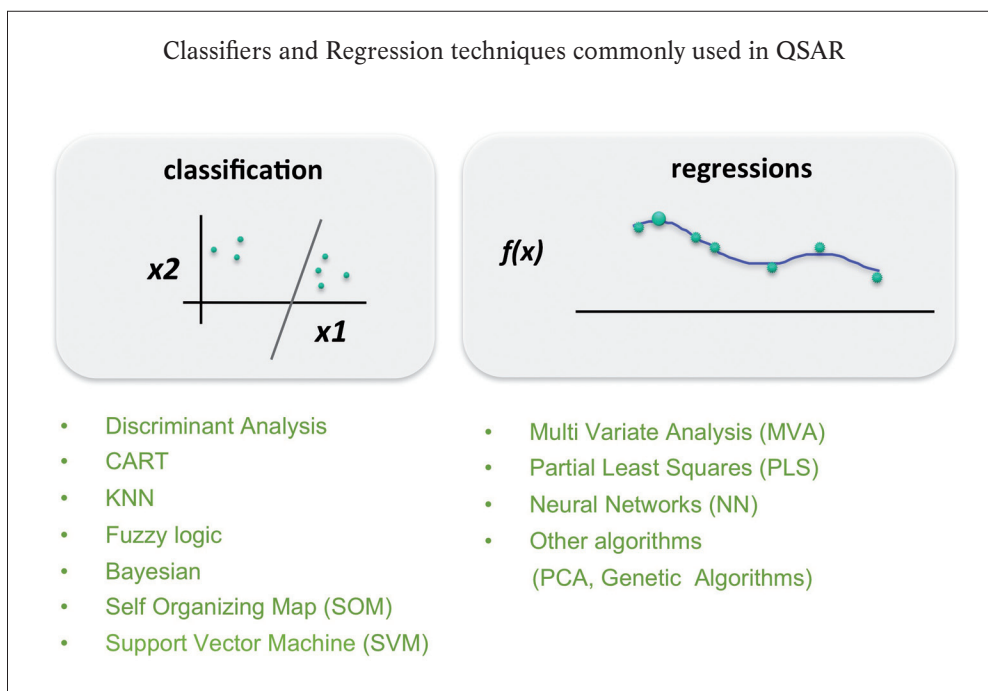


Figure 10: Mathematical and statistical techniques commonly used to build QSAR models.

have a model using molecular descriptors, even a quantitative one, and a continuous output of the model, like the acute toxicity value, which only at the last stage of the model takes this continuous value to obtain predictions for categories.

Thus, the definitions of classifiers as SAR and QSAR, and the distinction between regression or classifier algorithms are useful as general rules, to describe the general boundaries of the methodology. However, there are cases where the boundaries are indistinct. With the fuzzy logic approach it is possible to go from one paradigm to another quite easily. Even with other mathematical methods it is possible to talk about the probability that a certain compound is toxic or not, and thus to think about a classification problem (*difference between toxicity 0 = not toxic, and 1 = toxic*) using a continuous scale: for example, the probability that a certain chemical is toxic (*100% probability*) or not toxic (*0% probability*).

Different algorithms

There is a variety of methods for building QSAR models. These methods are called pattern recognition methods because their aim is to devise algorithms that could learn to distinguish patterns in a data set. They can be classified as supervised (*for example, Multiple Linear Regression, Discriminant Analysis, Partial Least Squares, Classification and Regression Trees, Neural Networks, etc.*) or unsupervised (*for example, Principal Component Analysis, Cluster Analysis, k-Nearest Neighbours, Nonlinear Mapping, etc.*), where supervision refers to the use of the response data which is being modelled. Unsupervised learning makes no use of the response, meaning that the algorithms seek to recognize patterns in the descriptor data only. The advantage of unsupervised learning is the lower likelihood of chance effects, due to the fact that the algorithm is not trying to fit a model. On the other hand, supervised learning does use the response data and care needs to be taken to avoid chance effects. Another significant difference between supervised and unsupervised learning methods is the ratio of compounds (p) to variables (n) in a data set. When $n \geq p$, some supervised learning techniques may not work due to the failure to invert a matrix, while others may give a false, but apparently correct, classification. Even though this is not a problem for unsupervised methods, the presence of extra variables that have no useful information may obscure meaningful patterns.

The nature of the response data that they are capable of handling is another important feature of modelling methods. In this context, there are two types of methods: those that deal with classified responses (*for example, mutagen / not mutagen, toxic / slightly toxic / non-toxic*) and those that

handle continuous data (*the response is a potency of an end-point*). For the modelling of categories, a wide range of classification methods exists, including: Discriminant Analysis, k-Nearest Neighbours (*KNN*), Classification and Regression Trees (*CART*), Support Vector Machine, etc. For the modelling of continuous data, the most widely used method is Multiple Regression Analysis (*MRA*), a simple approach that leads to an easily understandable result. MRA is a powerful means of establishing a correlation between independent variables (*molecular descriptors*) and a dependent variable (*biological activity*). The main characteristics of MRA are:

- Linear relationship between Y and several Xi descriptors;

$$Y = aX_1 + bX_2 + cX_n + \dots + \text{const.}$$

- Errors are minimized by least square;
- Polynomial terms may be included.

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

In addition, Artificial Neural Networks (*figure 11*) can be used for modelling both classified and continuous data. The main characteristics of ANN are:

- The structure is inspired by biology;
- ANNs are a set of connected nonlinear elements making transformation of input.

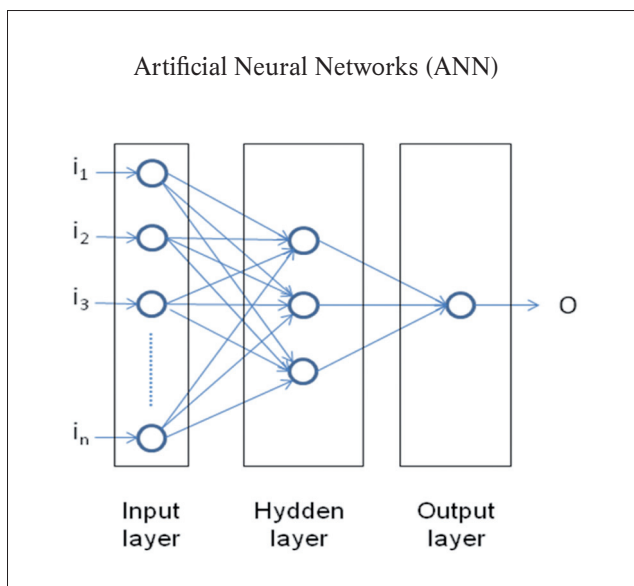


Figure 11: The organization of Artificial Neural Networks is similar to that of biological systems. The elements of the network (called neurons) are organized in layers and interconnected in a way that the output of a neuron is the input of the subsequent one.

Free tools & algorithms for QSAR modelling

Free and open sources tools and algorithm have been developed and currently maintained to build QSAR models. Some examples of tools developed specifically for building QSAR models are:

- OCHEM (<http://ochem.eu/>)
- CORAL (<http://icnanotox.org/2010/coral-correlation-and-logic-software/>)
- SARpy [15]
- OpenTox (<http://www.opentox.org/>)

Other more general free resources, including mathematical and statistical approaches which can be used for QSAR modelling are:

- WEKA (<http://www.cs.waikato.ac.nz/ml/weka/>)
- R (<http://www.r-project.org/>)

The way that a QSAR model is built up

In some cases, especially in case of genotoxicity models, the human expert identified fragments which can be related to the genotoxic effect. For instance, it is known that nitrosoamines are genotoxic. The visual examination of a series of chemicals sharing the same fragment may be used for this purpose. In this case the effect is simply the toxic effect (*genotoxic or not, for instance*), and the chemical information is simply the fragment. The algorithm is, in this case, the rule. Expert systems have been built up in this way. Examples of this kind of model include

- HazardExpert (<http://www.compudrug.com/>)
- Derek (<https://www.lhasalimited.org/>)
- Toxtree (<http://toxtree.sourceforge.net/>)

Most typically a QSAR model is built up starting with a set of chemicals with known property values.

The very first step in QSAR modelling is the translation of the structural information in some numerical values (*the molecular descriptors*).

Therefore, conceptually, table of all the chemical compounds is built up. For each chemical one needs the descriptors and the property values of that set of compounds. The compounds are typically arranged in a column, and there are several columns for the chemical descriptors, as well as one column for the property value. The chemical descriptors are the x (*the input*) of the model, and property is the y (*the output of the model*).

However, a major problem may result from the uncritical use of powerful mathematical tools; the risk is that the model does not working when applied to new compounds, because it is only capable of replicating the toxicity of the chemicals used to train the model. In model development the procedure is to use some chemical compounds with a known toxicity. These compounds are used as training set. Then, using chemical parameters and a suitable algorithm, the model is developed.

However, to check if the model is really a predictive one, an assessment has to be done. This obvious consideration applies to all kinds of models, the simple ones with a single parameter, and the more complex ones. The risk of chance correlation is higher when a high number of

descriptors or parameters is used and when few examples or molecules are used. This may also lead to over-fitting, a phenomenon in which the model gives high performance on the training set, but then results decrease dramatically when the model is applied to new chemicals, such as in an external validation set (as shown in figure 12).

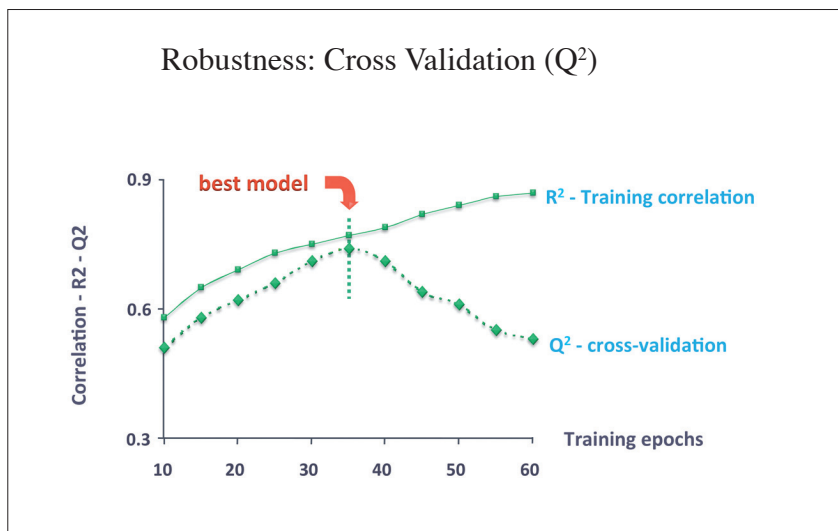


Figure 12: QSAR models have to be tested on molecules not used to build them in order to test their real predictive power. Cross validation consists of leaving out N compounds from the training set and rebuilding the model. The N compounds are then used as test set. This procedure is iterated many times. The parameters R^2 and Q^2 are both measure of correlation values; the difference is that Q^2 refers only to the compounds outside the training set of the model.

This leads to the need for dimension reduction, variable elimination and variable selection, which are different techniques for reducing the complexity of a problem in order to be able to recognize useful and informative patterns in the data. Dimension reduction is the process of reducing the number of random variables under consideration and is usually performed by a mathematical procedure called Principal Component Analysis (PCA) in which new variables, called principal components, are created from linear combinations of the original variables. Variable elimination is the process by which unhelpful or unnecessary variables are removed from a data set. Common procedures for variable elimination are Corchop [17] and unsupervised forward selection. Even after eliminating unnecessary variables from a data set, there may still be many variables to choose from when building a model. In this case variable selection is used, the aim of which is to choose descriptors that will be useful in some sort of mathematical model and will lead to a model that will generalize to other unseen compounds. There are many diverse procedures for variable selection and some are built in to the process of model building, such as forward stepping multiple regression.

The validation of the model

The models have to be validated. This simple statement may appear obvious today. QSAR models developed decades ago mainly addressed the discovery of certain relationships between a given parameter and the effect of interest. For instance, it was satisfactory to find that there was a linear relationship between LogP and aquatic toxicity. Furthermore, we can imagine a model aiming to understand the biochemistry of a given effect, in which all chemicals have known toxicity values; but we want to understand why. In this case the prediction is not for the property/effect, but for the process underlying the effect. Conversely, models for regulatory purposes require stringent validation procedures on the prediction of the property/effect. Sometimes, these two different purposes are confused, but they are conceptually different, as the above example shows.

For regulatory purposes there must be proof that this relationship applies to the prediction of the properties of *new* chemicals; thus, this has to be specifically addressed. This statement puts emphasis on the statistical validation of the model. This is clearly addressed within the five OECD principles for QSAR [18]. New statistical tools and evaluation procedures have been introduced, compared to the simple fitting measurement based on the training set. The importance of an external test set has been stressed in many cases. If the total number of compounds is low, this imposes limitations on the external validation. Internal validation is in any case recommended, and to this end a number of tools has been developed, such as leave-one-out, y-scrambling, etc. Leave-One-Out Cross Validation (*LOO or CV*) involves leaving out one compound, fitting the model to the remainder of the set, making a prediction for the left out compound and repeating the process for each of the compounds in the set. A variety of statistics can be generated using this procedure, for example LOO R^2 (called Q^2) and a predictive residual sum of squares (*PRESS*).

The disadvantage of LOO is that only a small part of the data set is omitted and if outliers occur in pairs or groups they will not be identified. A better approach is to leave out some larger portion of the set (*10 or 20%*) and to repeat this a number of times. This allows the generation of a set of predicted values for the compounds so that estimates may be made of the likely errors in prediction. The disadvantage of this approach is that it is computationally intensive and suffers from a combinatorial explosion as the sample size is increased. Some examples of techniques for the model validation are summarized below.

Table 1: Common approaches used to validate QSAR models.

| <i>Cross validation</i> | <i>Bootstrapping</i> | <i>Y-scrambling</i> |
|--|--|---|
| <p>Leave One Out</p> <ul style="list-style-type: none"> All the data are used for fitting except for one compound Predict the excluded sample Repeat it for all samples Calculate Q^2 or R^2_{cv} similarly to R^2 on the basis of these predictions <p>Problem: this approach may be too optimistic if there are many samples</p> <p>Leave Many Out</p> <ul style="list-style-type: none"> Use larger groups to obtain a more realistic outcome | <ul style="list-style-type: none"> Bootstrapping simulates what happens by randomly re-sampling the data set with n objects K n-dimensional groups are generated by a randomly repeated process which eliminates some objects The model obtained on the different sets is used to predict the values for the excluded sample From each bootstrap sample the statistical parameter of interest is calculated The estimation of accuracy is obtained by the average of all calculated statistics | <ul style="list-style-type: none"> Randomly permute Y responses while X variables are kept in the same order for several times |

Evaluation of a classifier

Typically, classifiers are evaluated using the Cooper statistic. In the simple case of a binary classification, there are two classes, such as toxic (*positive*) or not (*negative*). The results of a classifier could be therefore grouped in four categories: toxic compounds predicted as toxic (*True Positive or TP*) or as non-toxic (*False Negative or FN*) as well as non-toxic compounds predicted as non-toxic (*true negative or TN*) or as toxic (*False Positive or FP*). These four classes are usually represented in the so-called *confusion matrix* (as shown in figure 13).

Three main statistical parameters can be derived from the combination of these four cases for model evaluation:

Accuracy (A), also referred as concordance, is the measure of the correctness of prediction. This parameter gives a general evaluation of the errors made and is defined as the ratio between the

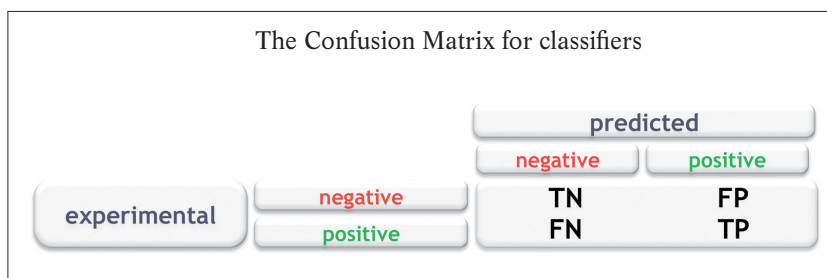


Figure 13: The confusion matrix of a binary classification. Positive classification is usually associated with active compounds (e.g. toxic) whereas negative is associated to inactive ones (non-toxic). Compounds correctly predicted are called True Positives (TP) or True Negatives (TN) depending on whether they are active or inactive. Active compounds predicted inactive are referred as False Negatives (FN), whereas inactive compounds incorrectly predicted are False Positives (FP).

compounds correctly predicted and the total number of compounds. A good model has high accuracy value.

$$A = (TP + TN) / TOTAL$$

Sensitivity (S) is the measure of the positive compounds correctly predicted. Especially for regulatory purposes, it is important not to declare safe a chemical which is actually toxic (FN). Sensitivity takes into account the number of FN and is defined as the ratio of the TP tests to the total number of positives. A good model has high sensitivity.

$$S = TP / P$$

Specificity (SP) is the measure of the negative compounds correctly predicted. Specificity takes into account the number of false positives and is defined as the ratio of the TN tests to the total number of negative compounds. Sometimes the $1 - SP$ parameter is reported.

$$SP = TN / N$$

It is our opinion that for regulatory purposes it is important to verify that the classifier has a high sensitivity, in order to reduce the number of false negatives.

Binary classifications are not all that is defined within REACH. For instance, a chemical can be *not bioaccumulative*, *bioaccumulative* or *very bioaccumulative* (three classes).

Evaluation of a regression model

Regression models are most typically evaluated using statistical parameters which take into account the errors of the model. These errors are measured on the basis of the training set, and this gives an idea of the model robustness. However, this is not sufficient since the main interest of REACH is to understand if a certain model can be used for prediction purposes. Thus, for regulatory purposes, additional statistical measurements are used for prediction. Some measurements use internal validation whereas other tools refer to an external test set.

The values predicted by the model (*on training, test and/or external validation set*) are put in correlation with the experimental values using a graph and then the coefficient of determination (R^2) is calculated giving an estimation of the model goodness.

In the case of classifiers the emphasis is on false positives and negatives. However, recently we underlined the importance of also paying attention to false negatives for regression models [3]. Indeed, regulators pay much more attention to false negatives. With regard to models for regulatory purposes, attention can be paid to false positives within a wider strategy in which intelligent testing methods are used, taking into account the sensitivity and selectivity properties of each individual element in the combined strategy. The usual approach to evaluating regression methods, using R^2 , clearly shows that we do not take into account whether or not the error has a sign, because we use the square. However, we should evaluate this. This was addressed in the EC projects DEMETRA and later on in project CAESAR.

The validation of the QSAR models and non-testing methods and regulatory needs

In the classical QSAR models it was assumed that a convincing explanation of the phenomenon was sufficient. The interest in the possible use of such a model to predict the properties of related

compounds was not the declared target of the study, and the statistical proof of the predictive power of the model was usually not sufficiently checked. Indeed, the older QSAR models were based only on the fitting description of the mathematical equation.

We emphasise that the target for these studies was not the prediction of the property of the chemical compound, but the understanding and modelling of the mechanism at the basis of the phenomenon. In these kinds of studies, all the property values of the set of compounds were known. What was unknown was when a certain phenomenon occurred. Even today there are studies being conducted to explore the possible reasons for the occurrence of a certain phenomenon and the emphasis may be placed on exploring the mechanism. Additionally, if we want to shift the model to the predictive field, we need appropriate ways to validate our model.

Another reason to develop or use QSAR models is to predict the property values of a certain chemical. In this case, the model is built on the basis of the property values of other chemicals, with known values, and the model calculates or derives the property value of the chemical on the basis of certain rules. In order to use a model for this purpose, a suitable check of the predictive performance of the model has to be done, as explained in the paragraph “The validation of the model”.

If the predictive performance of models is not statistically checked, there is a serious risk in using the model to predict a property value. The careful check of the predictive power is one of the principles for the correct use of the QSAR models defined by the OECD [29].

It is possible that a model aims both to predict the mechanism at the basis of the phenomenon and also to predict the property of a target chemical. Indeed, in a certain way all models would aim to achieve both of these goals. The main difference is the emphasis placed on one aspect over another; is the plausible explanation of the phenomenon, or the predictive power of the model more important? Indeed, a model can be optimized and carefully checked towards the suitable, high predictive power. However, the components of this model may be not fully understandable. This may be the case with integrated models, which combine different models (*link to section hybrid models*). It has been theoretically described and experimentally proved that these models are more robust and predictive. However, since they refer to different parameters which are statistically optimized, their interpretation is more difficult than simpler models based a few rules.

On the other hand, models which are based on plausible mechanisms may be more easily accepted. However, we have to remember that these models also have a statistical basis: according to what is

observed from related compounds, a certain mechanism can be *hypothesized*. This does not mean that the specific mechanism will occur for the target compound.

Unfortunately, phenomena occurring in nature and life are very complex. Currently we are capable only of gathering limited information on them.

In information technology, there is a concept of explicit and implicit knowledge. Explicit knowledge is knowledge which has been already codified into explicit rules. This is the case of QSAR models where, for instance, some fragments associated to carcinogenicity have been identified. However, these lists are not univocal; several of them exist, but not all fragments have been identified and thus there are chemical which are carcinogenic which are not recognized as toxic (*false negatives*). Furthermore, there are chemicals showing a carcinogenic fragment which are not toxic (*false positives*).

On the other hand, there are models based on the data, which extract the knowledge directly through a process of data mining and knowledge engineering. In some cases these models showed higher performance than the models based on explicit knowledge, due to the fact that they can incorporate information which has not yet been codified by human experts [15,16].

Hybrid models

As previously mentioned, it is likely that more than one model exists for the same endpoint. Thus, we may face the issue of how to compare the results. We may adopt simple or sophisticated strategies.

A simple approach is to take the worst case scenario. If we have two models, even if only one of them predicts the chemical as toxic, we can follow the precautionary principle and assume the toxicity of the chemical. As an alternative, we can adopt the approach of the majority vote, in case of several results.

If we are dealing with continuous values we could process them from a statistical point of view and take the average, or another statistical parameter. Alternatively, we could use the results only when they match, and disregard results where there is a conflict.

None of the above approaches take into consideration the reliability of the individual models, which may be different. Thus, a more sophisticated way is to take into account the model reliability. A possible way to handle this is through the Bayesian system. Such an approach has been adopted for instance within the OSIRIS project, to integrate results from different models. An even more sophisticated approach also takes into account the individual results on the basis of the chemical, for the individual models.

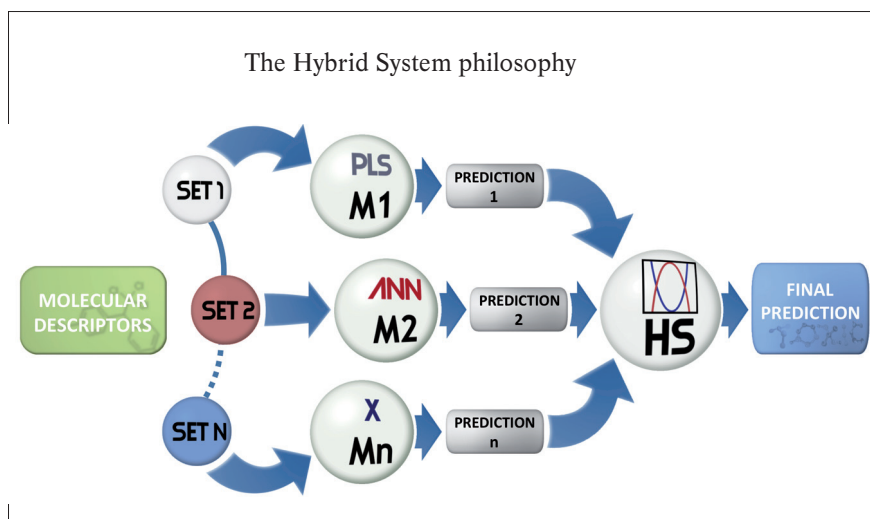


Figure 14: Multiple models can be utilised to obtain evaluation which will be then used as input by another model in a hybrid system approach.

This strategy has been adopted in the DEMETRA project, which developed and integrated a series of models [1]. The strategy of using the results of several QSAR models as input for a final model is depicted in *figure 14*.

Another example is the T.E.S.T. model, which contains integrated models, called consensus models. The consensus model takes into account the results of the reliable models only, and this depends on the chemical compound.

CAESAR also developed hybrid models, which integrated the results of different models, depending on the chemical. Here we can discuss two different examples, adopting two strategies.

The CAESAR model for bioconcentration (*BCF*) is based on two separate models, which are combined. These two models use different descriptors, and the algorithm using the descriptors is different. Thus the two models provide different results. The results of these models are used by a third model, the hybrid one, which uses the output of the two models instead of the chemical descriptors [4,5]. The strategy is similar to what represented in *figure 14*.

The architecture of a multi-step hybrid system
 Three models are used in series to predict the mutagenicity of molecules

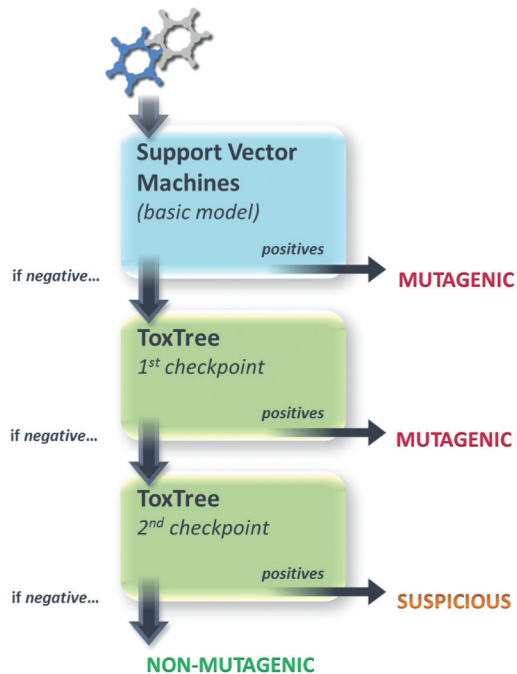


Figure 15: A three-steps hybrid system for mutagenicity prediction.

Performance of the hybrid system compared to each model alone

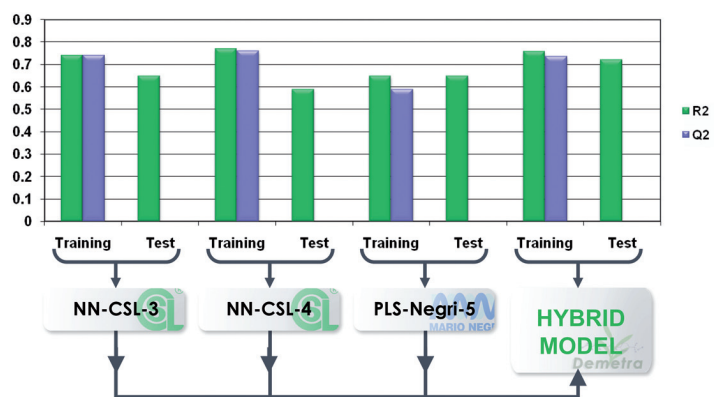


Figure 16: Hybrid system predictive performances compared to the single models.

The CAESAR model for mutagenicity uses two models in cascade (*figure 15*). The first model makes the prediction. If the output is “non-mutagenic”, at this point the chemical is processed by a second model. This is repeated once again, with a third model. At this point the output of the third model is: “non-mutagenic” or “suspicious”. Thus, in this case we have a sequence of models, which are switched depending on the output of the previous model.

These two strategies are quite well suited for two different purposes: a continuous output, or a category.

As a general comment, we notice that it is preferable to have combinations of models which are based on different approaches. This maximizes the exploitation of the approach. *Figure 16* shows the improvement of the results, using the hybrid models.

Further perspectives on in silico models

Models to explore biochemical mechanisms may follow different methods from those mentioned above, which address the prediction of the activity, and not necessarily of the mechanism. Expert modellers can explore complex situations using all their experience and subjective processes, even though the possibility of disseminating the procedure will be limited. Industry can of course use its own confidential data for internal purposes and the model will not suffer because of these conditions.

The scenario of QSAR models is very broad. Many techniques exist, which are not, strictly speaking, classical QSAR models. Methods such as docking offer the possibility of studying the interactions between a ligand and the receptor. Methods such as COMFA can investigate parts of the molecule which are involved in the toxicity process [19]. While QSAR models explore the hypothesis of a relationship between a certain chemical descriptor and the property, docking allows the introduction of specific knowledge to do with the biochemical environment in which the chemical should be active. Forces affecting the binding are used for modelling. The model is suitable when the property of interest is mediated by the binding, and is very appealing in its capability to show the direct biochemical interaction. However, in the event that the process is more complex, and several steps are involved, binding alone may overlook important parts of the phenomenon to be studied.

Models such as COMFA are useful to identify the steric and electrostatic factors of the molecule which affect the process, showing the specific parts of the molecule where this occurs. There are examples of the useful integration of different methods, to better explore the toxicity phenomena and the factors involved in the phenomenon.

Chapter references

1. Benfenati E., Clook M., Fryday S. and Hart A., QSARs for regulatory purposes: the case for pesticide authorization, in: Benfenati E. (Ed.), Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier Science Ltd, Amsterdam, The Netherlands (2007), 1-57
2. Dimitrov A., Dimitrova N., Parkerton T., Comber M., Bonnell M. & Mekenyan O., Base-line model for identifying the bioaccumulation potential of chemicals, SAR QSAR Environ Res. 2005 Dec;16(6):531-54.
3. Benfenati E., Boriani E., Craciun M., Malazizi L., Neagu D., Roncaglioni A., Databases for pesticide ecotoxicity, in: Benfenati E. (Ed.), Quantitative Structure-Activity Relationships (QSAR) for Pesticide Regulatory Purposes, Elsevier Science Ltd, Amsterdam, The Netherlands (2007), 59-81
4. Zhao C., Boriani E., Chana A., Roncaglioni A. and Benfenati E., A New Hybrid QSAR Model for Predicting Bioconcentration Factor (BCF), Chemosphere 2008, 73:1701-1707.
5. EC funded project CAESAR (Computer Assisted Evaluation of industrial chemical Substances According to Regulation) - <http://www.caesar-project.eu/>
6. EC funded project DEMETRA (Development of Environmental Modules for Evaluation of Toxicity of pesticides Residues in Agriculture) - <http://www.demetra-tox.net/>
7. US EPA DSSTox (Distributed Structure-Searchable Toxicity) Database Network - <http://www.epa.gov/ncct/dsstox/>

8. EC funded project CHEMOMENTUM - <http://www2.fz-juelich.de/jsc//grid/Chemomentum/>
9. EC funded project OpenMolGRID (Open Computing GRID for Molecular Science and Engineering) - <http://www.openmolgrid.org/>
10. The OpenTox Framework - <http://www.opentox.org/>
11. The IUPAC International Chemical Identifier (InChI) - <http://www.iupac.org/home/publications/e-resources/inchi.html>
12. Daylight SMILES (Simplified Molecular Input Line Entry System) - <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>
13. Devillers J., Genetic Algorithms in Molecular Modeling (Principles of QSAR and Drug Design), Elsevier Science Ltd, Amsterdam, The Netherlands, 1996, 327 pp.
14. Devillers J., Neural Networks in QSAR and Drug Design (Principles of QSAR and Drug Design), Academic Press, London, UK, 1996, 284 pp.
15. Ferrari T., Gini G., Bakhtyari N.G. & Benfenati E., Mining toxicity structural alerts from SMILES: A new way to derive Structure Activity Relationships, *in*: Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on, 2011, pp. 120-127.
16. Multicase Inc. Bioactive Software - <http://www.multicase.com/>
17. Livingstone D.J., Rahr E., Corchop - an Interactive Routine for the Dimension Reduction of Large QSAR Data Sets, *Quant. Struct.-Act. Relat.* 2006, 8(2):103-108.
18. OECD Principles for the Validation, for Regulatory Purposes, of (Quantitative) Structure-Activity Relationship Models. Paris, France. <http://www.oecd.org/dataoecd/33/37/37849783.pdf>
19. Roncaglioni A., Benfenati E., Computer-aided methodologies to predict endocrine-disrupting potency of chemicals *in* Shaw I (Ed.), CRC Press, Boca Raton (2009), 306-321.

Practical case

Many tools, many purposes, many QSAR models

No single QSAR strategy will be sufficient to address all toxicological endpoints. For example, some models call for binary outcomes, such as a chemical is or is not carcinogenic, or is or is not a skin sensitizer. In other cases the classification will depend on the dose and result in categorical outcomes, such as different toxicity levels expressed as low, medium, and high. In other cases the effect or property is expressed as a continuous value, such as the toxic dose.

We should consider that the same endpoint might require a continuum value or a class, depending on the regulation, or depending on the purpose of the evaluation. For instance, for classification and labelling purposes the class is suitable, while for risk assessment, when we have to evaluate an effect for a certain exposure concentration, a continuous value is the proper parameter in order to compute the risk.

To obtain predictions as to classes or continuum values, different algorithms are typically used. Also the way to describe the chemical structure may be different. For instance the presence of certain residues in the molecule has been used to identify carcinogenic compounds, while continuous

chemical descriptors are typically used to predict continuous properties. To get more information, refer to the Part A of this book.

A further matter related to endpoint complexity is that for certain endpoints certain *in silico* models are used, but the same technique does not apply to other endpoints. For instance, for endocrine disruptors models, docking software is often used. These models have some advantages because they clearly relate to the knowledge of the binding of a ligand to a receptor, such as the estrogen receptor. In this case there is a good level of knowledge of the structure of the receptor, and the mechanism is known. Thus, predictions can be made, based on this knowledge. However, in many other cases such detailed knowledge of the receptor basis of the effect is not available, and this approach cannot be used.

The consequence of such a complex picture, with many endpoints and models based on different approaches, is that it is not realistic to have a single approach. The use of models should be modulated considering these issues. Different approaches can be used for the same target to improve the reliability of the overall prediction. In some cases they have been combined into a unified system, to improve overall results.

Here we describe some cases, providing examples. However, specific assessments have to be done for specific models and specific endpoints.

A series of existing models

QSAR models have been developed for many applications:

- Physico-chemical properties, such as boiling point, water solubility, partition coefficients. For instance, the U.S. EPA Estimation Program Interface (*EPI*) suite comprises models for these endpoints [1].
- Environmental properties, such as bioconcentration factor, degradation, soil adsorption, photo degradation.
- Ecotoxicological properties, such as fish, daphnia, bird, bee toxicity.

Toxicities, such as carcinogenicity, mutagenicity, developmental toxicity and skin sensitization have been modelled by the EC funded project CAESAR [2] and initially implemented in the on line freely available CAESAR software; these models are now available through the VEGA platform (<http://www.vega-qsar.eu/>).

As previously mentioned, thousands of models have been published in the scientific literature, but in most of the cases they are not available. However, hundreds of QSAR models have been implemented and made available, on a commercial basis or free. ANTARES [3] listed more than 250 of them and through its official website gives a direct link to them.

Typically, the performance of the QSAR models is better for physico-chemical properties, and decreases with the increased complexity of the studied system. For certain human endpoints, such as carcinogenicity and developmental toxicity, the general position is that *in silico* models should not be used as unique, sufficient tool, but as support for the evaluation based on several methods (*ref JRC*).

Indeed, models for carcinogenicity produce a quite large error. About one out of three chemicals is wrongly predicted. Better results can be obtained if the applicability domain of the model is evaluated. At least three models for carcinogenicity should be used, because the results vary. In the event of agreement, the prediction is more reliable.

For aquatic toxicity, most of the models address acute toxicity, mainly in fish. Results are good for chemicals which do not carry residues which increase their toxicity, about 30% of the cases. Specific models for more toxic compounds should not be used in these cases.

Models for mutagenicity (*mainly Ames test*) generally give good results (*accuracy about 80%, which is close to the test reproducibility*).

Models for bioconcentration factors (*BCF*) give good results (*R² about 80%; error about 0.5 log units*). Care should be taken if the predicted value is close to the threshold, while if the predicted value is well above or below the BCF threshold, the prediction is much more reliable.

The EC project ANTARES is making detailed evaluation of the results of the QSAR models for different endpoints. The reader should look for updated results at the ANTARES web site.

Below we will discuss in more detail some examples, referring to publicly available software, VEGA. The VEGA platform offers several tools. Here we will focus on some of the QSAR models

of high quality, checked with external test sets. Furthermore, VEGA has other tools to assess the model reliability. Some of these tools can be used for read across.

Models for bioconcentration factor (BCF).

According to international guidelines “Bioaccumulation” is defined as the process where the chemical concentration in an aquatic organism achieves a level that exceeds that in the water as a result of chemical uptake through all routes of chemical exposure (*e.g. dietary absorption, transport across the respiratory surface, dermal absorption*). *Bioaccumulation typically takes place under field conditions and is a combination of chemical bioconcentration and biomagnification (the process by which lipid normalized chemical concentrations increase with trophic level in a food-chain)*. The extent of chemical bioaccumulation is usually expressed in the form of a bioaccumulation factor (BAF), which is the ratio of the chemical concentration in the organism (CB) and the water (CW), including the uptake in the diet.

Bioconcentration is the process where the chemical concentration in an aquatic organism achieves a level that exceeds that in the water as a result of the exposure of an organism to a chemical in the water but does not include exposure via the diet. Bioconcentration refers to a situation, typically derived under controlled laboratory conditions, wherein the chemical is absorbed from the water via the respiratory surface and/or the skin only. The extent of chemical Bioconcentration is usually expressed in the form of a Bioconcentration factor (BCF).

The bioconcentration factor is the concentration of test substance in/on the fish or specified tissues thereof, divided by the concentration of the chemical in the surrounding medium at a steady state. In the context of setting exposure criteria it is generally understood that the terms “BCF” and “steady-state BCF” are synonymous. A steady-state condition occurs when the organism is exposed for a sufficient length of time that the ratio does not change substantially.

Bioconcentration factors (BCFs) are used to relate pollutant residues in aquatic organisms to the pollutant concentration in ambient waters. Many chemical compounds, especially those with a hydrophobic component, partition easily into the lipids and lipid membranes of organisms and bioaccumulate.

BCF and BAF are described by the following formulas:

$$BCF = CB/CWD = k1/(k2 + kE + kM + kG) \quad [1]$$

$$BAF = CB/CWD = \{k1 + kD (CB/CWD)\} / (k2 + kE + kM + kG) \quad [2]$$

Where CB is the chemical concentration in the organism (g/kg-1), k1 is the chemical uptake rate constant from the water at the respiratory surface (L*kg-1*d-1), CWD is the freely dissolved chemical concentration in the water (g*L-1), kD is the uptake rate constant for chemical in the diet (kg*kg-1*d-1) and k2, kE, kM, kG are rate constants (d-1) representing chemical elimination from the organism via the respiratory surface, faecal egestion, metabolic biotransformation, and growth dilution, respectively.

BCF & REACH

The degree of information requested under REACH varies upon yearly tonnage of production and/or import. In particular among the ecotoxicological information in Point 9.3.2 bioaccumulation is mentioned in the aquatic species, preferably fish. The preferred experimental conditions for BCF test are those reported in the OECD 305 guideline. The number of likely fish recommended for the test is in the range 132 to 240, for a duration of 44-116 days and a cost for each experiment in the range of 50-100 k€.

According to the REACH framework the potential use of BCF information includes the following:

- Classification & Labelling (C&L): all substances should be assessed for environmental hazard classification. Bioaccumulation potential is one aspect that needs to be considered in relation to long-term effects.
- Prioritization (PBT, vPvB): bioaccumulation is one of the criteria used for the PBT/vPvB assessment. For a definitive conclusion, reliable measured BCF data are generally necessary (for fish or an invertebrate such as molluscs). However, a provisional assessment can be made against screening criteria. To define if a chemical is PBT or vPvB the thresholds are: for B BCF > 2000 L/kg (whole organism weight) = 3.3 in Log unit vB BCF > 5000 L/kg = 3.7 in Log unit.

- Chemical Safety Assessment (CSA): fish BCF and BMF (*Biomagnification Factor*) values are used to calculate concentrations in fish as part of the secondary poisoning assessment for wildlife, as well as for human dietary exposure. An invertebrate BCF may also be used to model a food chain based on consumption of sediment worms or shellfish. An assessment of secondary poisoning or human exposure via the environment will not always be necessary for every substance. In the first instance, a predicted BCF may be used for first tier risk assessment.

As reported in *Table 2*, potential use of BCF information in REACH satisfies classification and labelling (C&L) requirements, B assessment and chemical safety assessment (CSA) (*up to 100 ton/year*), while for production above 100 ton/year a specific definite value becomes necessary. Thus, both quantitative and qualitative (*classification*) evaluation might be requested.

Table 2: BCF Data requests depending on the tonnage of the substances

| tons/year | C&L | B and vB | CSA | BCF Value |
|-----------|-----|----------|-----|-----------|
| >1 | X | X | | |
| >10 | X | X | X | |
| >100 | X | X | X | X |

QSAR and BCF for REACH

The PBT and vPvB assessment of a substance shall be based on all relevant information available, which is normally the information that shall be submitted as part of the technical dossier, including the physicochemical, hazard and exposure information generated in the context of the CSA.

Other properties or estimations particularly in relation to LogKow (*also called LogP*) may be used to infer bioconcentration properties of chemical compounds; thus, apart from the general stimulus to QSAR in the REACH legislation, some QSAR-based estimation methods are already mentioned for deducing BCF properties.

Moreover, to properly assess the reliability of a QSAR model prediction, a very useful criterion for comparison is the experimental data variability. The variability of the BCF data reported in the literature is ± 0.75 log units [5]. The variability of the experimental data (*calculated as the average*

of the range assumed by the values for each compound) in the Arnot et al. [database](#) is 0.69 log units [6]. Considering only experimental data for fish species suggested by the OECD (*according to OECD guideline 305*) and with an overall reliability score of 1 (*the most reliable data*), the variability drops to 0.48 log units. For the EURAS database, considered a gold standard database, the variability of the experimental BCF values is 0.45 log units, which decreases to 0.42 log units for the substances included in this study.

The CAESAR Model

The CAESAR model for BCF (*now implemented within the VEGA platform*) is based on a dataset of 473 compounds with experimentally determined BCF values extracted from Dimitrov et al. 2005 [5]. This dataset was divided into a training set (378 compounds), used to develop the model and a test (95 compounds) used to assess the performance of the model in prediction. The final hybrid model integrates two models, based on 8 molecular descriptors overall.

The prediction error was about 0.5 Log unit (*Standard Deviation of the Error of Prediction*), which is of the same range as the experimental variability. The results of the CAESAR model do not change considering the different tautomers of the chemical structures. This model was originally implemented in the CAESAR freeware and now is available within the [VEGA](#) [7] platform.

The CAESAR model for Bioconcentration Factor

The model is constituted by a combination of 2 Radial Basis Function Neural Network (*RBF-NN*) models developed with 5 descriptors each, for a total of 8 descriptors (*2 are in common between the models*). More details about this model can be found in literature (*Zhao et al., 2008*). The model reached an $R^2 = 0.83$ on the training set, and $R^2 = 0.80$ on the test set.

BCF data have been taken from Dimitrov et al. (2005) [Errore. L'origine riferimento non è stata trovata.]. The original data have been individually checked by at least two partners of CAESAR, and about 10% of the compounds were discharged, as explained in Zhao et al., obtaining 473 compounds.

CAESAR developed more than 100 BCF predictive models based on different algorithms. Some of the best models have been combined to improve performances. The model, which has been implemented, is one of these integrated models. It uses 8 descriptors within an algorithm which has fixed parameters, fully transparent and available at request (coord@caesar-project.eu). Even though sophisticated algorithms have been used in the training phase, during the development of the model and the selection of the descriptors, the final model is a simple one.

These are the 8 descriptors used by the CAESAR model:

- MlogP - Moriguchi log of the octanol-water partition coefficient (*logP*)
- BEHp2 - Highest eigenvalue n. 2 of Burden matrix/weighted by atomic polarizabilities
- AEige - Absolute eigenvalue sum from electronegativity weighted distance matrix
- GATS5v - Geary autocorrelation - lag 5/weighted by atomic van der Waals volumes
- Cl-089 - Cl attached to C1(*sp*²)
- X0sol - Solvation connectivity index chi-0
- MATS5v - Moran autocorrelation - lag 5/weighted by atomic van der Waals Volumes
- SsCl - Sum of all (*-Cl*) E-State values in molecule.

The CAESAR model, like most of the QSAR models for BCF, used mainly LogP as fundamental descriptor. So it is quite similar to models like BCFBAF v3.00 [7] and many others. We used a series of LogP values calculated with four programs at pH 7, as explained in the past [9]. Table 3 reports the correlations between these calculated LogP values and the experimental logBCF, for the chemicals used in the CAESAR model. These results do not mean that one program gives more reliable LogP prediction; we simply explain the reasons for our selection in this specific case. When the model was developed, we also used logD as additional descriptor, calculating the partition coefficient in a series of acidic and basic pH, but the results were no better.

Table 1: Regression coefficient between logP calculated with different programs and BCF.

| Descriptor | Chemical meaning | Source | Model | R | R2 | F value |
|---------------|--|-----------------|---|-------|-------|---------|
| logP (ACD) | logP value calculated by ACD software | ACD software | $\log\text{BCF} = 0.305^* \log\text{PACD} + 0.767$ | 0.605 | 0.336 | 217.442 |
| logP (Kowwin) | logP value calculated by Kowwin software | Kowwin software | $\log\text{BCF} = 0.357^* \log\text{PKowwin} + 0.605$ | 0.657 | 0.432 | 266.931 |
| logP (MDL) | logP value from MDL descriptors | MDL descriptors | $\log\text{BCF} = 0.481^* \log\text{PMDL} + 0.290$ | 0.737 | 0.543 | 448.043 |
| MLOGP | Moriguchi octanol-water partition coefficient (logP) | Dragon software | $\log\text{BCF} = 0.555^* \text{MLOGP} + 0.117$ | 0.746 | 0.556 | 471.748 |

Lombardo et al. 2010 [4]

As also appears from table 1, the correlation between LogP and BCF is not enough to support the use of this single parameter with a simple model. This is the same message as in Figure 1, where experimental values were used.

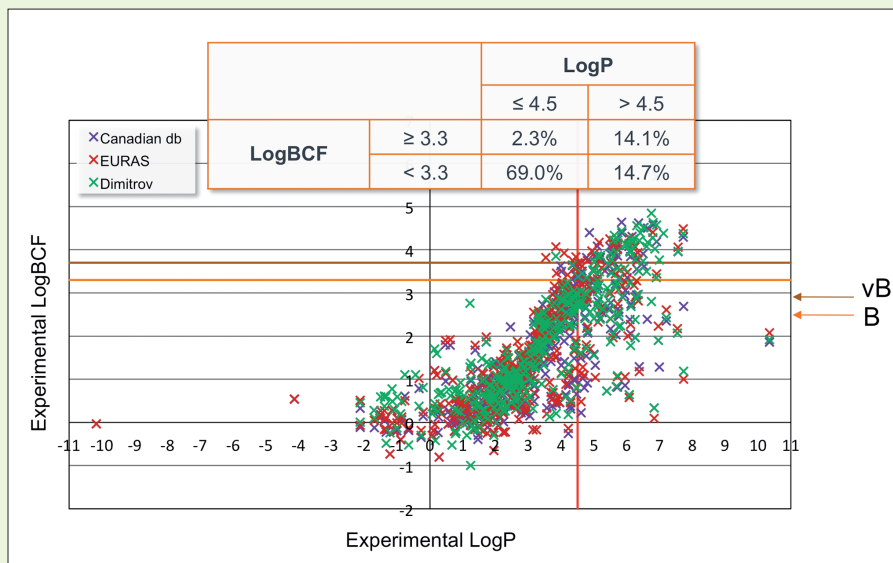


Figure 1: Comparison of experimental values of logP and LogBCF. The two thresholds for BCF indicated in the REACH legislation are shown. The screening threshold proposed by ECHA for logP 4.5 is also reported.

QSAR Model Reporting Format

Download the QSAR Model Reporting Format for the CAESAR models. The QSAR Model Reporting Format (QMRF) is a document explicitly requested for the registration of chemical substances within the REACH legislation. This document provides all the information (*including the dataset used to build and test the model*) to judge the scientific validity of the models. The QMRF of the CAESAR model for Bioconcentration factor can be downloaded from the official VEGA website.

Scientific validity

The CAESAR model has been assessed according to the OECD principles.

- A defined endpoint: The endpoint refers to the BCF value, which is an important endpoint for REACH. Only data produced according to official guidelines have been used [6]. We used the threshold defined by REACH for the characterization of the false negatives, and model selection.
- An unambiguous algorithm: Chemical structures have been checked individually by two persons, to have a correct starting point (*10% of the structures have been pruned, because not fully correct*). Bi-dimensional chemical descriptors have been used, in order to reduce the variability related to the three-dimensional descriptors. 8 descriptors are used for the model. The mathematical algorithm is unique and clearly defined. In our software, whichever format of the chemical structure is chosen by the user as input, the system transforms it into the same internal format, to guarantee the same univocal output of the model. This is not always true for other QSAR programs. We checked that tautomers do not significantly affect the model.
- A defined domain of applicability: The model is based on a data set of heterogeneous industrial chemicals of about 500 compounds. Some a priori conditions have been defined for the use of the model. The model does not work on mixtures of compounds. The model does not work on complexes. The model works on the neutral form of acids and bases. Some a posteriori restrictions have been introduced, evaluating the outliers. The model has higher uncertainty for sulfonic acids, pesticides, poly-halogenated compounds, chemicals with long aliphatic chains, or ter-butyl groups. An advanced tool for the automatic check of the reliability of the model prediction based on

the applicability domain is implemented within VEGA. The applicability check should be always checked according to REACH, and in particular if the value of the Applicability Domain Index (ADI) is < 0.7 , careful check of the QSAR results should be done by an expert.

- Appropriate measures of goodness-of-fit, robustness and predictivity: The model has been checked with a large set of statistical criteria, according to [Golbraikh at al. 2012](#) [8]. For external validation, we used an external set of 95 compounds, not used when the model was developed.
- A mechanistic interpretation, if possible: LogP is the most important descriptor, as expected, and is modulated by other descriptors. More discussion about this has been given by Lombardo at al. [4].

Classification approaches for BCF

Using a quantitative model like CAESAR as a basis for classification approaches for BCF has the main advantage that its use remains flexible, not linked to a specific threshold, such as those indicated in specific legislations (*for instance, a substance is considered bioaccumulative for REACH if the BCF value is greater than 2000, but for CLP the threshold is 500*). Therefore, it can still be used if these limits are modified or updated over the years, and the same model can be used for all thresholds. In this section, we analyse the use of the CAESAR model in classification according to the REACH. Depending on the tonnage of the chemical to be put on the market, REACH specifies different ways to report the BCF characterisation. As already explained, for lower tonnage the information is only categorical, to define the chemical as bioaccumulative or not; however, at higher tonnage (> 100 tonnes/y) BCF has to be given as a continuous value to be used for risk evaluation.

Table 5 shows the results of the model (considering only compounds in the applicability domain), used for classification in three classes with the B and vB limits indicated in REACH: 3.3 in log units for B and 3.7 for vB. To take account of the uncertainty related to experimental and predicted values, an offset of 0.5 log units was applied to the compounds whose predicted BCF values fell near the B and vB thresholds. In other words, we applied a conservative criterion, reflecting the fact that the data is affected by a given uncertainty. In Table 5, we note that when used as a classifier the CAESAR model has clear advantages over the single criterion of the LogP at 4.5 (see above) because: 1) it can predict three classes; 2) the accuracy of the prediction is much higher (always above 90% even on the second validation set, while accuracy for LogP as from Table 2 is about 84%).

Table 5: Classification with the CAESAR model. Three sets are reported: training, first validation and second validation set. The percentage of the total of compounds predicted is given without considering those outside the applicability domain. In brackets, the number of compounds for each class. The total number of compounds is also reported.

| Training set | | | | Observed logBCF | | | First validation set | | | Observed logBCF | | | Second validation set | | | Observed logBCF | | | | | | | | | | | | | | | | | | | |
|-------------------------|-----------|-------|------|-----------------|-------------------------|-----------|----------------------|------|------|-------------------------|-----------|-------|-----------------------|------|-----------|-----------------|------|-----|------|-----|-----|-------|-----|-----|--|--|--|--|--|--|--|--|--|--|--|
| 327 comp. | | | | nB | B | vB | 81 comp. | | | nB | B | vB | 119 comp. | | | nB | B | vB | | | | | | | | | | | | | | | | | |
| Predicted logBCF | nB | 82.46 | 3.38 | 0.31 | Predicted logBCF | nB | 90.00 | 3.75 | 0.00 | Predicted logBCF | nB | 88.24 | 4.20 | 0.84 | nB | (270) | (11) | (1) | (72) | (3) | (0) | (105) | (5) | (1) | | | | | | | | | | | |
| | B | 1.54 | 2.15 | 0.92 | | B | 0.00 | 1.25 | 1.25 | | B | 0.84 | 1.68 | 2.52 | | | | | | | | | | | | | | | | | | | | | |
| | vB | 0.62 | 1.23 | 7.38 | | vB | 1.25 | 0.00 | 2.50 | | vB | 0.00 | 0.84 | 0.84 | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Table 6 shows the confusion matrix using the CAESAR model as a classifier, with the 0.5 offset explained above. The percentage of false negatives decreases, but false positives increase. This solution is more conservative, as explained.

Table 6: Classification with the CAESAR model adding a 0.5 log units offset. Three sets are reported: training, first validation and second validation set. The percentage of the total of compounds predicted is given without considering those that are outside the applicability domain. In brackets, the number of compounds for each class. The total number of compounds is also reported. To take account of the endpoint variability, the predicted values are modified adding an offset of 0.5 log units for the compounds near the B and vB thresholds.

| <i>Training set</i> | | | | <i>First validation set</i> | | | | <i>Second validation set</i> | | | | | | |
|-------------------------|-----------|----------|-----------|-----------------------------|-------------------------|-----------|-----------|------------------------------|-----------|-------------------------|-----------|-------|------|------|
| <i>Observed logBCF</i> | | | | <i>Observed logBCF</i> | | | | <i>Observed logBCF</i> | | | | | | |
| 327 comp. | nB | B | vB | 81 comp. | nB | B | vB | 119 comp. | nB | B | vB | | | |
| Predicted logBCF | nB | 73.70 | 0.31 | 0.00 | Predicted logBCF | nB | 77.78 | 0.00 | 0.00 | Predicted logBCF | nB | 81.51 | 1.68 | 0.00 |
| | | (241) | (1) | (0) | | | (63) | (0) | (0) | | | (97) | (2) | (0) |
| | B | 8.87 | 3.06 | 0.31 | | B | 11.11 | 3.70 | 0.00 | | B | 7.56 | 1.68 | 0.84 |
| | | (29) | (10) | (1) | | | (9) | (3) | (0) | | | (9) | (2) | (1) |
| | vB | 2.14 | 3.36 | 8.26 | | vB | 1.23 | 2.47 | 3.70 | | vB | 0.84 | 2.52 | 3.36 |
| | | (7) | (11) | (27) | | | (1) | (2) | (3) | | | (1) | (3) | (4) |

The performance in classification of the CAESAR model (without and with the 0.5 correction) was compared with that of BCFBAF v3.00 (see Tables 7 and 8). Figure 4 shows the results of the CAESAR model used as classifier.

Table 7: Classification with the BCFBAF v3.00 model. Three sets are reported for the compounds of the dataset: training, validation and external. The percentage of the total of compounds predicted is given without considering those outside the applicability domain. In brackets, the number of compounds for each class is reported. The total number of compounds is also reported.

| <i>Training set</i> | | | | <i>Validation set</i> | | | | <i>External set</i> | | | | | | |
|-------------------------|-----------|----------|-----------|------------------------|-------------------------|-----------|-----------|------------------------|-----------|-------------------------|-----------|-------|-------|-------|
| <i>Observed logBCF</i> | | | | <i>Observed logBCF</i> | | | | <i>Observed logBCF</i> | | | | | | |
| 450 comp. | nB | B | vB | 103 comp. | nB | B | vB | 82 comp. | nB | B | vB | | | |
| Predicted logBCF | nB | 82.00 | 3.33 | 2.00 | Predicted logBCF | nB | 81.55 | 2.91 | 1.94 | Predicted logBCF | nB | 39.02 | 12.20 | 3.66 |
| | | (369) | (15) | (9) | | | (84) | (3) | (2) | | | (32) | (10) | (3) |
| | B | 2.00 | 0.67 | 2.00 | | B | 0.97 | 0.97 | 0.00 | | B | 10.98 | 4.88 | 4.88 |
| | | (9) | (3) | (9) | | | (1) | (1) | (0) | | | (9) | (4) | (4) |
| | vB | 2.22 | 0.89 | 4.89 | | vB | 3.88 | 0.97 | 6.80 | | vB | 6.10 | 3.66 | 14.63 |
| | | (10) | (4) | (22) | | | (4) | (1) | (7) | | | (5) | (3) | (12) |

Table 8: Classification with the BCFBAF v3.00 model adding a 0.5 log units offset. Three sets are reported for the compounds of the dataset: training, validation and external. The percentage of the total of compounds predicted is given without considering those that are outside the applicability domain. In brackets is the number of compounds for each class. The total number of compounds is also reported. To take account of the endpoint variability, the predicted values are modified adding an offset of 0.5 log units for the compounds near the B and vB thresholds.

| Training set | | Observed logBCF | | | Validation set | | Observed logBCF | | | External set | | Observed logBCF | | |
|------------------|----|-----------------|------|------|------------------|-----|-----------------|-------|------|------------------|-----|-----------------|------|-------|
| 450 comp. | | nB | B | vB | 103 comp. | | nB | B | vB | 82 comp. | | nB | B | vB |
| Predicted logBCF | nB | 77.56 | 2.22 | 0.44 | Predicted logBCF | nB | 78.64 | 1.94 | 0.97 | Predicted logBCF | nB | 26.83 | 6.10 | 1.22 |
| | | (349) | (10) | (2) | | | (81) | (2) | (1) | | | (22) | (5) | (1) |
| | B | 4.44 | 1.11 | 1.56 | | B | 2.91 | 0.097 | 0.97 | | B | 12.20 | 6.10 | 2.44 |
| | | (20) | (5) | (7) | | (3) | (1) | (1) | | (10) | (5) | (2) | | |
| | vB | 4.22 | 1.56 | 6.89 | | vB | 4.85 | 1.94 | 6.80 | | vB | 17.07 | 8.54 | 19.51 |
| | | (19) | (7) | (31) | | (5) | (2) | (7) | | (14) | (7) | (16) | | |

| 542 compounds | | | | |
|---------------------|----|------------------|-------|-------|
| CAESAR BCF model | | Predicted LogBCF | | |
| | | nB | B | vB |
| Experimental LogBCF | nB | 83.39% | 1.29% | 0.37% |
| | B | 4.80% | 1.66% | 0.55% |
| | vB | 0.74% | 1.48% | 5.72% |

Figure 4: Results of the CAESAR model as classifier. Comparison of accuracy, using CAESAR and BCFBAF v3.00, for their three respective sets (training, validation and external). * Modified: using an offset of 0.5 for values close to the thresholds (see text).

The read-across model

Besides the QSAR model described above, the VEGA platform makes available a model developed with the k-nearest neighbour (KNN) technique. In this case the BCF value is derived in a way similar to the read-across strategy: the software identifies the chemicals which are more similar to the target compounds, and then calculates the BCF value of the target chemical using the BCF values.

The integration and interpretation of the results

VEGA offers several models for BCF. Three of these models are derived from the QSAR CAESAR platform. Indeed, CAESAR starts from 2 QSAR models, based on LogP and other descriptors as described above. The results of these two models are then used as input for the third, hybrid model, which provides a third value, which is not the average of the worst case [9]. A fourth model is based on the KNN algorithm, as explained above.

The user should check the consistency of the results of the different models and must evaluate whether or not the result of the QSAR or read-across model is reliable¹.

VEGA is a tool supporting the human expert in the assessment of the chemical properties. It combines QSAR and read-across tools. A completely independent algorithm is at the basis of the evaluation for read across. This algorithm shows similar compounds, assesses the QSAR results on the similar compounds, and analyses certain relevant chemical features in the target compound and its related compounds. Some automatic evaluation for read across is done, but the user should also analyse the results independently.

For the evaluation of the results, we recommend taking advantage of both of these tools, QSAR and read across. The user should analyse the results shown on the similar compounds, and the additional documentation provided on the occurrence of relevant chemical moieties.

By using more models and inputs, uncertainty can be reduced. For this reason we recommend carefully considering all data provided by VEGA, not only the result of the prediction. Even if the ADI value is not high, the user can get sufficient information on the BCF class considering the similar compounds, their experimental values, and the figures showing the LogP versus LogBCF values. The agreement between the CAESAR and KNN values reinforces the reliability.

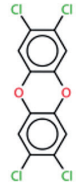
¹ The VEGA guidance provides all the necessary information to evaluate the reliability and is accessible through the VEGA official website in the section “how to interpret results” - <http://www.vega-qsar.eu/interpretation.html>

In cases of uncertainty, we recommend using an additional model, such as EPISuite or T.E.S.T., considering that agreement between predictions reinforces the reliability.

Case studies

In order to help users better understand how to interpret the results of the CAESAR/VEGA models, six case studies have been implemented in the VEGA websites using the CAESAR model for bioconcentration factor integrated in the VEGA platform. Two reliable and four unreliable cases have been considered and explained in detail as shown in Figure 5. The 6 case studies are available in the “*how to interpret results*” section of the VEGA website (<http://www.vega-qsar.eu/interpretation.html>).

EXAMPLE 1



NAME
2,3,7,8-Tetrachlorodibenzo-p-dioxin

CAS
1746-01-6

SMILES
O1c3ccc(cc3(Oc2cc(ccc2)Cl)Cl)Cl







PREDICTED VALUE
3.02

Model assessment:

Unable to provide a prediction due to the presence of one or more fragments related to model outliers. The following relevant fragments have been found: O linked to aromatic and 3 Br/Cl linked to aromatic (SO 05).

Let's consider the read across evaluation:

- There are chemicals with very good similarity. Look at the similarity values > 0.95, and at the chemical structures.

| | |
|--|--|
|  CAS: 29446-15-9 Dataset id: 47 (training set) SMILES: <chem>O1c3cccc3(Oc2cc(ccc2)Cl)Cl</chem> Similarity: 0.993 Experimental value: 2.88 [log(L/kg)] Predicted value: 1.93 [log(L/kg)] |  CAS: 30746-58-8 Dataset id: 53 (training set) SMILES: <chem>O1c3cccc3(Oc2c1c(c(c2Cl)Cl)Cl)Cl</chem> Similarity: 0.946 Experimental value: 3.5 [log(L/kg)] Predicted value: 3.17 [log(L/kg)] |
|  CAS: 33857-26-0 Dataset id: 49 (test set) SMILES: <chem>O1c3ccc(cc3(Oc2ccc(cc12)Cl)Cl)Cl</chem> Similarity: 0.986 Experimental value: 2.22 [log(L/kg)] Predicted value: 1.93 [log(L/kg)] |  CAS: 39227-58-2 Dataset id: 52 (training set) SMILES: <chem>O1c3cccc3(Oc2c1c(cc(c2Cl)Cl)Cl)Cl</chem> Similarity: 0.945 Experimental value: 2.97 [log(L/kg)] Predicted value: 2.7 [log(L/kg)] |
|  CAS: 67028-18-6 Dataset id: 54 (training set) SMILES: <chem>O1c3ccc(cc3(Oc2c1cc(c(c2Cl)Cl)Cl)Cl)Cl</chem> Similarity: 0.949 Experimental value: 3.35 [log(L/kg)] Predicted value: 3.09 [log(L/kg)] |  CAS: 56348-72-2 Dataset id: 230 (test set) SMILES: <chem>O1c1ccc(c(c1)Cl)Clc2ccc(c(c2)Cl)Cl</chem> Similarity: 0.848 Experimental value: 4.17 [log(L/kg)] Predicted value: 3.86 [log(L/kg)] |

- All these chemicals differ only for the number or the position of the chlorine atoms.
- Figure 1 shows that the target chemical is within the typical relationship between logP and BCF.
- Figure 2 shows that the BCF and logP values of the target compound are close to the values of the similar compounds. A trend appears.
- The QSAR model underestimates BCF, of about 0.3-0.9 log units.
- The similar compounds with higher BCF values have four chlorine atoms.

Conclusion: The compound is out of the applicability domain only for the presence of O linked to aromatic and 3 Br/Cl linked to aromatic, which is associated to higher uncertainty of the predictions; **nevertheless, the read across assessment supports with a reasonable documentation that the QSAR prediction is realistic, even though probably slightly underestimated, considering the results on the other tetrachlorodibenzo-p-dioxins.**

Figure 5: How to interpret the output of the models included in VEGA. The figure shows the first (out of six) examples provided in the VEGA websites as case studies to learn how to interpret results. For each of the six molecules utilized, the VEGA evaluation has been reported as well as several aspects considered for the evaluation of the reliability of the results.

The models for mutagenicity

The endpoint

Chemical substances, mixtures of substances and physical agents (e.g., radiation) can induce alterations in the genome of either somatic or germinal cells. There are several mutagenic endpoints of concern; these include point mutations (i.e., sub microscopic changes in the base sequence of DNA) and structural or numerical chromosome aberrations. Structural aberrations include deficiencies, duplications, insertions, inversions, and translocations, whereas numerical aberrations are gains or losses of whole chromosomes (e.g., trisomy, monosomy) or sets of chromosomes (haploidy, polyploidy).

Certain mutagens, such as alkylating agents, can directly induce alterations in the DNA. Mutagenic effects may also come about through mechanisms other than chemical alterations of DNA. Among these are interference with normal DNA synthesis (as caused by some metal mutagens), interference with DNA repair, abnormal DNA methylation, abnormal nuclear division processes, or lesions in non-DNA targets. Evidence that an agent induces heritable mutations in human beings could be derived from epidemiologic data indicating a strong association between chemical exposure and heritable effects. It is difficult to obtain such data because any specific mutation is a rare event, and only a small fraction of the estimated thousands of human genes and conditions are currently useful as markers in estimating mutation rates.

Mutagenicity & REACH

Mutagenicity studies are required for all the tonnage bands. Unlike most other endpoints, a negative mutagenicity result in vitro can be considered sufficient evidence for non-mutagenic potential but positive results must be confirmed in vivo. At 1-10 tonnage level (Annex VII) the in vitro gene mutation study in bacteria (Ames test) is required. At Annex VIII level (10-100 ton), two additional in vitro studies are required: a cytogenicity study and a gene mutation study in mammalian cells. If

there is any positive result within these in vitro tests, in vivo mutagenicity studies are required. At higher tonnage in vivo studies are needed.

The model

This model has been implemented in the CAESAR freeware and now is available at the VEGA website. The CAESAR model for mutagenicity is based on a data set that includes 4225 compounds. For developing classification models, this data set was divided in two classes: 80% (3380 chemicals) used for building the model and 20% (845 chemicals) left for testing.

For regulatory purposes, an integrated model was arranged combining two complementary techniques: a machine learning algorithm (Support Vector Machines - SVM), to build an early model with the best statistical accuracy, equipped with an expert facility for false negatives (FN) removal based on known structural alerts, to refine its predictions. Figure 6 shows the architecture of the model.

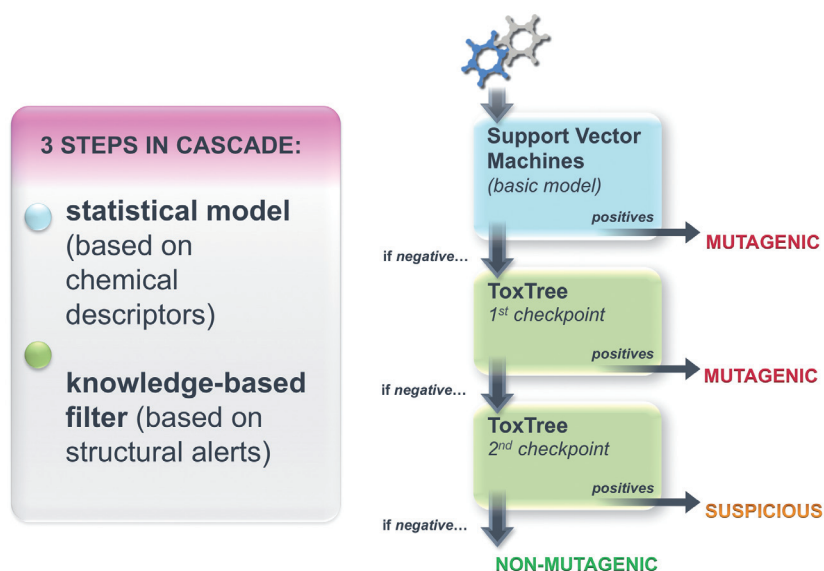


Figure 6: The combined approach for mutagenicity

After the SVM step, a set of rules (Figure 7), taken from the Benigni/Bossa [11] rules, have been utilized.

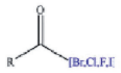
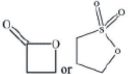


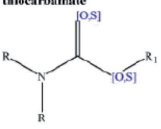
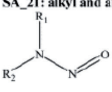
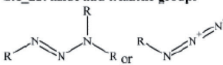

| | |
|--|--|
| <p>SA_1: acyl halides</p>  | <p>R= any atom/group, except OH, SH</p> |
| <p>SA_6 propiolactones or propiolactones</p>  | |
| <p>SA_9: alkyl nitrite</p>  | <p>R= any alkyl group</p> |
| <p>SA_13: hydrazine</p>  | <p>R=any atom/group</p> |
| <p>SA_16: alkyl carbamate and thiocarbamate</p>  | <p>R= Aliphatic carbon or hydrogen R1 = Aliphatic carbon</p> |
| <p>SA_18: Polycyclic Aromatic Hydrocarbons</p> | <p>Three or more fused rings, not heteroaromatic</p> |
| <p>SA_19: Heterocyclic Polycyclic Aromatic</p> | <p>Three or more fused rings, heteroaromatic</p> |
| <p>SA_21: alkyl and aryl N-nitroso groups</p>  | <p>R1= Aliphatic or aromatic carbon, R2= Any atom/group</p> |
| <p>SA_22: azide and triazene groups</p>  | <p>R=any atom/group</p> |
| <p>SA_27: nitro-aromatic</p>  | <p>Ar = Any aromatic/heteroaromatic ring Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded. Chemicals with a sulfonic acid group (-SO₃H) on the same ring of the nitro group are excluded.</p> |

Figure 7: Structural Alerts utilized in the second step of the CAESAR hybrid model for mutagenicity. The depicted fragments, which are part of the Benigni/Bossa ruleset for mutagenicity, have been implemented within the first of the two knowledge-based model integrated in the CAESAR model for mutagenicity.



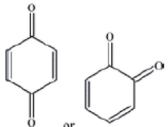


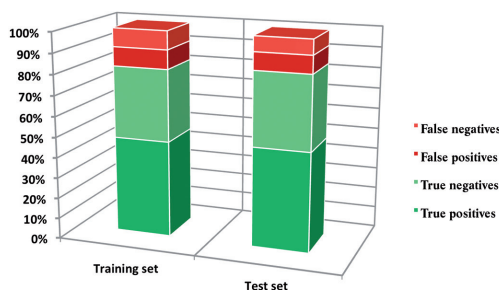
| | |
|--|---|
| <p>SA_7: epoxides and aziridines</p>  | <p>R= any atom/group</p> |
| <p>SA_8: aliphatic halogens</p>  | <p>R= any atom/group</p> |
| <p>SA_12: quinones</p>  | |
| <p>SA_28 bis: aromatic mono- and dialkylamine</p>  | <p>Ar = Any aromatic/heteroaromatic ring R1= Hydrogen, methyl, ethyl R2 = Methyl, ethyl Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded. Chemicals with a sulfonic acid group (-SO₃H) on the same ring of the nitro group are excluded.</p> |
| <p>SA_29: aromatic diazo</p>  | <p>Ar = Any aromatic/heteroaromatic ring Chemicals with a sulfonic acid group (-SO₃H) on both rings linked to the diazo group are excluded</p> |

Figure 8: Structural Alerts utilized in the third step of the CAESAR hybrid model for mutagenicity. As for the first set, these structural alerts have been obtained from the Benigni/Bossa ruleset for mutagenicity and carcinogenicity.

A further set of rules (*Figure 8*) is used after this step, but these rules are not accurate: a number of false positives may be expected, and for this reason the prediction of the model is given as “suspicious”.

The expert filter, to be applied only on compounds presumed safe by SVM, wraps two sets of structure alerts (*SA*) (*selected from the Benigni/Bossa rulebase*) with different distinguishing features: the former has the aim of enhancing the prediction accuracy attempting a precise identification of misclassified FN. The latter (*the ‘suspicious’ one*) goes on with the FN removal to an extent that does not noticeably downgrade the original prediction accuracy (*by generating too many false positives -FP- as well*). To point out this distinction, compounds picked out by the first checkpoint are classified as ‘mutagenic’, and those picked out by the second are classified as ‘suspicious’. Unaffected ones are finally classified as ‘non-mutagenic’.



- Good accuracy (considering reproducibility of the experimental data about 85%)
- A cost-sensitive model was also evaluated to reduce FN

Figure 9: The results of the CAESAR classification model for mutagenicity with SVM. The percentage of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) are here reported for both the Training Set and the Test Set. Considering the percentage of TP and TN, the molecules correctly predicted are little less of the 80%, which is an acceptable result, considering that the reproducibility of experimental data is around 85%.

Figure 9 shows results obtained with SVM, while Figure 10 compares the results obtained with CAESAR SVM (*only the first step of the CAESAR hybrid model*) with Toxtree.

The third step of the CAESAR hybrid model classifies the molecules as non-mutagens or suspicious mutagens. Depending on the type of approach the user wants to adopt, they can be considered either mutagens (*conservative approach*) or non-mutagens (*non-conservative approach*). Figure 11 shows the performance of the CAESAR model in both cases.

| CAESAR Test Set | Toxtree | SVM model |
|-----------------|---------|-----------|
| ACCURACY: | 78% | ✓ 83% |
| SENSITIVITY: | 86% | ✓ 87% |
| SPECIFICITY: | 69% | ✓ 79% |

Figure 10: Comparison of the results between CAESAR SVM and Toxtree. The performances of the SVM model included in CAESAR mutagenicity model (as first step) and Toxtree (based on Benigni/Bossa rulebase) have been compared. As shown, the SVM gave better results for the three parameters commonly used to evaluate binary classifiers' performances (Accuracy, Sensitivity and Specificity).

| CAESAR Test Set | SUSPICIOUS taken as NON-MUTAGENIC | SUSPICIOUS taken as MUTAGENIC |
|-----------------|-----------------------------------|-------------------------------|
| ACCURACY: | 83.3% | 82.1% |
| SENSITIVITY: | 88.3% | 90.9% |
| SPECIFICITY: | 77.1% | 71.2% |

CONFIDENT CHOICE
Accuracy close to the reliability of the experimental test (85%)
[non-conservative approach]

PRUDENT CHOICE
Sensitivity boosted over 90%
[conservative approach]

Figure 11: Differences in the CAESAR model performances while applying conservative and non-conservative approaches. Applying the conservative approach (suspicious considered as mutagens) lead to a higher accuracy and, as expected, to an increased specificity. Adopting a non-conservative approach (suspicious considered as non-mutagens) leads as expected to a higher sensitivity while decreasing the accuracy.

The models for carcinogenicity

The endpoint

The process of carcinogenesis involves the transition of normal cells into cancer cells via a sequence of stages that entail both genetic alterations (*i.e. mutations*) and non-genetic events. Genotoxic modes of action involve genetic alterations caused by the chemical interacting directly with DNA to result in a change in the primary sequence of DNA.

Non-genotoxic modes of action include epigenetic changes, *i.e.*, effects that do not involve alterations in DNA but that may influence gene expression, altered cell-cell communication, or other factors involved in the carcinogenic process.

Chemicals are defined as carcinogenic if they induce tumours, increase tumour incidence and/or malignancy or shorten the time to tumour occurrence. Chemicals can induce cancer by any route of exposure (*e.g., when inhaled, ingested, applied to the skin or injected*), but carcinogenic potential and potency may depend on the conditions of exposure (*e.g., route, level, pattern and duration of exposure*). Carcinogenic chemicals have conventionally been divided into two categories according to the presumed mode of action: genotoxic or non-genotoxic.

Different experimental data to do with carcinogenicity studies are available, starting from *in vitro* methods to long-term bioassays in experimental animals, to human epidemiologic studies. All this information is important for the overall assessment.

Since the etiopathogenesis of many tumours is DNA mutations, often mutagen substances are also carcinogens. Even if strongly correlated, these phenomena have to be distinguished according to REACH legislation.

Carcinogenicity & REACH

Between different endpoints, carcinogenicity is one of the most essential ones in assessment of human health safety. The objective of investigating the carcinogenicity of chemicals is to identify potential human carcinogens, their mode(s) of action, and their potency. The golden standard test under REACH for carcinogenicity is OECD 451. OECD guideline no. 451 recommends the use of 400 animals (*rats and mice*) for the studies. Standard information requirements are specifically described for substances produced or imported in quantities of ≥ 1000 tons/year.

The precise information requirements will differ from substance to substance, according to the toxicity information already available and details of use and human exposure for the substance in question. A carcinogenicity study may be proposed by the registrant or may be required by the Agency if:

- the substance has a widespread dispersive use or there is evidence of frequent or long-term human exposure; and
- the substance is classified as mutagen category 3 or there is evidence from the repeated dose studies that the substance is able to induce hyperplasia and/or pre-neoplastic lesions.

If the substance is classified as mutagen category 1 or 2, the default assumption would be that a genotoxic mechanism for carcinogenicity is likely. In these cases, a carcinogenicity test will normally not be required.

These carcinogenicity assays are expensive, require the sacrifice of many animals and also require many years per experiment, raising economic and ethical issues. The development of *in silico* methods to reduce these costs is thus a priority.

QSAR and Carcinogenicity for REACH

The big challenge in solving the general carcinogenicity prediction problem is to construct a model that would be able to predict carcinogenicity for a wide diversity of molecular structures, spanning an undetermined number of chemical classes and biological mechanisms. Quantitative models based on SMILES [12] for prediction of carcinogenicity were successfully developed [13,14].

Many statistical approaches can be used for prediction of complex endpoint such as carcinogenicity. The CAESAR models are in the area of the data mining models which address complex endpoints. Others models, which have been developed, are based on toxic residues codifying human expert knowledge, such as [OncoLogic](#) [15], [HazardExpert](#) [16], [Derek](#) [17], [ToxTree](#) [18], or data mining based on fragments, such as [MultiCase](#) [19,20].

The predictive power of models is one of the most important characteristics in QSAR modelling. In a recent paper [Benigni et al.](#) [21] pointed out that the prediction reliability should be checked by means of an external test set with new chemicals not used in modelling. The state of art and perspectives of predictive models for carcinogenicity are reported in [a recent paper](#) [22]. It was stressed that the models for regulatory purposes should be connected with high sensitivity, i.e., the ability to correctly identify true positives and minimize false negatives.

The CAESAR model

Within CAESAR, the data mining approach has been improved by using a highly verified set of compounds (all chemical structures have been double-checked, and experimental data verified in cases of unusual findings, compared to similar compounds), and adopting a wide series of chemical descriptors. Different algorithms have been developed resulting in a series of models, one of which is implemented and reported here due to its higher performance.

Preliminary results of carcinogenicity modelling using CP ANN algorithm obtained in the scope of CAESAR project are described in an [article](#) [23], where a QSAR model has been built starting from a dataset of 805 non-congeneric organic compounds, characterized by 27 MDL descriptors.

Among statistical approaches, artificial neural networks (ANNs) appeared to be one of the most suitable and promising for prediction of complex endpoints such as carcinogenicity for non-congeneric datasets of chemicals. The main advantage of neural network modelling is that complex, non-linear relationships can be modelled without any assumptions about the form of the model. Large datasets can be examined. [Neural networks are able to cope with noisy data and are fault-tolerant](#) [24].

Here we presented categorical or qualitative models for prediction of carcinogenic potency of non-congeneric chemicals using CP ANN method. Our models have been developed in accordance

with principles of validation adopted by OECD within the European Commission (EC) funded project CAESAR (*Computer Assisted Evaluation of industrial chemical Substances According to Regulation*) [2].

In our study an external dataset of 738 chemicals was composed and external validation of models made. In the paper it is shown how the number of correctly predicted carcinogens can be increased using correlation between the threshold of categorical models and sensitivity and specificity. We address the issue of threshold effects on the overall performance of models.

The results of the CAESAR classification model for Carcinogenicity

The CP ANN model presented in our study demonstrated good prediction statistics on the test set of 161 compounds with sensitivity of 75%, specificity of 61%-69% in addition to accuracy of 69%-73%. A diverse external validation set of 738 compounds confirmed the robustness of our models regarding a large applicability domain, yielding accuracy of 60.0%-61.4%, sensitivity of 61.8%-64.0%, and specificity of 58.4%-58.9%.

According to these results, currently the role of in silico models in this endpoint can be limited to considering it as a supplement to existing methods in gathering evidence rather than a substitute. Its utility is consequently in support of the overall assessment.

The user is also advised that, since some of the models are based on datasets focused on a limited chemical space, particular attention should be placed for this endpoint in the evaluation of similar compounds already present in the studied datasets and the model's ability to correctly predict them. As a conclusion, carcinogenicity models can be used as a support in risk assessment, e.g. in setting priorities among chemicals for further testing.

The models for Developmental toxicity

The endpoint

Developmental toxicity refers to any effect interfering with normal development, both before and after birth. This includes embryotoxic/fetotoxic effects such as reduced body weight, growth and developmental retardation, organ toxicity, death, abortion, structural defects (*teratogenic effects*), functional effects, peri- and postnatal defects, and impaired postnatal mental or physical development up to normal pubertal development. This important endpoint is more problematic to assess than other endpoints [25]. Developmental toxicity involves several issues and a variety of experimental methods may be adopted. The complexity, length, and cost of the experiments and the late recognition of the importance of this endpoint have resulted in a low number of available studies.

Developmental toxicity & REACH

The standard data requirements for developmental toxicity under the REACH Regulations are as follows:

- A reproduction/developmental toxicity screening test (OECD TGs 421 or 422), usually required for substances produced or imported in quantities between 10 and 100 tons/year (*according to Annex VIII of the REACH legislation*).
- A prenatal developmental toxicity study (EU B.31, OECD TG 414) in one species, usually required for substances produced or imported in quantities of ≥ 100 tons/year (*according to Annex IX and X of the REACH legislation*). A study in a second species may be considered necessary at these levels of production/importation.

However, these tonnage-related standard data requirements may be adapted, either by being reduced (*e.g. a data waiver*), deferred or extended. Factors that can influence the testing requirements include structural relationships with other chemicals, the results of other toxicity studies, presence

of mutagenic and carcinogenic properties, available data from humans exposed to the substance, concerns for endocrine disruption and the patterns of use and human exposure. The golden standard test under REACH for developmental toxicity is OECD 414, which recommends the use of 80 adult animals (*rats or rabbits*) for this study. Studies according to OECD 421 or 422 are used for screening only. If positive, they may not be considered sufficient for classification & labelling and quantitative risk assessment. If negative for a substance produced or imported in quantities between 10 and 100 tons/year, no further testing is needed. A negative outcome of these screening studies for substances produced or imported in quantities over 100 tons/year implies an OECD 414 study needs to be performed.

Developmental toxicity is correlated with reproductive toxicity, and the two endpoints share the highest costs. The studies for these two endpoints are estimated to require more than one half of the total costs associated with REACH. In fact, the European Chemicals Bureau made estimations of the resources needed for these studies and quantified them as 38% of the animals and 30% of the total cost for the two-generation reproductive toxicity studies and 23% of the animals and 24% of the resources for the developmental toxicity studies.

A generation reproductive toxicity test may cost up to 750,000\$ and require 3,200 rats, a developmental toxicity test costs about one-tenth that amount and requires 150 animals, for a total cost of 22 million of experimental animals for only these two REACH endpoints [26].

QSAR and Developmental toxicity for REACH

Developmental toxicity is a complex endpoint; it depends by a multitude of factors acting together. Developmental effects should be considered in relation to adverse effects occurring in the parents. Since adverse effects during or after pregnancy may result in a secondary consequence of maternal toxicity, reduced food or water intake, maternal stress, lack of maternal care, specific dietary deficiencies, poor animal husbandry, intercurrent infections etc., it is important that the effects observed should be interpreted in conjunction with possible concomitant maternal toxicity. The nature, severity and dose-response of all effects observed in progeny and parental animals should be considered and compared together to achieve a balanced integrated assessment of available data on all endpoints relevant for developmental toxicity.

In an intricate scenario such as this, it is a great challenge to develop useful *in silico* methods able to predict the proprieties of chemicals linked to this endpoint.

Some results have been previously published on the same dataset used for Caesar models [27,28]. These studies applied two different methods: CART decision tree and logistic regression. The prediction statistics for CART were as follows: accuracy 57-63%; sensitivity 58-64%; specificity 57-66%. The statistics for logistic regression were similar: accuracy 60-62%; sensitivity 60-63%; specificity 59-62%.

Other studies on developmental toxicity have modelled individual animal endpoints [10,12]. Accuracy in these cases ranged from 45 to 88% for reproductive toxicity, and sensitivity from 10 to 72% (*however, for the last value accuracy was exceptionally low, only 45%*).

The results from the CAESAR models (*see below*) are good compared to those in the literature.

The CAESAR models

The dataset extracted from Arena et al. [27] includes 292 compounds divided into training and test set. Chemical compounds were categorized into toxicant or non-toxicant according to FDA risk factors.

Table 9: Classification between developmental and non-developmental toxicant within CAESAR project

| FDA classes | Definition | CAESAR Binary class |
|-------------|---|----------------------------|
| Category A | Negative human studies | Non developmental toxicant |
| Category B | Negative animal studies & No human studies executed OR Positive animal studies & Negative human studies | |
| Category C | Positive animal studies & No human studies executed OR No studies at all | Developmental toxicant |
| Category D | Positive human studies | |
| Category X | Animal OR human studies show abnormalities AND/OR Evidence of foetal risk based on human experience | |

The chemical descriptors were calculated in collaboration with Dr Todd Martin of U.S. EPA in a public program of his agency.

Several models have been developed: in one case (*model A*) the software used for the model is WEKA (*Waikato Environment for Knowledge Analysis*), an open source workbench. In this case we used 13 chemical descriptors. The algorithm used for modelling is Random Forest that constructs a “forest” of random “trees” (*this model has been implemented in the CAESAR software*).

A second model (*model B*) was developed using Adaptive Fuzzy Partition (*AFP*) – AFP was used to develop classification models implementing a fuzzy partition algorithm. It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces. The aim of this algorithm is to select the descriptor and the cut position that allow the discovery of the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the chemical activity values in an active subspace A and in its neighbouring subspaces. In this case we used 6 chemical descriptors. For the feature selection, a hybrid selection algorithm (*HSA*), which combines the genetic algorithm (*GA*) concepts and a stepwise regression, was used to select the best descriptors for classifying developmental toxicity dataset.

Figure 12 summarizes the results obtained on training and test sets using both models.

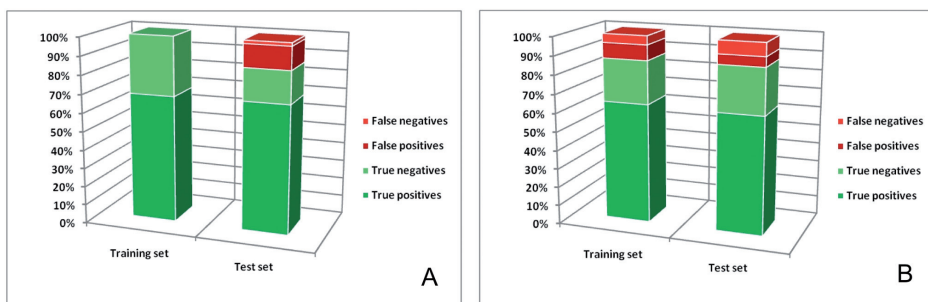


Figure 12: Results of two classification models for developmental toxicity. Percentage of TP, TN, FP and FN obtained on both training and test set using two predictive models. A) Random Forest Model built using WEKA. B) Model obtained using the Adaptive Fuzzy Partition (AFP) method.

The RF model shows good accuracy in the test set (84%), and in fitting (100%), with few false negatives (95% sensitivity). The AFP model shows high accuracy in the test set (88%) and in fitting (87%) with a good balance between sensitivity (90%) and specificity (82%). The results shown in [Table 2](#) with the AFP model were obtained after removing one compound (*etoposide*), because the AFP model could not predict it. Indeed, the descriptor values for etoposide are outside the domain applicability of the AFP model for one descriptor. The quality of the models also appears good with internal validation: the accuracy in cross-validation (*leave-several-out*) was 77% for the RF model, and 72% for the AFP. We used ten-fold cross-validation for RF and leave-one-out for AFP.

Thus, the CAESAR models offer an improvement over previous models based on the same set of compounds, and perform well compared with other studies on developmental toxicity. Our set of compounds includes a heterogeneous list of chemicals, from different classes. The same applies to the other models on developmental toxicity mentioned before.

The models for skin sensitization

The endpoint

A skin sensitizer is a substance that will induce an allergic response following skin contact. Substances are classed as skin sensitizers if there is evidence in humans that the substance can induce sensitization by skin contact in a substantial number of persons, or where there are positive results from an appropriate animal test.

A range of in vivo methods exist that have been proven to be very accurate in terms of the predictive identification of chemicals that possess skin sensitizing properties. For many years, guinea pigs were the species of choice for sensitization tests. Two types of tests were developed: adjuvant tests in which the skin sensitization effect is potentiated by the injection of Freund's Complete Adjuvant (*GPMT*), and non-adjuvant tests (*The Buehler test*).

The Local Lymph Node Assay (*LLNA*) is based upon the characteristics of induced proliferative responses in draining lymph nodes following topical exposure of mice to chemicals. The endpoint is the stimulation index (*SI*), which gives a ratio of thymidine incorporation in lymph nodes from dosed animals compared to the incorporation in lymph nodes from vehicle treated control animals. The test is positive when the stimulation index (*SI*) is greater than 3 for any of the dose concentrations. The EC₃ value, interpolated from the dose response curve, is the effective concentration of the test substance required to produce a three-fold increase in the stimulation index compared with vehicle-treated controls.

Furthermore progress in understanding the mechanisms of skin sensitization, including effects on the production of cytokines by the different cell types within the skin, provides the opportunity to develop in vitro tests as an alternative to in vivo sensitization testing. With the forthcoming elimination of in vivo tests for cosmetic testing in the European Union, several opportunities that have been exploited for in vitro test development focus on key elements of the sensitization process. One unifying characteristic of chemical allergens is the requirement that they react with proteins for the effective induction of skin sensitization. The majority of chemical allergens are electrophilic and react with nucleophilic amino acids.

Skin sensitization & REACH

Skin sensitization is an endpoint that needs to be assessed within REACH. The test of first choice under REACH for skin sensitization is the Local Lymph Node Assay (*LLNA*).

According to OECD data, testing for skin sensitization accounts for over 5% of the total use of animal tests. OECD guideline no. 406 recommends the use of 30 animals (*guinea pigs*) while OECD guideline no. 429 recommend the use of 25 animals (*mice*) for the *LLNA* assay.

In REACH, sensitizing potential needs to be assessed for chemicals above the 1 ton threshold according to Annex VII

QSAR and Skin sensitization for REACH

One potential alternative approach to skin sensitization hazard identification is the use of (*quantitative*) structure activity relationships (QSARs) coupled with appropriate documentation and performance characteristics. This represents a major challenge. Current thinking is that QSARs might best be employed as part of a battery of approaches that collectively provide information on skin sensitization hazard.

The CAESAR Models

Global QSAR model

This CAESAR model for skin sensitization is based on a data set that includes 209 compounds extracted from the paper by [Gerberick et al. \[29\]](#). To develop classification models, this data set was subdivided into two classes, sensitizer (*S*) and non-sensitizers (*N*), which gave a good distribution of the numbers of compounds in each class. The class *S* merges the first four ranges established by ECETOC²: Extreme ($EC3 < 0.1\%$), Strong ($0.1\% < EC3 < 1\%$), and Moderate ($1\% < EC3 < 10\%$) and Weak ($EC3 > 10\%$) ranges; the class *N* regroups all compounds belonging to the non-sensitizers. The data set was subdivided into a training set (about 80% of the compounds) used to develop the model and a test (including the remaining 20% of the data) used to assess the performance of the model in prediction.

The CAESAR model for skin sensitization was developed using Adaptive Fuzzy Partition (AFP) – AFP was used to develop classification models implementing a fuzzy partition algorithm. It models relations between molecular descriptors and chemical activities by dynamically dividing the descriptor space into a set of fuzzy partitioned subspaces. The aim of this algorithm is to select the descriptor and the cut position that allow us to find the maximal difference between the two fuzzy rule scores generated by the new subspaces. The score is determined by the weighted average of the chemical activity values in an active subspace *A* and in its neighbouring subspaces.

2 ECETOC - European Centre for Ecotoxicology and Toxicology of Chemicals (<http://www.ecetoc.org/>)

Table 10: Classification of skin sensitizers as defined by ECETOC

| EC3 (%) | LLNA Class |
|-----------|---------------------|
| NC | Non Sensitizer |
| ≥ 10 | Weak Sensitizer |
| 1-10 | Moderate Sensitizer |
| 0.1-1 | Strong Sensitizer |
| < 0.1 | Extreme Sensitizer |

This model was derived from 8 descriptors from Dragon software (nN; GNar; MDDD; X2v; EEig10r; GGI8; nConj; O-058). For the feature selection, a hybrid selection algorithm (HSA), which combines the genetic algorithm (GA) concepts and a stepwise approach, was used to select the best descriptors for classifying skin sensitization dataset.

Figure 13 summarizes the results obtained with this model. The percentage of TP, TN, FP and FN are shown for both training and test set.

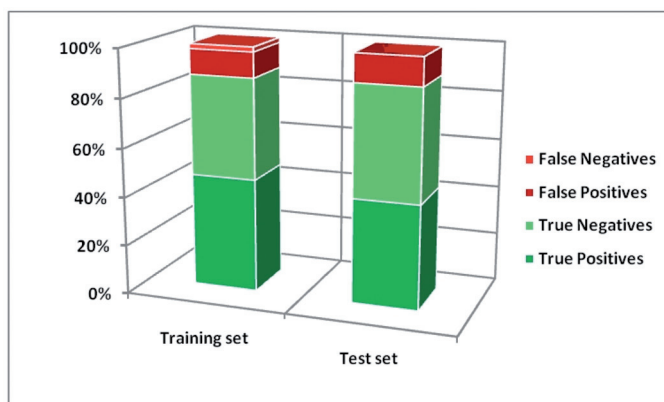


Figure 13: Results obtained with the predictive model for skin sensitization (Global Model). The percentage of molecules correctly classified (TP and TN) is around 80% for both training and test set, which represents a good result.

Local QSAR model

A complementary approach was developed to enable potential mechanisms of toxic action to be assigned to chemicals thought to be capable of skin sensitization. This approach was based on the work of Aptula and Roberts [30] who had previously suggested that for a chemical to be a skin sensitizer it must be capable (either directly or after some abiotic or metabolic transformation) of

one of five electrophilic-nucleophilic reactions. The approach undertaken was to devise SMARTS⁵ patterns capable of identifying the electrophilic mechanisms previously assigned to the 210 Gerberick LLNA dataset [29]. The 44 chemical TIMES-SS LLNA data, in which each chemical also had a mechanism of action assigned to it by the same authors [31] was then used to validate the applicability of the SMARTS patterns (Table 11).

Table 11: Confusion matrix for the TIMES-SS validation set of chemicals. This table show the confusion matrix obtained using expert assignment from reference [31].

| | Ac | pro-MA | SB | SN2 | non |
|--------|----|--------|----|-----|-----|
| Ac | 2 | | | | |
| pro-MA | | 5 | | | 4 |
| SB | | | 4 | | |
| SN2 | | | | 6 | |
| non | | | | | 23 |

Table 12: Skin sensitization mechanism of action - expert assignment vs. SMARTS assignment. The columns of this confusion matrix represent the expert assignment, whereas the rows are the assignment made with the developed SMARTS. On the main diagonal lie the molecules correctly assigned by the SMARTS method.

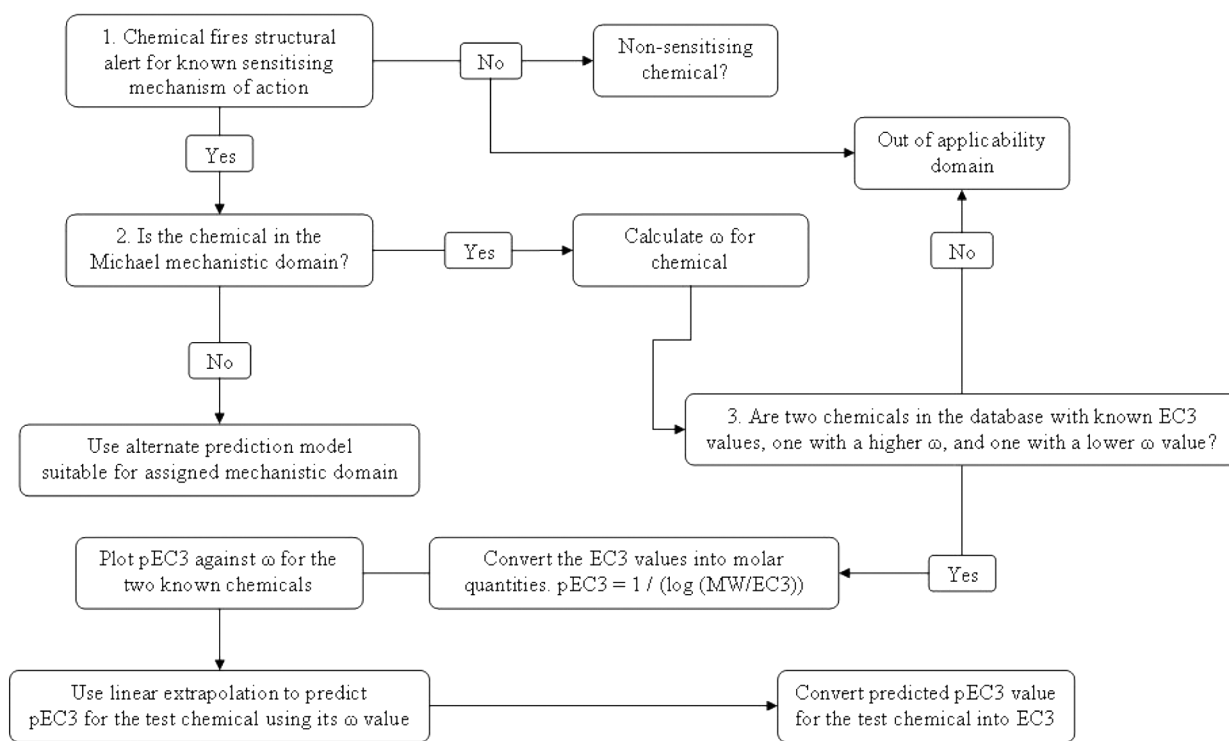
| | Ac | MA | pro-MA | SB | pro-SB | SN2 | pro-SN2 | SNAr | SN1 | non |
|---------|----|----|--------|----|--------|-----|---------|------|-----|-----|
| Ac | 24 | | | 1 | | 1 | | | | |
| MA | | 29 | | | | | | | | 2 |
| pro-MA | | | 25 | | | | | | | |
| SB | | | | 36 | | | | | | 3 |
| pro-SB | | | | | 4 | | | | | |
| SN2 | 2 | 2 | | | | 45 | | | | |
| pro-SN2 | | | | | | | 2 | | | |
| SNAr | | | | | | | | 3 | | |
| SN1 | | | | | | | | | | 1 |
| non | | | | | | | | | | 22 |

3 SMiles ARbitrary Target Specification (SMARTS) - <http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>

The results of this approach have been reported by Roberts at al. in 2007 [32]. *Table 12* represents the confusion matrix showing the classification statistics for the training data. The columns indicate the expert assignment from reference [29] and the rows report the class from the SMARTS pattern.

The ability to assign a chemical to one of the suggested electrophilic mechanisms allowed for the development of a novel quantitative read-across approach for the prediction of the skin sensitizing potential of chemicals assigned to the Michael mechanistic domain (Figure 14).

Figure 14: Read across procedure utilizing the electrophilic index (ω) as a quantitative measure of a chemical's electrophilicity.


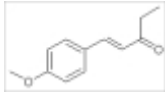
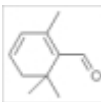

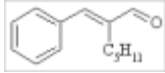


Read across procedure utilising the electrophilic index (ω) as a quantitative measure of a chemical's electrophilicity

This approach utilized the electrophilic index (ω) which has been suggested to account for a chemical's inherent electrophilicity [33]. This descriptor was calculated from the frontier molecular orbitals; these are the orbitals considered to be directly involved in the formation of the covalent

bond between the nucleophilic sulphur containing protein side chains and the electrophilic chemical thought to be the initial event required for skin sensitization. Examples of the predictions obtained by the method can be seen in ... and full details of how to carry out the calculations and predictions for the remaining chemicals assigned to the Michael domain can be found in [reference \[34\]](#).

Table 13: Example of prediction obtained using the electrophilic approach.

| Name | Structure | Experimental EC3 | Predicted EC3 | ω |
|---|---|------------------|---------------|----------|
| trans-2-hexenal |  | 5.5 | NP* | 1.608 |
| 1-(4-methoxyphenyl)-1-penten-3-one |  | 9.3 | 9.87 | 1.734 |
| safranal (1,1,3-trimethyl-2-formylcyclohexa-2,4-diene) |  | 7.5 | 5.29 | 1.796 |
| diethyl maleate |  | 5.8 | 8.74 | 1.804 |
| α -amyl cinnamic aldehyde |  | 11 | NP* | 1.839 |

* NP means a prediction has not been made is not a chemical more electrophilic (larger ω) or less electrophilic (smaller ω) in this small.

Chapter References

1. U.S. EPA Estimation Program Interface (EPI) suite - <http://www.epa.gov/oppt/exposure/pubs/episuite.htm>
2. *EC funded* project CAESAR (Computer Assisted Evaluation of industrial chemical Substances According to Regulation) - <http://www.caesar-project.eu/>
3. Life+ funded project ANTARES (Alternative Non-Testing methods Assessed for REACH Substances) - <http://www.antares-life.eu/>
4. Lombardo A., Roncaglioni A., Boriani E., Milan C. and Benfenati E., Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish, Chem Cent J. 2010 Jul 29;4 Suppl 1:S1.
5. Dimitrov S., Dimitrova N., Parkerton T., Comber M., Bonnell M. and Mekenyan O., Base-line model for identifying the bioaccumulation potential of chemicals, SAR QSAR Environ Res, 16, 531-554, 2005.
6. Arnot J.A. & Gobas F.A.P.C., A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ Rev 14: 257-297, 2006.
7. VEGA (Virtual models for property Evaluation of chemicals within a Global Architecture) - <http://www.vega-qsar.eu>.
8. Golbraikh A. & Tropsha A., Beware of q²!, J Mol Graph Model, 2002 Jan, 20(4):269-76.
9. Zhao C., Boriani E., Chana A., Roncaglioni A., Benfenati E., A new hybrid system of QSAR models for predicting bioconcentration factors (BCF), Chemosphere. 2008 Dec; 73(11):1701-7. Epub 2008 Oct 26.
10. Schüürmann G., Ebert R.U., Chen J., Wang B., Kühne R., External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean, J Chem Inf Model. 2008 Nov;48(11):2140-5.

11. Benigni R., Bossa C., Jeliaskova N., Netzeva T. and Worth. A, The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree, European Commission report EUR 23241.
12. Toropov A.A. and Benfenati E., SMILES in QSPR/QSAR Modeling: results and perspectives, *Curr Drug Discov Technol.* 2007 Aug;4(2):77-116.
13. Toropov A.A., Toropova A.P., Benfenati E. and Manganaro A., QSAR modelling of carcinogenicity by balance of correlations, *Mol Divers.* 2009 Aug; 13(3):367-73. Epub 2009 Feb 4.
14. Toropov A.A., Toropova A.P. and Benfenati E., QSAR modelling for mutagenic potency of heteroaromatic amines by optimal SMILES-based descriptors, *Chem Biol Drug Des.* 2009 Mar; 73(3):301-12.
15. Woo Y-T. and Lai D.Y., OncoLogic: A mechanism-based expert system for predicting the carcinogenic potential of chemicals. In *Predictive Toxicology*, edited by Helma C. Boca Raton FL, USA: CRC Press; 2005:385-413.
16. Lewi D.F., Bird M.G., Jacobs M.N., Human carcinogens: an evaluation study via the COMPACT and HazardExpert procedures, *Hum Exp Toxicol.* 2002 Mar; 21(3):115-22.
17. Marchant C.A., Prediction of rodent carcinogenicity using the DEREK system for 30 chemicals currently being tested by the National Toxicology Program. The DEREK Collaborative Group, *Environ Health Perspect.* 1996 Oct; 104 Suppl 5:1065-73.
18. Benigni R., Bossa C., Tcheremenskaia. O and Worth A., Development of structural alerts for the in vivo micronucleus assay in rodents, EUR 23844 EN 2009, 1-43.
19. Klopman G., Chakravarti S.K., Zhu H., Ivanov J.M. and Saiakhov R.D., ESP: a method to predict toxicity and pharmacological properties of chemicals using multiple MCASE databases, *J Chem Inf Comput Sci.* 2004 Mar-Apr;44(2):704-15.
20. Matthews E.J. and Contrera J.F., A new highly specific method for predicting the carcinogenic potential of pharmaceuticals in rodents using enhanced MCASE QSAR-ES software, *Regul Toxicol Pharmacol.* 1998 Dec; 28(3):242-64.

21. Benigni R. and Bossa C., Predictivity of QSAR, *J Chem Inf Model*. 2008 May;48(5):971-80. Epub 2008 Apr 22.
22. Benfenati E., Benigni R., Demarini D.M., Helma C., Kirkland D., Martin T.M., Mazzatorta P., Ouédraogo-Arras G., Richard A.M., Schilter B., Schoonen W.G., Snyder R.D. and Yang C., Predictive models for carcinogenicity and mutagenicity: frameworks, state-of-the-art, and perspectives, *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev*. 2009 Apr;27(2):57-90.
23. Fjodorova N., Vracko M., Tušar M., Jezierska A., Novic M., Kühne R. and Schüürmann G., Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses, *Mol Divers*. 2010 Aug; 14(3):581-94. Epub 2009 Aug 15.
24. Taskinen J. and Yliruusi J., Prediction of physicochemical properties based on neural network modelling, *Adv Drug Delivery Rev* 2003, 55:1163-1183.
25. Van der Jagt K., Munn S., Tørsløv J. and De Bruijn J., Alternative approaches can reduce the use of test animals under REACH, Addendum to the report “Assessment of additional testing needs under REACH. Effects of (Q)SARS, risk based testing and voluntary industry initiatives” 2004.
26. Höfer T., Gerner I., Gundert-Remy U., Liebsch M., Schulte A., Spielmann H., Vogel R. and Wettig K., Animal testing and alternative approaches for the human health risk assessment under the proposed new European chemicals regulation, *Arch Toxicol*. 2004 Oct; 78(10):549-64. Epub 2004 May 29.
27. Arena V.C., Sussman N.B., Mazumdar S., Yu S. and Macina O.T., The utility of structure-activity relationship (SAR) models for prediction and covariate selection in developmental toxicity: comparative analysis of logistic regression and decision tree models, *SAR QSAR Environ Res*. 2004 Feb;15(1):1-18.
28. Sussman N.B., Arena V.C., Yu S., Mazumdar S. and Thampatty B.P., Decision tree SAR models for developmental toxicity based on an FDA/TERIS database, *SAR QSAR Environ Res*. 2003 Apr;14(2):83-96.
29. Gerberick G.F., Ryan C.A., Kern P.S., Schlatter H., Dearman R.J., Kimber I., Patlewicz G.Y. and Basketter D.A., Compilation of historical local lymph node data for evaluation of skin sensitization alternative methods, *Dermatitis*. 2005 Dec;16(4):157-202.

30. Aptula A.O., Patlewicz G. and Roberts DW., Skin sensitization: reaction mechanistic applicability domains for structure-activity relationships, *Chem Res Toxicol.* 2005 Sep;18(9):1420-6.
31. Roberts D.W., Patlewicz G., Kern P.S., Gerberick F., Kimber I., Dearman R.J., Ryan C.A., Basketter D.A. and Aptula AO, Mechanistic applicability domain classification of a local lymph node assay dataset for skin sensitization, *Chem Res Toxicol.* 2007 Jul; 20(7):1019-30. Epub 2007 Jun 8.
32. Roberts D.W., Patlewicz G., Dimitrov S.D., Low L.K., Aptula A.O., Kern P.S., Dimitrova G.D., Comber M.I., Phillips R.D., Niemelä J., Madsen C., Wedebye E.B., Bailey P.T. and Mekenyan O.G., TIMES-SS—a mechanistic evaluation of an external validation study using reaction chemistry principles, *Chem Res Toxicol.* 2007 Sep;20(9):1321-30. Epub 2007 Aug 23.
33. Parr R.G., Donnelly R.A., Levy M. and Palke W.E., Electronegativity: The density functional viewpoint, *J. Chem. Phys.* 68, 3801 (1978)
34. Enoch S.J., Cronin M.T., Schultz T.W. and Madden J.C., Quantitative and mechanistic read across for predicting the skin sensitization potential of alkenes acting via Michael addition, *Chem Res Toxicol.* 2008 Feb; 21(2):513-20. Epub 2008 Jan 12.

Emilio Benfenati^a,
Rodolfo Gonella Diaza^a,
Andrea Gissi^b

*a. Istituto di Ricerche Farmacologiche
Mario Negri, Via La Masa 19, 20156,
Milano, Italy*

*b. Università degli Studi di Bari "Aldo Moro",
Piazza Umberto I 1, 70121, Bari, Italy*

The applications and the possible uses of QSAR models

The applications of QSAR models: industrial, pharmaceutical and regulatory applications

Academia has been one of the major sources of the QSAR models so far published. Most typically for academia the target is theoretical, and the applications are taken as examples. For instance, studies may address a certain relationship between the toxicity and the descriptor, proposing new descriptors as tools to better capture chemical structures, or describing a new algorithm to explore the possible links between toxicity and descriptors. In this way there have been a large number of models, with many possible combinations of endpoints, molecular descriptors, algorithms, etc. This represents a huge wealth of experience. Unfortunately, most of these models have never been used, due to their poor availability and the reduced experience with their use.

Commercial software typically codifies a certain high level of modelling capability, offering to the user an easy tool. In this case the software has been refined and gives the prediction of several endpoints in a user-friendly environment. Industry has been the main target of these kinds of tools.

Indeed, industry is using QSAR models for internal purposes, in the development of new compounds, and also to document the properties of their chemicals.

Besides commercial models, there are also public free tools with similar characteristics. The main targets for the commercial and free models are industry and regulators.

It is useful to analyse the differences on the use of QSAR models within industry and for regulatory purposes. *Table 14* summarizes some specifics of these two different approaches.

Table 14: Specific needs and characteristics of Industrial and Regulator approaches

| <i>Industry Approach</i> | <i>Regulator Approach</i> |
|---|--|
| <ul style="list-style-type: none">• Beneficial properties• Human properties• Millions of candidates• Focus on selected candidate drugs• Complex tools• Commercial tools• No false positives• Confidential data• Commonly used | <ul style="list-style-type: none">• Negative properties• Human AND Environmental properties• Hundreds/thousands chemicals• Focus on few chemicals• Simpler tools• Free tools• No false negatives• Public data preferred• Acceptability issue |

Regulatory Context

In the USA, QSAR models have been used for decades to evaluate physico-chemical, environmental, ecotoxicological, and toxicological properties. The US Environmental Protection Agency (EPA) makes available a series of QSAR models, such as EPISuite, and T.E.S.T. Furthermore, US EPA has tested a number of models (*reported in Table 15*).

Table 15: QSAR models tested by US EPA for several endpoints.

| <i>Endpoint</i> | <i>Software/Methods</i> |
|-------------------------------------|---|
| Carcinogenicity | SARs (<i>OncoLogic</i>) |
| Skin absorption | QSARs (<i>linear regression models</i>) |
| Environmental/Ecological effects | EPISUITE and PBT Profiler |
| Physicochemical properties | KOWWIN |
| Acute and Chronic Toxicity | ECOSAR |
| Environmental Fate | Modules in EPISUITE |
| Endocrine Disruption Knowledge Base | EDKB |

Interestingly, CAESAR QSAR is also available through the web site of US EPA. Indeed, the CAESAR models, developed within an EC funded project, have been implemented in collaboration with the US EPA.

Other regulatory bodies have applied and developed QSAR models, such as the US FDA (as shown in Table 16).

Table 16: US Agencies actively using and or testing in silico models.

| <i>Agency</i> | <i>Tests</i> |
|--|---|
| ATSDR (<i>Agency for Toxic Substances and Disease Registry</i>) | <ul style="list-style-type: none"> • Toxicity prediction - QSARs based on PBPK • Benchmark Dose (<i>BMD</i>) for human health effects |
| FDA (<i>Food and Drug Administration - Dept. of Health & Human Services</i>) | <ul style="list-style-type: none"> • Carcinogenicity - data from regulatory submissions used to develop MULTICASE |
| NTP (<i>National Toxicology Program</i>) | <ul style="list-style-type: none"> • Carcinogenicity - tested commercial software |
| NIOSH (<i>National Institute for Occupational Safety and Health</i>) | <ul style="list-style-type: none"> • Use of SARs for hazard alerts for Current Intelligence Bulletins |

Canada is also active in the use and evaluation of QSAR models, as illustrated in *Table 17*.

Table 17: Canadian efforts and results in using and evaluating in silico methods.

| <i>Environment Canada*</i> | <i>Health Canada**</i> |
|---|--|
| <p>Number of software have been tested:</p> <ul style="list-style-type: none"> • ECOSAR • TOPKAT • Probabilistic and computational NNs (<i>PNN</i>) • ASTER • OASIS <p>PNN was found to be the <u>most reliable</u></p> <p>TOPKAT also excellent but with smaller prediction space</p> | <p>Canadian Environmental Protection Act (<i>CEPA</i>) Complex Hazard (<i>ComHaz</i>) Tool has been used.</p> <p>Sources of Information include:</p> <ul style="list-style-type: none"> • (<i>Q</i>)SAR models (<i>TOPKAT</i>, <i>CASETOX</i>) • Chemical structures of concern, Structure Activity Relationship (<i>SAR</i>) models, such as <i>DEREK</i>, surrogate/analogue approaches (<i>Leadscope</i>, <i>visual grouping</i>) |

* *Environment Canada* - <http://www.ec.gc.ca>

** *Health Canada (Healthy Environments & Consumer Safety Branch)* - <http://www.hc-sc.gc.ca/>

In Europe different regulations hold different positions with regard to in silico models:

- the cosmetics directive (76/768/EEC) established that by 2013 all in vivo models will be banned for testing, and many of them have been banned for cosmetics already;
- The REACH legislation (1907/2006/EC) foresees use of in silico models.

In other cases, such as plant protection products and pharmaceuticals, tests on animals have to be done, at least on the parent compound. However, some tools have been developed to study ecotoxicity of metabolites and degradation products of pesticides (*more info can be found at the EC project DEMETRA website*).

Thus, in Europe there are different regulations which exhibit different attitudes towards in silico models. It may be useful to analyse the REACH legislation in greater detail, because it focuses more on the possible use of QSAR models.

The REACH legislation

In December 2006 the European Community adopted a new regulation addressing the production circulation of chemical substances in the European territory, and their potential impacts on both human health and the environment. This new regulation is called REACH (*Registration, Evaluation, Authorisation and restriction of Chemicals*) and states that, for each chemical circulating in the European territory, a complete dossier on physico-chemical, biological and toxicological properties has to be compiled. In order to prevent an over-usage of animal testing, REACH regulation foresees and promotes the use of alternative methods (*such as QSAR*) stating that:

Before new tests are carried out to determine the properties listed by REACH, all available in vitro data, in vivo data, historical human data, data from valid (Q)SARs and data from structurally related substances (read-across approach) shall be assessed first.

For more information about REACH, refer to ANNEX 5 (*Ref to annexes “Annex 5. REACH”*). The full text of the REACH legislation is available through the European Chemicals Agency (*ECHA*) official website (<http://echa.europa.eu/en>).

Use of QSAR models within a REACH perspective

Sometimes people say that alternative methods, such as QSAR models, are considered to save animals. This is incorrect. In fact, there are several reasons to use in silico methods:

1. **Innovation:** REACH explicitly encourages innovation in toxicity evaluation. The development of alternative methods is one of the purposes of REACH. This is clearly stated in the first article of the regulation. Within REACH, innovation refers to new chemical products, of course, but also to new methods to protect human health and the environment. The legislation acknowledges that there are current limitations on the methods of assess toxicity and environmental impact. For these reasons it anticipates periodic updates, to take new methods into account. In silico models usually are considered a surrogate of existing methods. Actually, in

silico models should also be considered as a way to cover existing gaps of knowledge. The legislation sets out conditions specifically for the use of QSAR models, and the European Chemicals Agency (ECHA) offers detailed guidance on their use. Even the introductory ‘Guidance in a nutshell’ on substance registration advises industry to ‘collect QSAR estimated results for the substance if suitable models are available’ as an initial step.

- 2. Time for the experiments:** The REACH legislation requires industry to evaluate the toxicity not just of new chemicals, but also of the tens of thousands of existing chemical substances that are currently in use but have never been subject to regulatory testing. Many argue that to achieve this by traditional in vivo testing would take decades.
- 3. Lack of laboratories and resources:** There is a lack of laboratories in Europe capable of performing in vivo tests for such a large number of substances. For instance, the Italian FEDERCHIMICA conducted a survey on the available laboratory in Italy. It was found that certain endpoints, listed by REACH, are not covered by Italian laboratories, and very few can cover most of the necessary endpoints.
- 4. Reduction of costs:** In vivo experiments for all registered compounds would be costly. In this context, in silico models are potentially vital because they use computer technology to connect, use and extend existing experimental data. When experimental data sets are already available from in vivo or in vitro tests, then in silico methods can be used to assess thousands of chemicals in a day. In previous regulation there was no need to use in silico methods. It is only with new legislation that the need has become apparent.

“...there is no choice but to use in silico technologies, just to get the work done.”

Bob Diderich, OECD, interviewed for the ORCHESTRA documentary: ‘QSARs in REACH?’

Uses, issues and priorities.’

In silico methods therefore offer an important means for European industry to meet the REACH demands, and to do so at a cost that will not destroy SMEs and/or move chemical industries and testing to other parts of the world. In silico methods can be used on an office

computer, so even with the costs of expert advice, they are cheaper than in vivo testing by orders of magnitude.

5. Animal testing as ‘a last resort’: REACH demands the use of existing data where possible, and states that further animal testing can only be used ‘as a last resort’:

“The challenge is to have scientifically sound information on the potential hazards of substances whilst at the same time minimizing unnecessary animal testing. One of the fundamental aims of REACH is to promote alternative methods for assessing hazards of substances and to see animal testing as a last resort. All parties involved should take this very seriously...”

Geert Dancet, Executive Director of ECHA, [Press Release: August 2009](#)

Reducing or avoiding the use of many additional millions of vertebrates to test existing chemicals is not just a policy priority, it also matches the desire of many industry shareholders, managers and consumers, and is a major concern for citizens and NGOs.

6. **Prioritization:** In silico methods offer a valuable tool for prioritising substances according to their toxicity, so that costly experimental tests can be focussed on those substances with higher probability of toxicity and higher risk. Well-informed prioritisation can reduce economic costs, time delays and the use of vertebrates. Thus, QSAR can be very useful for ECVHA and Member States, to direct attention to the more risky compounds.
7. **The focus on ‘weight of evidence’:** REACH demands that industry review all the available evidence, which in many cases will include evidence produced by QSAR models. The demand to review all the evidence is part of a regulatory shift away from relying on a single in vivo study, towards reviewing the ‘weight of evidence’ developed from a range of complementary sources.

QSAR models are recommended by ECHA as a valuable source of evidence, and supporting evidence, in a ‘weight of evidence’ approach. In this QSAR is capable of spotting data gaps in order to identify certain critical molecular features which deserve more attention. QSAR has also been instrumental in some cases in noticing erroneous experimental values, due to the fact that QSAR is supported by a broad set of experimental values.

8. **Pro-active decision-making:** The predictive ability of in silico methods enables a pro-active approach to toxicity within product development. This would be very useful to chemical industry. Toxicity evaluation can be brought ‘upstream’ in the product development and decision making processes, so that chemicals are selected, and products are developed, to be less toxic or non-toxic. In this way future toxicity, and therefore future costs and future impacts on health and the environment, can be reduced or avoided.
9. **The future of toxicology:** REACH creates a potential turning point in regulatory toxicology. Looking ahead, and beyond Europe, the Tox21 and ToxCast initiatives in the USA are expected to reshape the toxicological procedures for evaluation of chemicals. By providing up-to-date toxicity data for thousands of chemicals, these major projects will put in silico models in a central role. Experience and understanding of in silico models will be essential for analysing and making full use of those data in future toxicity evaluations.

The REACH requirements for QSAR

REACH legislation foresees the use of alternative in silico methods such as QSAR and Read-Across. Regarding Qualitative or Quantitative structure-activity relationship ((Q)SAR), Annex XI states that:

Results obtained from valid qualitative or quantitative structure-activity relationship models ((Q)SARs) may indicate the presence or absence of a certain dangerous property. Results of (Q)SARs may be used instead of testing when the following conditions are met:

- results are derived from a (Q)SAR model whose scientific validity has been established,
- the substance falls within the applicability domain of the (Q)SAR model,

- results are adequate for the purpose of classification and labelling and/or risk assessment, and
- adequate and reliable documentation of the applied method is provided.

Thus, these are the requirements according to REACH. Three of them refer to the QSAR model, and one to its use for a specific compound.

These principles are also expressed in four of the OECD principles for the QSAR validation, whereas there is no explicit mention of the fifth OECD principle within REACH.

Model validity

The first requirement points out that the method should be scientifically valid. We notice that it is not requested that the model be validated. Validation is a formal process, which takes many years. The validation process of a QSAR model would probably end after REACH. The validity has to be assessed with regard to scientific criteria, considering the performance of the model in its results in prediction (*ECHA - Guidance for the implementation of REACH*).

The OECD principles specify some features of the QSAR model, in order to assess if it works or not; these include the statistical characteristics of the model itself and its predictive properties. We notice that, for regulatory purposes, during in the early years of the QSAR development the interest was in the properties of the model addressing the results in fitting, i.e. based on the chemical used to build up the model, the emphasis is placed on the possibility for a model to predict the property of a new compound. In other words, it should be shown that it works for the purpose (*proof-of-principle*).

How to evaluate the scientific validity of the QSAR model?

Within Annex XI REACH requires that a QSAR model is “scientifically valid” (*it does not say validated*). A first proof of the scientific validity can be the fact that the method has been published

in a scientific journal through the peer-review process. In this case, the method has been evaluated by other scientists, who found the method suitable for publication.

The preliminary document on (Q)SAR characterisation, compiled by the body formerly known as the European Chemicals Bureau (*ECB*), lists a series of statistical parameters to be used for the model evaluation. Different tools apply to a model which is a classifier, or to a model which is a regression method. In the first case the output of the model is a class or category, such as toxic, or mutagen.

According to the OECD Guidance Document on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment, the term validation is defined as “the process by which the reliability and relevance of a particular approach, method, process or assessment is established for a defined purpose”.

In this definition, the performance of a model refers to how well it fits, its robustness and predictive ability, whereas purpose refers to the scientific purpose of the QSAR, as expressed by the defined endpoint and applicability domain. So a QSAR can be valid, because the model has scientific relevance, without being relevant for a given regulatory purpose: in fact, the regulatory relevance of the model expresses the usefulness of the predicted endpoint in relation to the information needed for the regulatory purpose.

The applicability domain

Requirements 1, 3 and 4 serve to verify if the model is valid, transparent and adequate for REACH. However, all these factors, which refer to the model evaluation, are not sufficient. Requirement 2 in Annex XI calls for showing that the model is appropriate for the chemical it has been applied to. This requirement refers to the applicability domain of the model. Thus, conceptually this requirement refers to the possible application of the model to the chemical compound of interest, while the other requirements evaluate the model per se.

Thus, according to REACH both the quality of the model and its adequacy for a given chemical form the basis for the evaluation of a QSAR model. This requirement is also present in the OECD principles.

How to evaluate the applicability domain?

There are some chemometric (*chemometrics is a statistical area which combines statistics and chemistry*) tools which use the chemical descriptors and/or fragments of the chemicals used to build up the model, and compare if the chemical descriptors and/or fragments of the target chemical are similar. An example of this approach is given by the freely available software AMBIT, developed within the Cefic Long-range Research Initiative (*LRI*). A major disadvantage of this approach is that it is based only on the chemical information.

Another approach is to evaluate the metabolism or toxicity pathway of the chemical of interest. However, this can be applied only in cases where they are known.

Another recent tool has been developed within the ORCHESTRA and ANTARES projects. The tool, working within the VEGA platform, takes into account both the chemometric information and the toxicity predictions made by the model, and in particular what kinds of errors have been made by the model. Thus, this approach is based not only on the input space (*the chemical descriptors and fragments*), but also the output space of the model, which is the predicted property. Furthermore, this tool is based not only on the a priori data and information, like the other approaches, but also on the a posteriori result of the model.

The tool uses several perspectives for its reasoning, and it combines different parameters into a single value. This value ranges from 0 to 1, and it is associated with an acceptability evaluation of the model results for a certain chemical. The user knows if the model can or cannot be used for a certain compound. In some cases a warning is given, recommending expert opinion. In all cases the reasons for the reliability is given, and it can be evaluated in a transparent way.

This tool is useful to explore the results of the model, linking the prediction with results obtained on similar compounds. This supports the model reliability in a transparent way, and extends the use of the QSAR as a tool for data exploration on similar compounds, which is also very useful as a basis for read across. Indeed, a major issue in read across is the evaluation of similarity.

A valid QSAR will be associated with at least one defined applicability domain in which the model makes estimations with a defined level of accuracy (*reliability*); when applied to chemicals within its applicability domain, the model is considered to give reliable results. There is no unique measure of model reliability; in fact it should be regarded as a relative concept, depending on the context in which the model is applied.

However, it is always important to question whether a specific QSAR is appropriate for the compound of interest. This involves first considering if the chemical of interest is within the scope of the model, according to the defined applicability domain. Clearly, the more explicit the definition of the model domain, the easier this question will be to answer. The second consideration consists in evaluating the suitability of the defined applicability domain for the regulatory purpose. This question arises because most currently available models were not tailor-made for current regulatory needs and inevitably incorporate biases which may or may not be useful, depending on the context of prediction. Such biases do not affect the validity of the model, but they affect its applicability for specific purposes. The third aspect to be considered is how well the model predicts chemicals that are similar to the substance of interest. This question provides a simple way of checking whether a model is appropriate by verifying its predictive capability for one or more analogous compounds that are similar to the compound of interest and for which measured values exist. Finally, it is important to assess if the model estimate is reasonable, taking into account other information. This inevitably necessitates an expert judgment, which should be clearly rationalized.

The model adequacy of QSAR prediction

The third requirement refers to the specific context of the application of the model. REACH clearly states that there are several possible ways to use QSAR models. So far, most of the debate has only addressed the possible use of QSAR for the dossier preparation. However this is not correct. QSAR can be used also for classification and labelling, and in this case different conditions apply. Furthermore, QSAR can be used for prioritization, as we discussed.

Even in the case of the use of QSAR for the registration, the debate focussed on the use of QSAR as key study, as the only source of information. However, QSAR can be used, even for registration, as additional study within a broader context, such as weight of evidence.

When applying these conditions in the context of a chemical assessment, it is necessary to consider the completeness of the overall information.

Furthermore, if a registrant intends to use QSAR data instead of experimental data, the adequacy of the QSAR result should be documented. Different types of QSAR Reporting Formats (*QRFs*) are being developed to provide a standard framework for summarizing and structuring key information

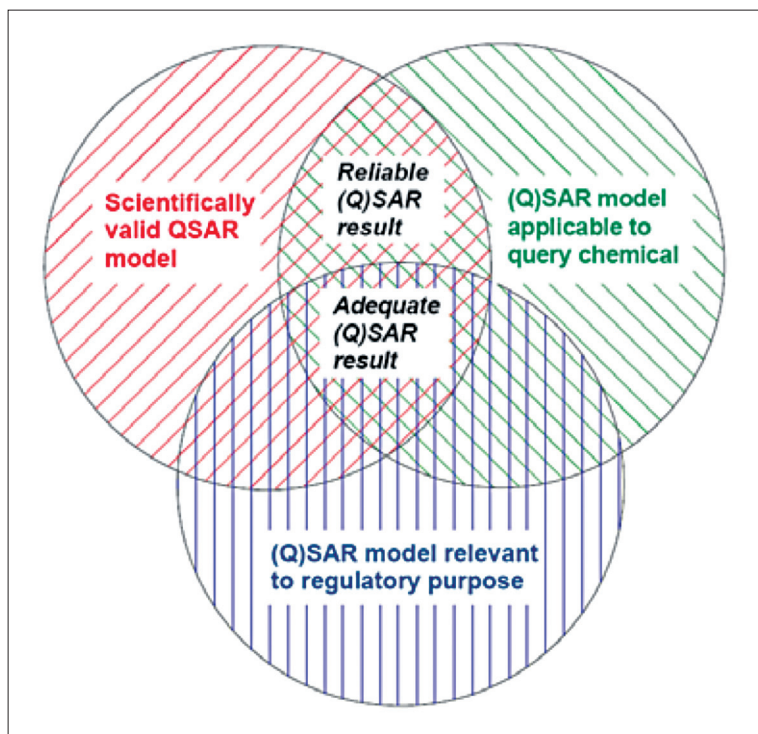


Figure 1: Interrelated concepts of QSAR validity, reliability, applicability, adequacy, regulatory relevance. The circles refer to (Q)SAR models whereas the intersections refer to (Q)SAR results with certain features. In order for a (Q)SAR result to be reliable for a given chemical, it should be generated by a scientifically valid (Q)SAR that is also applicable to the chemical of interest. This (Q)SAR estimate may or may not be adequate (fit for purpose), depending on whether the endpoint predicted is relevant to the particular regulatory purpose, and whether the estimate is sufficiently reliable for that purpose.

about QSAR models and their predictions. In the first one, the QSAR Model Reporting Format (QMRF), is stored the description of a particular QSAR model (*i.e. description of the algorithm, of its development and validation based on the OECD principles*). The second one, the QSAR Prediction Reporting Format (QPRF), explains how an estimate has been derived by applying a specific model or method to a specific substance (*i.e. information on the endpoint, identities of close analogues, etc.*). The last one, Totality of Evidence Reporting Format (TERF) or Weight of Evidence Reporting Format (WERF), has not been developed yet, but it will be useful to integrate the QSAR estimates with other sources of information based on Weight of Evidence considerations.

The documentation and the model transparency

The fourth requirement asks for transparency. This is reasonable, since all documentations at the basis of the assessment of the properties of a chemical should be clearly available and verifiable.

One of the driving forces of REACH was to have the correct knowledge of the properties of the chemical substances on the market. If some of the information is hidden this clearly goes against the spirit of REACH. The availability of the components of the model is also within the OECD principles, which explicitly ask for the definition of the endpoint (*the property*) in the first principle and the equation in the second principle.

The transparency should refer to all components of the model: the toxicity data, the chemical structures, the chemical descriptor and the algorithms. Indeed, it may be critical to have two different QSAR models, some of the parts of which are obscure, or confidential, which produce two different results. It would not be possible to check the reason for the different results, with the consequence of a lack of reliability. We notice that this requirement implies that models which have confidential or restricted components may be not suitable for REACH. The restriction may be with regard to the property values used to build up the model, or the information on the chemical part, or the mathematical equation. So far there has been no official position against the use of confidential models in Europe. The US FDA accepts the use of commercial software, which includes confidential data as a basis.

QSAR models in the regulatory perspective

Regulatory QSAR is a specific type of QSAR which is very demanding because it relates to the law. This introduces requirements, some internal to the QSAR model process, others external. There are two conceptual approaches for the use of the QSAR models:

- in the first case the QSAR model mimics the current method;
- in the second case the QSAR model addresses the lack of knowledge which still is not covered by the current method, such as animal models (*Figure 2*).

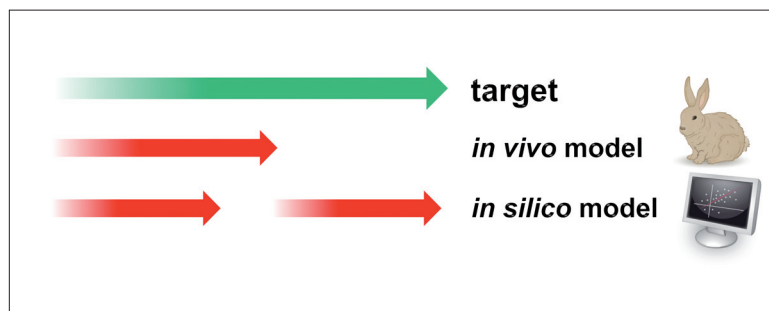


Figure 2: *In silico models can be used to cover gaps in experimental data.*

In the first case we refer to the approach, so far adopted, of using animal models. In very few cases does the current methodology refer to *in vitro* methods, like mutagenicity studies using the Ames test. In this approach the QSAR model should adhere as much as possible to the traditional model. In the second approach a new strategy is adopted, recognizing that current approaches do not offer good solutions. Depending on the endpoint, the two approaches may be accepted. Surely for endpoints where the current methods are broadly used, the first approach seems more attractive. For the second case we also recall what we discussed about the usage of QSAR models, and the call of innovation introduced by REACH. Here we will analyse in more detail the approach of using QSAR models as a substitute for existing animal methods.

Typically the *in vivo* experiment used to produce the datum has to refer to a certain suitable guideline. Furthermore, the use and acceptability conditions also have to have regard to the law; for instance, the toxicity value has to be used for classification and labelling, referring to certain toxicity classes with given thresholds; in other cases the value has to be used for risk assessment, and thus has to be a continuous value to be compared to the exposure value. Thus, while the typical QSAR scheme is as follows:

Data → Information → Knowledge

in which *Data* are the raw data (*toxicity and chemical ones*), *Information* is the elaborated data, as processed during the modelling phase, and *Knowledge* is the final output of the model, now the scheme has to keep into account the specific context and is modified as follows:

Method → Data → Information → Knowledge → Use

in which *Method* is the method to get the experimental data, as defined by the law, and *Use* is the use defined by the law of the toxicity value or class. This has dramatic consequences on the

modelling process, in each phase. The data must refer to a certain method; the uncertainty of this method should be known; the modelling process should keep in mind the thresholds identified by the law, and the model results should be checked against the specific use, considering the number of false negatives (*compounds which are predicted safe while they do not*), as well as the uncertainty of the QSAR model, etc. Different laws have different thresholds and limits. This defines specific boundaries and requirements both for the input and the output of the QSAR model. Not all kinds of inputs are equally correct. Similarly, not all outputs are good. We have to remember the REACH requirements discussed above, referring to the requirement of the use, suitable for REACH.

The QSAR for specific, regulatory purposes is like a bridge offering a new way to proceed. The predicted toxicity value should be suitable for the intended use. A good bridge has to take into account the specific constraints on both sides of the bridge, and the intended traffic. This means that the bridge has to be robust and quality checked for a demanding traffic flow, providing the soundest possible basis.

Different perspectives for a broader QSAR scenario

There may be a tendency to consider QSAR for regulatory purposes as the top level of QSAR models. This is inexact. Above we specified that QSAR for regulatory purposes represent a specific application. Other applications exist. In different circumstances different criteria may apply. Even though it is obvious that the attention to the quality of the model is an important factor in all cases, the purpose of the model should modulate the relevance of the different aspects of the QSAR model, and how it is built.

For instance, some models may address the drug design, and in this case the model may be tailored to minimize false positives, not false negatives. Indeed, for the drug industry it is important to identify good drugs, avoiding proceeding with expensive experiments with chemicals which later on will fail as drugs. In another phase of the process the toxicity evaluation will be introduced, and this will need different criteria.

In this evolving, complex scenario, new tools are being introduced. It is important to exploit all possible efforts to elucidate the reasons for toxicity, in order to reduce the use of toxic compounds. No single method is sufficient to cope with such a huge task. Rather, a wise integration of different tools is the right solution. This requires effort to facilitate the exchange of information arising from different tools, enabling the dialogue between different components of a more complex system, and thus helping its constituent parts to interact.

The QSAR model and the role of the expert

As we said, there are many models and many purposes, which may lead to different results, and different understanding of the same results. Indeed, there are some models which are flexible, and are developed as tool for exploration of the properties, giving several parameters to be selected and optimised. Conversely, other models have fixed parameters and the user will derive only one result. These two strategies represent different approaches.

As previously mentioned, some models require manual intervention by the expert, in order to better optimize, for instance, the tri-dimensional structure of the chemical to be modelled. Thus, different results may be expected. Other models, like the OECD toolbox, also require the careful identification of the suitable parameters, to be chosen by the expert on a case by case basis. Different results are obtained in this case too, depending on the parameters which are defined by the user.

Models like EPISuite, Toxtree, VEGA, and also many commercial packages have been optimized by the developers, and do not require the selection of the parameters, in most cases. These models are specific for selected endpoints, and some models include many of them. However, the possible models are finite. Conversely, the OECD toolbox aims to be a general framework which would ideally be applicable to all endpoints, and for this reason a high degree of flexibility is given.

As a result, the OECD toolbox requires a high degree of experience, while the other pre-optimized models are intentionally more user-friendly, easier, and more reproducible. Reproducibility is centrally important. For regulatory purposes the model uncertainty is very important, and the same result should be obtained in all countries by all users (*such as regulatory bodies and industries*). For this reason, the OECD toolbox needs a certain level of experience.

HOW MUCH EXPERTS AGREE ON THE EVALUATION OF THE RESULTS FROM QSAR MODELS?

Rodolfo Gonella Diaza^a, Anna Lombardo^a, Alberto Manganaro^a, Emilio Benfenati^a, Ilias Kotinas^b

^a Istituto di Ricerche Farmacologiche Mario Negri, Laboratory of Environmental Chemistry and Toxicology, Via G. La Masa 19, 20156 Milan, Italy
^b University of Patras, Human Computer Interaction Group, University campus, 26504 Rio Patras, Greece

Introduction

Within the EC project ORCHESTRA (<http://www.orchestra-qsar.eu>), in collaboration with the Gruppo di Lavoro sui metodi QSAR of the Italian Ministry of Health (Ministero della Salute), we did an exercise on the evaluation of the results of the QSAR models for bioconcentration factor (BCF).

The purpose of this exercise has been to see if there is an agreement or not on the human experts' evaluation of the reliability of the QSAR results. This exercise should not be considered a way to verify if the results of a certain QSAR model are correct or wrong. See the poster on the comparison of several QSAR models done on 860 compounds at QSAR 2012.

Methods

The exercise has been done on three freely available models for BCF: VEGA, EPISuite and T.E.S.T..

We used three chemicals (Fig. 1). We asked experts to evaluate the results which were previously obtained by us. The experts were evaluating the same report and information, and were not asked to use the model.

We obtained replies from about 30 experts from Europe and USA. Replies were anonymous.

| Molecule / Models | Mol. 1 | Mol. 2 | Mol. 3 |
|-------------------|------------------------------------|------------------------------------|------------------------------------|
| VEGA | 3.13 | 1.56 | 2.72 |
| EPISuite | Meylan: 2.83 Arnot&Gobas: 2.733 | Meylan: 1.47 Arnot&Gobas: 1.499 | Meylan: 4.08 Arnot&Gobas: 3.014 |
| T.E.S.T. | 2.79 | 1.75 | 2.18 |

Fig. 1. The three chemicals used for the exercise with the predicted values for each model.

Results

1 The acceptability of the results depends on the chemical (the three chemicals of the exercise are quite different in their complexity) and on the documentation in support to the prediction that the model can give. For example, the user has the information about the behaviour of the target compound in relation to logP, compared with the full set of other chemicals (Fig. 2).

Furthermore, the user can visualize the trend with similar compounds (Fig. 3) and the most similar compounds are also presented (Fig. 4).



Fig. 2. Comparison between MlogP and log BCF (VEGA output).

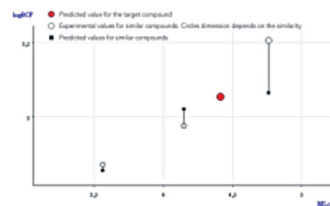


Fig. 3. Comparison between MlogP and log BCF for the 3 most similar compounds and the target compound (VEGA output).

| (a) | (b) |
|---|--|
| <p>CAS: 108-67-0 Dataset: 141 (training set) SAR: ES: ClogP; ES: ClogP Similarity: 0.988 Experimental value: 2.55 (log₁₀ K_{ow}) Predicted value: 2.18 (log₁₀ K_{ow})</p> | <p>2 Mol. 1 Mol. 2 Mol. 3</p> |
| <p>CAS: 96-45-0 Dataset: 67 (training set) SAR: ES: ClogP; ES: ClogP Similarity: 0.917 Experimental value: 2.59 (log₁₀ K_{ow}) Predicted value: 2.2 (log₁₀ K_{ow})</p> | <p>108-79-0 Mol. 1 Mol. 2 Mol. 3</p> |
| <p>CAS: 88-23-2 Dataset: 67 (training set) SAR: ES: ClogP; ES: ClogP Similarity: 0.968 Experimental value: 3.43 (log₁₀ K_{ow}) Predicted value: 2.57 (log₁₀ K_{ow})</p> | <p>100-67-4 Mol. 1 Mol. 2 Mol. 3</p> |
| <p>CAS: 91-57-8 Dataset: 20 (test set) SAR: ES: ClogP; ES: ClogP Similarity: 0.888 Experimental value: 2.41 (log₁₀ K_{ow}) Predicted value: 2.72 (log₁₀ K_{ow})</p> | <p>100-67-4 Mol. 1 Mol. 2 Mol. 3</p> |

Fig. 4. The lists of similar compounds for VEGA (a) and nearest neighbor method of T.E.S.T. (b).

- Users were able to evaluate the results provided by each model in terms of the expected reliability for a certain compound. It means that the evaluated QSAR models are able to provide clues useful to appreciate when a prediction is reliable (Fig. 5).
- There was a good agreement between the replies for the different models, showing a certain consensus.
- In case of EPISuite, for the first compound and also for the others, some experts criticized the lack of applicability domain. For this reason we implemented the Meylan model within VEGA.

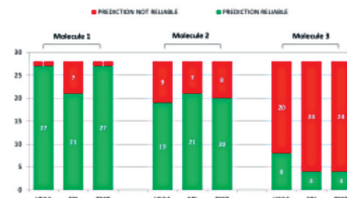


Fig. 5. Results of the exercise on model reliability evaluation.

Conclusions

- Experts users can extract useful information from the advanced QSAR models.
- Users can appreciate that there are chemicals which are more difficult to be predicted.
- A good agreement can be reached when the prediction seems reliable. Consensus between more models has been evaluated as useful. Thus, efforts on the integration of different perspectives can provide an advantage to the user.

We acknowledge all experts who kindly participated to the review exercise
We acknowledge the EC for funding the project ORCHESTRA (FP7-226521), and the Italian Ministry of Health



There is another important matter in which the expert plays an important role: the evaluation of the result. In the event of the same result being obtained by the same model for the same chemical, there may be a degree of confidence. In this case it is difficult to evaluate if the lack of confidence is generally related to the QSAR models, or to the specific QSAR model, or to the specific compound. The ORCHESTRA project studied and put effort in improving a better understanding of this issue (*e.g. through a series of exercises directed to regulators*). Results of these studies have been summarized and presented at QSAR2012¹ (Figure 3).

Generally speaking, in Europe the QSAR models for regulatory purposes are a novelty, compared with the USA, where they have been used for decades. In Europe there is certain scepticism with regard to QSAR models, for cultural reasons. In some cases there may also be a role played by industry consultants who historically sold evaluations using animal models used within their laboratories. In any event, it is clear that Europe's acceptance of the results from QSAR models is growing cautiously. Criticisms are that the models are uncertain, that different models may give different results, and that the results are unclear. In this situation, the expert on the model can be a figure facilitating its acceptance, since the expert can recognize critical matters related to the use of the model.

Furthermore, documentation is fundamental. The transparency of the model and the way it is reported is a requirement, for regulatory purposes, in order to make the process verifiable. This is clearly mentioned in the Annex XI of REACH. Transparency is related to reliability, and has already been discussed.

Figure 3: "How much experts agree on the evaluation of the results from QSAR models". The ORCHESTRA project organized an on line exercise, providing three molecular structures and the report on the prediction of bioconcentration factor (BCF) obtained from three software (VEGA², EPISUITE³ and TEST⁴). Experts in QSAR and toxicology were invited to evaluate the results obtained by this software and to judge the reliability of these results.

1 [QSAR2012: 15th International Workshop on Quantitative Structure-Activity Relationships \(QSAR2012\) in Environmental and Health Sciences, June 18th - 22nd, 2012, Tallinn, Estonia.](#)

2 [VEGA: Virtual models for property Evaluation of chemicals within a Global Architecture.](#)

3 [EPISUITE: U.S. EPA Estimation Program Interface Suite.](#)

4 [TEST: U.S. EPA Toxicity Estimation Software Tool.](#)

An ideal framework for the development of QSAR for regulatory purposes

A huge united effort has to be made to produce models suitable for regulatory purposes. Scientists should avoid the attitude of restricting aspects of QSAR models. Contributions to each part of the model should be acknowledged. Industry can play a major role by making data accessible. Indeed, industry is the main source of data, and it should discuss internally a strategy regarding access to data which maintains the ownership of the data for sensitive values. Regulators should promote initiatives working towards a larger dissemination and use of data and models.

On the basis of the above considerations, the QSAR community should increase synergies to better cope with the task. The VEGA platform has been developed in this spirit. As we have seen, no single approach can solve all problems. Furthermore, there will surely be new tools, which are continuously being introduced. We should take advantage of different tools in the attempt to obtain validated models. The ideal situation should be making freely available all model components. In this way it will be possible to get the most out of each component and adopt different strategies, combining different models. If the different models are proprietary, the powerful integration of the different components will be much more difficult.

Models should be fully transparent, including data. It is expected that data will be more and more available publicly, and present in several sources, such as the IUCLID database, thanks to the REACH legislation, the OECD toolbox, etc. This process will attract further data. However, there will be other data not publicly available, for obvious reasons, such as confidential data on chemicals under development. Companies with confidential data may have a use for QSAR models as test sets.

Ideally, chemical structures of the substance used in the QSAR models for regulatory purposes should be carefully checked. A manual check is still preferable, but it poses problems in the case of data sets of thousands of chemicals.

Several chemical formats are currently used in QSAR models. When a transformation from one format to another is made, the chemical structure should be checked, because it is possible to introduce modifications in the chemical structures. Care should be taken considering the exact chemical structure suitable to describe the correct stereochemistry of the chemical used in the laboratory experiment, when the chemical has chiral centres.

Ideally, chemical descriptors should come with their correct description, including the mathematical equation/way to get them. We mentioned that the change of software versions to calculate descriptors might be a problem. This can be solved by producing freely available tools to calculate chemical descriptors, and leaving available the old version when a new one is completed. When a new version of the software is used to calculate the chemical descriptors, the reproducibility of the results should be checked. More reproducible descriptors are preferable. However, this point is not so easy, because even 2D descriptors may give different results, on the basis of tautomerism, for instance, and there are several programs for log P which produce different results [1].

Ideally, all mathematical parameters of the model algorithm should be known. For a model algorithm, XML is the preferable format. In the ideal situation, within a unified XML system, the user can import toxicity data from one source, get the structure from a second source, apply the model and obtain the results seamlessly, while even referring to different sources. Some projects addressed this goal [2,3].

The debate and the open issues

Another change occurred during QSAR evolution; indeed, more general models were introduced, addressing heterogeneous chemical classes, while the original QSAR approach was based on classes of highly homogeneous compounds.

Besides the modifications related to the introduction of different technical and scientific tools, there has been a change related to the mentality at the basis of the QSAR. While the classical approach was to have a certain hypothesis (such as the role of lipophilicity), and generate a model reflecting such a hypothesis, some of the new tools are based on model generation without any a priori hypothesis: the system is capable of exploring a wide set of possible relationships and identifying the more likely ones, producing new hypotheses. The higher number of toxicity data, for a larger set of chemicals, the more difficult it becomes to manually explore all the possibilities. Computers represent a valuable aid in screening.

There has been a debate on this change. Those who prefer the classical approach object that the new approach generates a model which is too complex and incomprehensible. The reply of the other modellers is that the important aspect is to have a predictive model, and that even

descriptors, which seem simple, are complex, and the hypothesis of mechanisms must be proved. Thus, log P is actually calculated through software which calculates tens of parameters, invisible to the final user and each of them difficult to put in a relationship with the toxicity. Another example is the parameter HOMO (*highest occupied molecular orbital*) energy, calculated with complex mathematical equations, based on a series of assumptions.

We believe that both approaches should be pursued since both offer good results. The ideal situation would be to combine the different approaches, since in this way it is possible to reduce the knowledge gap. It is a pity that too much effort is made to show that one approach is better than the other. For the user's benefit all possibilities should be explored and efforts should be made to improve the model performance.

Related to reproducibility is the fact that the parameters of the model algorithm and the descriptors should be fixed. This requires a clear codification of the algorithms and descriptors. Algorithms, even those which are apparently complex, can be codified in a way that they result in a series of coefficients and variables. In the development phase, complex tools can be used, such as genetic algorithms, to find best descriptors and neural networks for model optimisation. Then, once the model is defined, it can be implemented as a relatively simple equation, as has been done for instance within the DEMETRA models, in which complex hybrid models resulted in much simpler equations with certain descriptors [4,5]. An important aspect is that descriptors should be fixed as well. This can be critical, because typically descriptors are calculated with software, which are proprietary or otherwise out of the control of the QSAR modeller. The software developers of the chemical descriptors quite frequently change the algorithms in the different versions, and this may bring about irreproducible results.

We described above what an ideal situation would be. Keeping in mind the goal of having a wide range of tools to improve the safety of industrial chemicals for regulatory purposes, there are several issues, which have been discussed above, that address the OECD principles. There are some advantages to having model components freely available. The main advantages are:

Wider use of the QSAR models by regulators, industry, non-governmental organisations, and scientists. The free availability can of course make the tool applicable regardless of a chemical's tonnage, improving societal protection. Nowadays, due to the cost of the assessment, for a chemical on the market in amount lower than a tonne, the REACH provisions are not applicable, and only a minimum set of evaluations is done for chemicals produced between one and ten tonnes, which are the category with the highest number of chemicals. The cost of the models will represent a barrier.

Better possibility to integrate different tools, making optimal exploitation of any specific model component. Nowadays it may happen that a certain model is powerful because it employs more data, yet the algorithm itself may be poor. Or in another case the algorithm is powerful, but the chemical descriptors poor. In the ideal situation, scientists may combine and test different components, achieving better results.

The whole process will be more transparent, convincing more users. Commercial components are usually restricted. The reproducibility will also be improved, providing some reassurance that a crucial part of the model will not disappear from the marketplace.

Obviously other solutions are present and forthcoming, involving commercial solutions. Commercial companies may offer a more user-friendly environment, assistance, and dedicated solutions for specific problems of industrial interest. For instance, a chemical company may wish to explore a large number of compounds, using some confidential data it has. In this case, focussed models can be built for the specific target.

Chapter References

1. Benfenati E., Gini G., Piclin P., Roncaglioni A. and Vari M.R., Predicting LOGP of pesticides using different software, *Chemosphere*. 2003 Dec;53(9):1155-64.
2. Chemomentum: Grid Services Based Environment to Enable Innovative Research - <http://www.unicore.eu/community/projects/chemomentum.php>.
3. OpenMolGRID: Open Computing GRID for Molecular Science and Engineering <http://www.openmolgrid.org>.
4. Quantitative structure-activity relationships (QSAR) for pesticide regulatory purposes, edited by Benfenati E., Elsevier, Amsterdam, The Netherlands. 510pp, 2007.
5. DEMETRA: Development of Environmental Modules for Evaluation of Toxicity of pesticide Residues in Agriculture - <http://www.demetra-tox.net>

Annex 1

QSAR and SAR models: Basic definitions

Quantitative Structure-Activity Relationship (QSAR) models correlate the properties and molecular structure of a chemical with its biological effect on human health and/or on relevant species in an ecosystem. The correlation can then be used in the prediction and assessment of new substances. In the so-called Structure-Activity Relationship (SAR) approaches the quantitative aspect of the phenomenon is not addressed, and the study refers to categories (*e.g. Toxic/Non-toxic*). The expression (Q)SAR is commonly used to cover both cases. In case that the modelled feature is a property, the expression QSPR is also used. Here we will refer to QSAR for simplicity's sake, and discuss the case of SAR explicitly later on.

Sometimes the expression *in silico* methods is used. This refers to the fact that computers are used, and computers have silicon in the hardware. However, *in silico* methods is a broader expression, which includes tools such as docking methods, which are not properly QSAR and are aimed at studying the interactions between the receptor and the ligand.

Giuseppina Gini^a, Luigi Cardamone^a,
Magdalena Gocieva^a, Marina Mancusi^b,
Rima Padovani^b, Lorenzo Tamellini^b

a. Politecnico di Milano, Piazza L. da Vinci 32, 20135,
Milano, Italy

b. Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129,
Torino, Italy

Annex 2

An introduction to toxicology

Toxicology (from the Greek words *toxikos* and *logos*) is the study of the adverse effects of chemicals on living organisms. It is the study of symptoms, mechanisms, treatments and detection of poisoning, especially the poisoning of people and the environment pollution.

Poisons are substances that can cause damage to organisms, leading to a deterioration of their main vital functions, illness, or death. Virtually, every chemical substance may be harmful or lethal when a sufficient concentration is absorbed by an organism. Paracelsus, sometimes called the father of toxicology, wrote: “All things are poison and nothing is without poison, only the dose permits something not to be poisonous.” That is to say, substances often considered toxic can be benign or beneficial in small doses, and conversely an ordinarily benign substance can be deadly if over-consumed.

The toxic dose differs a lot according to the specific chemical that is considered. Some substances induce death by concentration of few micrograms per kilogram, while others may be quite toxic even if their concentrations are much higher (some grams per kilogram). At the same time it is not easy to quantify the toxicity of a substance because several factors need to be considered to understand and eventually predict the main phenomena in the field of toxicology. Thus several

indexes have been proposed to estimate the toxicity of a substance and to make comparison among different chemicals. One of the most used indexes is LD₅₀ (*Lethal Dose*), which is the dose (*mg/kg body weight*) that is responsible of the death of the 50% of the animals exposed to the different chemical agents.

| Substance | Animal, Route | LD ₅₀ |
|--|--------------------------------|--------------------------------------|
| Vitamin C (ascorbic acid) | rat, oral | 11,900 mg/kg |
| Grain alcohol | young rat, oral | 10,600 mg/kg |
| Table Salt | rat, oral | 3,000 mg/kg |
| THC (main psychoactive substance in Cannabis) | rat, oral | 1,270 mg/kg males; 730 mg/kg females |
| Caffeine | rat, oral | 192 mg/kg |
| Nicotine | rat, oral | 50 mg/kg |
| Strychnine | rat, oral | 16 mg/kg |
| Aflatoxin B1 (from <i>Aspergillus flavus</i>) | rat, oral | 0.048 mg/kg |
| Batrachotoxin (from poison dart frog) | human, sub-cutaneous injection | 0.002-0.007 mg/kg (estimated) |
| Polonium 210 | human, inhalation | 0.00001 mg/kg (estimated) |
| Botulinum toxin (Botox) | human, oral, injection | 0.000001 mg/kg (estimated) |

Table 2.1: LD50 values for different substances

When a chemical agent or one of its metabolites produces a toxic effect, it has to interact with specific sites of the organism and to be present in adequate concentration for a sufficient period of time. As a consequence it is important to know which effects a particular substance may cause, information about its chemical structure, the characteristics of the exposure (administration mode, exposure time and rate) and the features of the organism.

In particular, the toxic actions of all substances are carried out by the alteration of biochemical and physiologic processes of cells. Cellular death is the direct consequence of the damage induced by chemical agents and it may lead to serious effects on the tissue in question. In fact many toxic responses stem from cellular death and the loss of efficiency of crucial organs that affect the functionality of the whole organism. Other responses exist which are not due to cellular death, but depend on the unbalancing of the physiological and biochemical processes.

Chemicals influence such processes by different modes of action that can be synthesized as follows:

- interference in normal ligand-receptor interactions;
- interference in membrane functions;
- interference in cellular energy production;
- binding/influence on biomolecules;
- alteration of calcium homeostasis;
- toxicity due to the death of specific cells;
- nonlethal genetic alteration of somatic cells.

Many substances have toxic effects because they interfere in the normal ligand-receptor interaction. Ligand-receptor interaction can be defined as the interaction between a molecule and a protein on or within a target cell. Receptors are macromolecular tissue components that a drug or a chemical agent (*ligand*) interact with to produce a biological effect [12]. This particular kind of bond is reversible and highly selective, that is to say also small changes in chemical structure may drastically reduce or cancel the binding effect. An example of this mode of action is given by neurotoxic substances, because the functions fulfilled by the central nervous system (*CNS*) are highly dependent on neurotransmitter-receptor interaction.

Other substances may alter the functionality of excitable membranes, of membranes of cellular organelles, of lysosomal membranes and of mitochondrial membranes. The insecticide DDT, for example, interferes in the closing of sodium channel and then alters speed of repolarization, which is specific to excitable membranes.

In other cases chemicals become toxic interfering in the normal oxidation process of carbohydrates that leads to the production of Adenosine Tri-Phosphate (*ATP*). Some of them interfere in the oxygen supply to tissues whereas others, like cyanide, stop the use of oxygen in tissues, because of the affinity to a specific enzyme. The subsequent reduction of *ATP* stores may cause a range of consequences, such as the alteration of the functionality of the membranes and of the ion pumps, and the inhibition of protein synthesis, leading to the loss of the functionality and the death of cells.

Some toxicants often bind or at least affect the normal use of biomolecules (*proteins, lipids and nucleic acids*). One of the most famous examples is carbon monoxide, which binds iron in haemoglobin with high affinity, causing the reduction of oxygen supply to tissues. Other agents promote the production of intermediates, especially free radicals, that bind biomolecules (like lipids and nucleic acid), inducing the loss of the functionality of the cell membrane and the alteration of basic intracellular functions, such as protein synthesis.

The inference in the normal processes that are responsible of the intracellular calcium homeostasis seems to play a crucial role in cellular damage or death caused by chemical agents. Some membrane abnormalities develop when great amounts of calcium ions accumulate in isolated cells, as a consequence of the toxic effect of some chemicals. Furthermore, calcium is a second messenger in the regulation of numerous functions. For example, the normal organization of cellular cytoskeleton is altered when calcium concentration rises or some endonucleases, activated by calcium ions, may induce the DNA fragmentation and the chromatin condensation, which are important processes involved in cellular apoptosis.

Another kind of toxic effect is the selective death of cells within an organ or a tissue, which sometimes looks like particular pathologic processes. The human embryo, for example, is very sensitive to the action of many toxicants. The administration of a particular drug against nausea (*Thalidomide*) to pregnant women may cause the loss of some undifferentiated embryonic cells, leading to abortion or to some congenital malformations.

Finally, a particular group of chemical substances (*xenobiotics*) binds DNA molecules, inducing cellular death or promoting a complex series of events that may generate cancer. Such substances are called genotoxic carcinogens. Most of the lesions chemically induced to DNA are repaired, but some of them may be missed or incorrectly repaired, causing the introduction of an altered gene, which the new cellular generation will inherit. If the mutation affects a somatic cell, the genetic lesion will not be transmitted to the future generation, but it could generate a cancer. Genotoxic substances seem to be able to induce cancer, since they activate some proto-oncogenes, whose expression is strictly controlled in normal cells. However the induction of cancer is a process which depends on several factors, since also substances that are not genotoxic may increase the incidence of the pathology, perhaps by some mechanisms which are different from DNA damage (*cancer promoters*).

Furthermore it is necessary to note that the classification of the mode of action referring to a specific process or site of action is difficult to make and above all it is not exclusive. The cyanide binds a particular enzyme with a specific affinity, so this bond looks like the ligand-receptor

interaction. On the other hand this interaction inhibits the enzymatic activity, causing the reduction of energy stores, which may influence biomolecular use and alter the calcium homeostasis. Another example is given by lead, whose several toxic effects may be partly imputed to the bond with specific proteins, while others are not easily attributable to a specific biochemical mode or a particular enzyme.

Finally it is important to emphasize that the mode and the site of action of the majority of substances are not well known yet. This knowledge gap is actually the main problem that has to be faced when assessing the toxicity of a chemical by alternative methods, such as QSARs. Nowadays several steps forward have been made in the field of the development of structure-activity relationships, thanks to innovative algorithms and calculation tools, but great efforts should still be made to improve toxicological knowledge in order to promote a wider and more effective use of such methods.

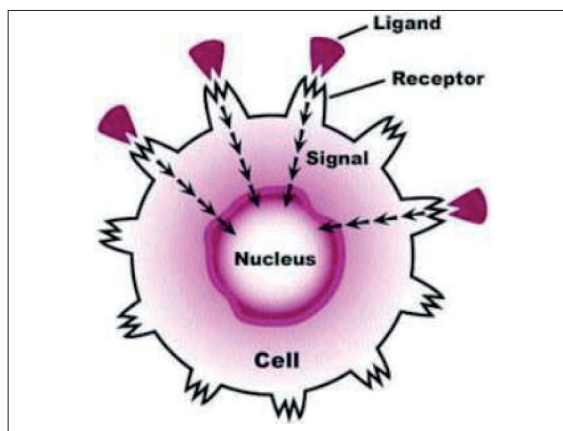


Figure 2.1: Functional scheme of the ligand-receptor interaction. This interaction is very selective and it influences the inner cell mechanisms.

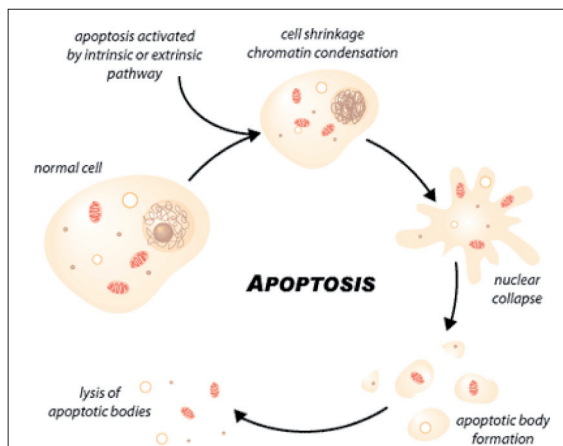


Figure 2.2: Description of the apoptosis process. Although many pathways and signals lead to apoptosis, there is only one mechanism that actually causes the death of the cell in this process; after the appropriate stimulus has been received by the cell, that cell will undergo the organized degradation of cellular organelles by activated proteolytic caspases, through different phases: cell shrinkage and rounding, chromatin condensation into compact patches against the nuclear envelope, DNA fragmentation, nucleus breaking into several discrete chromatin bodies or nucleosomal units due to the degradation of DNA, irregular buds formation on the cell membrane (known as blebs), cell breaking apart into several vesicles called apoptotic bodies, which are then phagocytosed.

Andrey A. Toropov^a, Alla P. Toropova^a,
Emilio Benfenati^a, Giuseppina Gini^b

*a. Istituto di Ricerche Farmacologiche Mario Negri, 20156,
Via La Masa 19, Milano, Italy*

*b. Department of Electronics and Information,
Politecnico di Milano, Via Ponzio 34/5,
20133 Milano, Italy*

Annex 3

Rigid and flexible topological indices: variety of the representations of the molecular structure

113

Introduction

The Monte Carlo method is one of the universal computational technology <http://mathworld.wolfram.com/MonteCarloMethod.html>. Rigid and flexible topological indices are a tool for representation of the molecular structure for physical, chemical, biological and medicinal problems. The rigid topological index has the same value for all endpoints, whereas flexible index is a mathematical function of collection of substances and selected endpoint.

Quantitative structure - property/activity relationships (QSPR/QSAR) is a field of theoretical chemistry that is starting by works of H. Wiener [1-4]. QSPR/QSAR are “main employer” of the topological indices, both rigid and flexible. The main idea of these studies is the use of a molecular graph or more exactly a matrix of topological distances calculating special coefficients (descriptors), which can be correlated with the properties of organic compounds.

From the beginning of 1980s, a number of different descriptors conceptually related to the Wiener number started to increase [5-14]. Most these descriptors or indices were based on two matrixes: the adjacency matrix and the above mentioned matrix of topological distances in a molecular graph.

Considering for instance the molecular graph of 2-methyl butane, with numbering of vertices as in *Figure 3.1*

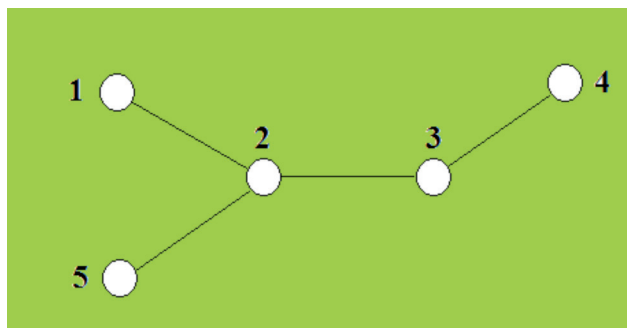


Figure 3.1: The molecular graph of 2-methyl butane (CAS 78-78-4)

the adjacency matrix $A(G)$ and matrix of topological distances $D(G)$ are the following

$$\begin{array}{c}
 \begin{array}{c}
 \mathbf{1} \quad \mathbf{2} \quad \mathbf{3} \quad \mathbf{4} \quad \mathbf{5} \\
 \mathbf{1} \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
 \mathbf{2} \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\
 \mathbf{3} \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\
 \mathbf{4} \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\
 \mathbf{5} \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0}
 \end{array} \\
 \mathbf{A(G)}
 \end{array}
 \end{array}
 \end{array}
 \end{array}
 \end{array}
 \end{array}$$

$$\begin{array}{c}
 \begin{array}{c}
 \mathbf{1} \quad \mathbf{2} \quad \mathbf{3} \quad \mathbf{4} \quad \mathbf{5} \\
 \mathbf{1} \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{2} \\
 \mathbf{2} \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{1} \\
 \mathbf{3} \begin{bmatrix} \mathbf{2} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{2} \\
 \mathbf{4} \begin{bmatrix} \mathbf{3} & \mathbf{2} & \mathbf{1} & \mathbf{0} & \mathbf{3} \\
 \mathbf{5} \begin{bmatrix} \mathbf{2} & \mathbf{1} & \mathbf{2} & \mathbf{3} & \mathbf{0}
 \end{array} \\
 \mathbf{D(G)}
 \end{array}
 \end{array}
 \end{array}
 \end{array}
 \end{array}$$

In spite of the existence of a large number of descriptors [15], the main idea of their calculation can be illustrated with the Wiener number (W) and connectivity indices of zero-order (0c) and first-order (1c), the latter is also known as Randić index [8,12-15]. For example, these topological indices for 2-methyl butane (Figure 3.1) are calculated with the formulae:

1.
$$W = 0.5 \sum \sum dij = 18$$
2.
$${}^0\chi = \sum \delta_i^{-1/2} = \delta_1^{-1/2} + \delta_2^{-1/2} + \delta_3^{-1/2} + \delta_4^{-1/2} + \delta_5^{-1/2} =$$

$$= (1)^{-1/2} + (3)^{-1/2} + (2)^{-1/2} + (1)^{-1/2} + (1)^{-1/2} = 4.284$$
3.
$${}^1\chi = \sum \delta_i \delta_j^{-1/2} = (\delta_1 \delta_2)^{-1/2} + (\delta_2 \delta_3)^{-1/2} + (\delta_3 \delta_4)^{-1/2} + (\delta_2 \delta_5)^{-1/2} =$$

$$= (1 \cdot 3)^{-1/2} + (3 \cdot 2)^{-1/2} + (2 \cdot 1)^{-1/2} + (3 \cdot 1)^{-1/2} = 2.269$$

where δ_i is vertex degree equal to the sum of the elements of the adjacency matrix in a row, and d_{ij} is the element of the D(G) matrix.

The detailed information on the other structural descriptors applied in the last years for carrying out QSPR/QSAR analyses can be found in [15, 16].

However, the unlimited growth of a number of various descriptors is not an ideal situation for the user of QSPR/QSAR models, for practical reasons.

Furthermore, there are theoretical issues. Some are general ones. For instance: which descriptors are the best ones? Others are more specific, relative to some descriptors. For instance, why degree in Eq. (2) and Eq.(3) have been selected as $-1/2$, and not $\pm 1/7$, or 3.14, etc.? Indeed, the QSAR analysis of several molecular properties [17] has shown that in most cases the optimal value of the degree, in models based on descriptors similar to the Eq.(2) or Eq. (3), is different from $-1/2$.

Another simple question about formulae as in Eq. (2) and/or Eq.(3) is the following: which numbers of neighbours connected to a vertex in molecular graph are the best basis for constructing correlation between property/activity and descriptors?

QSPR/QSAR analysis [18-21] has shown that the correction of the adjacency matrix (e.g. ethyl isopropyl sulphide CAS 5145-99-3, Figure 3.2) by means of changes of zeros on diagonal by special

selected coefficients x and y (Figure 3.3) can produce considerable improvement of correlation between the connectivity indices (${}^0\chi$ and ${}^1\chi$) and various parameters:

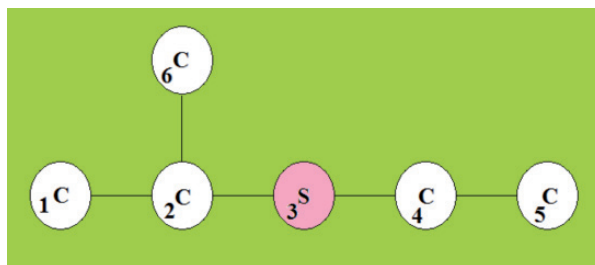


Figure 3.2: The hydrogen-suppressed graphs (HSG) the ethyl isopropyl sulphide (CAS 5145-99-3).

Adjacency matrix

| | 1 | 2 | 3 | 4 | 5 | 6 | δ_i |
|---|---|---|---|---|---|---|------------|
| 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 1 | 0 | 1 | 0 | 0 | 1 | 3 |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 2 |
| 4 | 0 | 0 | 1 | 0 | 1 | 0 | 2 |
| 5 | 0 | 0 | 0 | 1 | 0 | 0 | 1 |
| 6 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |

Modified adjacency matrix

| | 1 | 2 | 3 | 4 | 5 | 6 | δ_i |
|---|-----|-----|-----|-----|-----|-----|------------|
| 1 | x | 1 | 0 | 0 | 0 | 0 | $1+x$ |
| 2 | 1 | x | 1 | 0 | 0 | 1 | $3+x$ |
| 3 | 0 | 1 | y | 1 | 0 | 0 | $2+y$ |
| 4 | 0 | 0 | 1 | x | 1 | 0 | $2+x$ |
| 5 | 0 | 0 | 0 | 1 | x | 0 | $1+x$ |
| 6 | 0 | 1 | 0 | 0 | 0 | x | $1+x$ |

It has to be noted that these changes of the adjacency matrix involve changes of the δ_i . Similar modifications have been carried out for matrix of topological distances [22]. Also in this case an improvement of the QSPR models has been reached. These descriptors have been named “variable” or “flexible”, however, we shall use the term “optimal descriptors”, because descriptors which are discussed here are calculated by means of the Monte Carlo method optimization.

The mentioned optimal descriptors [18-22] have been based on the hydrogen-suppressed molecular graph.

Construction of optimal descriptors

Most probably, the first optimal descriptor, based on hydrogen-suppressed graph, has been suggested by Randić [23, 24]. The main idea of the approach was to use diagonal entries of the adjacency matrix for taking into account of influence of heteroatoms, similarly to the well known generalization of the Hückel molecular orbitals method [23].

The optimal descriptors have been successfully used as a tool of the QSPR modelling properties of aliphatic alcohols [25], nitrogen-containing compounds [26], and sulfides [27]. Later on, optimal descriptors based on the hydrogen-filled graph have been suggested [28].

In Refs. [23-27] as a target function the standard error of estimation (s) has been used (*i.e. searching for minimum of the s*), whereas in Refs. [28,29] as a target function the correlation coefficient (r) has been used (*i.e. searching for maximum of the r*). In [29] the comparison of these two target functions has been carried out. It has shown that the maximization of correlation coefficient gave models with preferable statistical quality. By the way, such optimisation procedure (maximum of correlation coefficient) required less number of cycles (*iterations*).

Optimal descriptors based on HSG and HFG

As an example of the general scheme of the hydrogen-suppressed graphs based approach we consider in *Figure 3.2* for ethyl isopropyl sulfide (CAS 5145-99-3).

Accordingly to [30], the use of $x=+0.25$ and $y=-0.95$ in the calculation of the optimal connectivity index ${}^1\chi(x,y)$ for the correlation with normal boiling points of 21 sulfides gave the more accurate model in comparison with the “rigid” ${}^1\chi(0,0)$.

An optimal descriptor is a modification of rigid coefficients (*e.g., vertex degree, path of length 2, etc.*), in the calculation of a “classic” bi-dimensional descriptor, by a numerical value that provides improvement of the statistical characteristics of the correlation between the descriptor and the property/activity of interest. In other words, each descriptor is a mathematical function of the

representation of the molecular structure (MSR). Any MSR contains molecular invariants (MI), which define the molecular individuality. An MSR can be represented by

$$4. \quad D = F(MSR) = F(MI_1, MI_2, \dots, MI_m)$$

where MI_k is the k-th molecular invariant ($k=1, m$, the m is the total number of molecular invariants in the molecule). The descriptor that is calculated with Eq. 4 is the “rigid” version which is similar to descriptors calculated with Eq.(1), Eq.(2), and Eq.(3).

Formula (4) can be modified by replacing “rigid” components MI_k by flexible ones $CW(MI_k)$:

$$5. \quad D = F(MSR) = F(CW(MI_1), CW(MI_2), \dots, CW(MI_m))$$

where $CW(MI_k)$ is the correlation weight of the k-th molecular invariant. The descriptor D calculated with Eq. 5 is a flexible version of the descriptor calculated with Eq. 4. The correlation weights $CW(MI_k)$ are numerical coefficients used in the calculation with Eq.(5).

The correlation coefficient between a descriptor calculated with Eq.(5) and the property/activity (PA) of interest is also a mathematical function of the CWs,

$$6. \quad R(PA, D) = R[PA, F(CW(MI_1), CW(MI_2), \dots, CW(MI_m))]$$

where $R(PA, D)$ is the correlation coefficient between the PA and D , calculated with Eq. (5).

By means of the Monte Carlo method optimisation procedure one can calculate the $CW^*(MI_1)$, $CW^*(MI_2)$, ... $CW^*(MI_m)$, which, being placed in Eq.(6) give maximum of the $R(PA, D)$ for the training set of the compounds under consideration. The predictability of the model can be tested with an external set of compounds.

As an important possibility of the scheme based on the Eq. (6), these correlation weights can be calculated not only for numerical invariant of a molecular graph such as vertex degrees [28], extended connectivity of increasing orders [31], paths of length [9] 2,3,..., valence shells of increasing orders [9], but also for not numerical features of a molecular structure, such as presence of different atoms, presence/absence of different cycles, and so on. The optimization of correlation weights of the local and global graph invariants (OCWLGI) is the basic principle of

building up optimal descriptors considered in this article. Thus, these descriptors can be named OCWLGJ-descriptors.

Owing to this possibility (involving of local and global invariants in the modelling) one can estimate a measure of the influence for a given molecular attribute (invariant of molecular graph) on the property/activity of interest, that can be used as a hint on the study of mechanism related to the phenomena (property/activity) under consideration.

A comparison of the hydrogen-suppressed graph based and the hydrogen-filled graph (HFG) based optimal descriptors has been carried out in Ref. 32. It has been shown that the optimal descriptor based on the hydrogen-filled graph gave better model for the normal boiling points of alkyl alcohols.

OCWLGJ-descriptors based on GAO

The graph of atomic orbitals (*GAO*) is an attempt to take into account the structure of atoms in QSPR/QSAR analysis [31-37].

The conversion of the hydrogen filled graph into *GAO* can be carried out by the scheme:

1. Each vertex of the hydrogen filled graph is replaced by the group of atomic orbitals. Such groups of the atomic orbitals are listed in *Table 3.1*.
2. Elements of the adjacency matrix of the graph of atomic orbitals are defined as

$$a_{ij} = \begin{cases} 1, & \text{if } i\text{-th and } j\text{-th } GAO \text{ vertices fall in different atoms} \\ & \text{in HFG and these atoms have joint edge in HFG} \\ 0, & \text{otherwise} \end{cases}$$

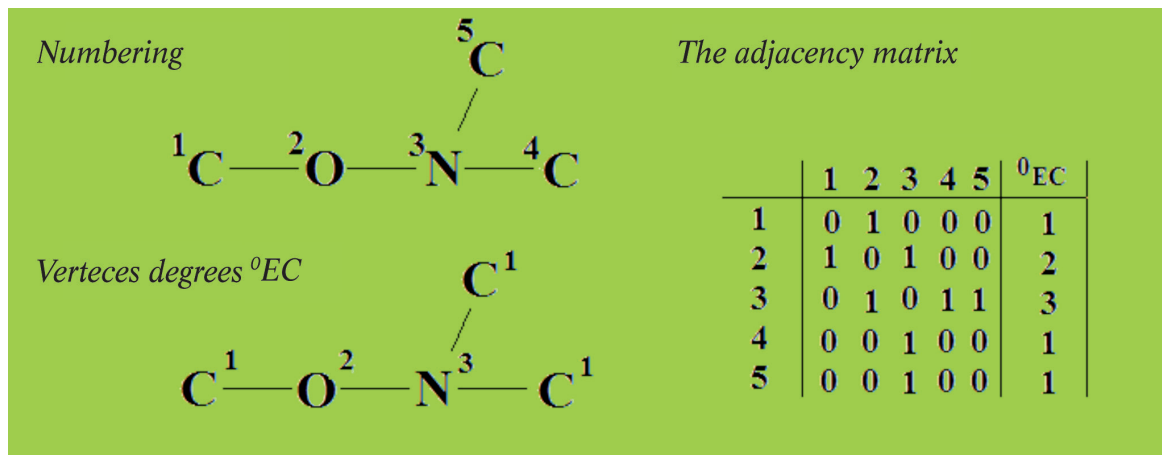
The groups of atomic orbitals for some chemical elements are listed in *Table 3.1*. In fact, the groups of atomic orbitals is the electron configuration of a chemical element.

Table 3.1: Groups of atomic orbitals, for some chemical elements, used in constructing the graph of atomic orbitals

| Chemical elements | Group of Atomic Orbitals |
|-------------------|---|
| H | 1s ¹ |
| C | 1s ² , 2s ² , 2p ² |
| N | 1s ² , 2s ² , 2p ³ |
| O | 1s ² , 2s ² , 2p ⁴ |

For the hydrogen - suppressed graph of Trimethylhydroxylamine (CAS 5669-39-6) with vertex numbering shown in Figure 3.4, this conversion gives the GAO shown in Figure 3.5.

Figure 3.3: The numbered HSG of Trimethylhydroxylamine (CAS 5669-39-6)



For a training set of graphs of atomic orbitals, one can carry out the same Monte Carlo method optimisation of correlation weights of the invariants using the same algorithms [31-38]. However, the models based on the HSG or HFG and GAO-based model are different. For the same list of compounds the number of different vertexes as well as vertex degree values for the GAO representation is larger. It improves statistical quality of a model (calculated with the optimal descriptors) for the training set, but statistical quality for the external test set can be poor [31]. Consequently, one should be careful with the GAO, since this representation of the molecular structure sometimes leads to the overtraining [31]. However, quite good GAO-based models are also possible [31-34].

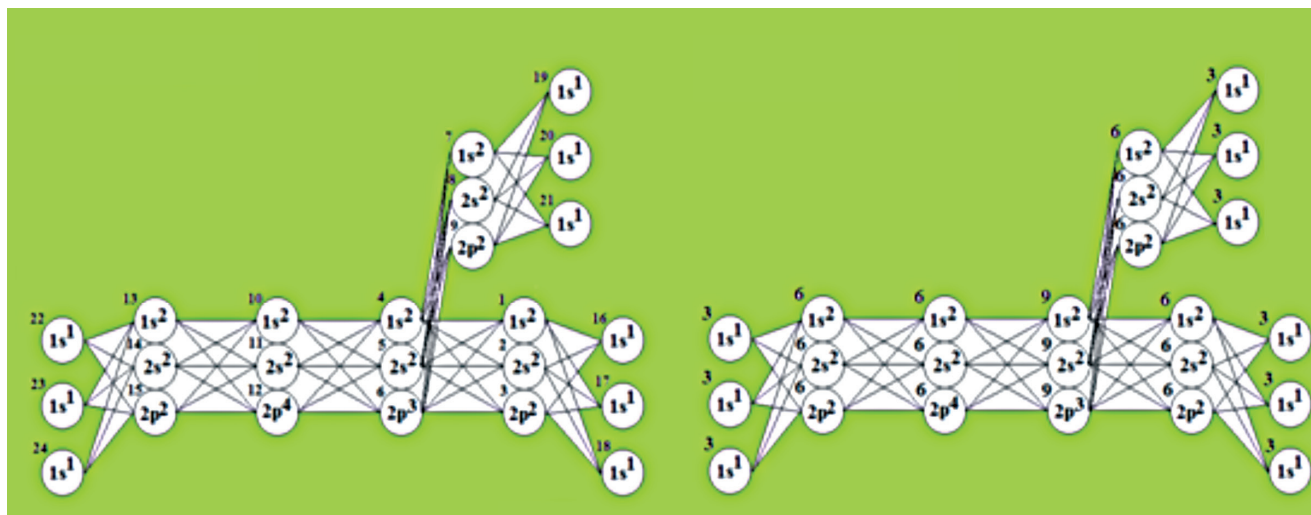


Figure 3.4: The numbered, accordingly HSG from Figure 3.3, GAO of Trimethylhydroxylamine (CAS 5669-39-6)

Collection of QSPR/QSAR based on OCWLG I-descriptors (optimal descriptors)

QSPR/QSAR models in this study are characterized by the number of compounds in a set (n); the square of correlation coefficient (r^2); standard error of estimation (s); and Fischer F-ratio.

The generalized form of the optimal descriptor is the following:

$$7. \quad DCW = CW(G) + \sum_{k=1}^N CW(A_k) + \sum_{k=1}^N W(VI_k)$$

where N is the number of vertex in molecular graph, i.e. the number of atoms in the case of HSG (Figure 3.2) and HFG (Figure 3.4) and the number of AO in the case of GAO (Figure 3.5); $CW(A_k)$ is the correlation weight of chemical element (or AO in the case of the GAO) which is an image of the k -th vertex; $CW(VI_k)$ is the correlation weight of the invariant of k -th vertex such as vertex degree;

Morgan extended connectivity [31,38], valence shells [9, 47], the number of paths of length 2 or 3 [47]; CW(G) is the correlation weight of a global invariant of the molecular graph, such as the number of cycles [7, 47], hydrogen bond indices [39], etc. As an alternative of the additive scheme for optimal descriptor (i.e. Eq. 7), the multiplicative scheme [31] can be used. Table 2 contains a collections of QSPR/QSAR models which were built up with OCWLGI-descriptors.

Table 3.2: A collection of QSPR/QSAR models based on the optimal descriptors

| <i>Description</i> | | <i>Statistical characteristics</i> | <i>Reference</i> |
|-----------------------------------|---|--|------------------|
| <i>Physicochemical parameters</i> | | | |
| QSPR 1 | QSPR models for normal boiling points of alkanes, alkylbenzenes, and polyaromatic hydrocarbons | n = 70, r ² = 0.9988, s = 5.8° C, F = 57437 (training set) n = 70, r ² = 0.9985, s = 6.7° C, F = 45154 (test set) | [38] |
| QSPR 2 | QSPR modelling of the constants of stability of 110 biometal M ²⁺ complexes with α-amino acids and phosphate derivatives of adenosine | n = 55, r ² = 0.9843, s = 0.279, F = 3328 (Training set) n = 55, r ² = 0.9935, s = 0.248, F = 4027 (Test set) | [39] |
| QSPR 3 | QSPR model of normal boiling points of alcohols | n=29, r ² =0.9906, s=2.9°C F=5733 (training set) n=29, r ² =0.9896, s=3.0°C F=2595 (test set) | [32] |
| QSPR 4 | QSPR model of normal boiling points of acyclic carbonyl compounds | n = 100, r ² = 0.972, s = 6.12°C, F = 3464 (training set) n = 100, r ² = 0.975, s = 6.00°C, F = 3905 (test set) | [40] |
| QSPR 5 | QSPR model of normal boiling points of normal boiling points of haloalkanes (GAO) | n = 138, r ² = 0.9841, s = 9.80°C, F = 3464 (training set) n = 138, r ² = 0.9854, s = 7.39°C, F = 3905 (test set) | [35] |
| QSPR 6 | The challenged study involved predictions of normal boiling points for organic molecules of varied composition. These molecules included species with both linear and cyclic structures, comprise ketones, esters, aldehydes, nitriles, amines, alcohols, and hydrocarbons and a wide variety of atoms, such as C, H, O, N, Si, Cl, Br, F, P, and S | n = 126, r ² = 0.9279, s = 33.3°C, F = 1599 (training set); n = 32, r ² = 0.8819, s = 39.1°C, F = 224 (test set); | [41] |

| | | | |
|-----------------------------|--|---|------|
| QSPR 7 | The short list of additional examples relevant to this review includes: Normal boiling points of alkanes | n = 67, r ² = 0.9984, s = 1.126°C, F = 39180 (training set); n = 67, r ² = 0.9910, s = 2.553°C, F = 7118 (test set); | [31] |
| QSPR 8 | Flory-Huggins parameter for binary polymer-solvent mixtures | n = 30, r ² = 0.9990, s = 0.028, F = 27537 (training set); n = 30, r ² = 0.9972, s = 0.053, F = 10294 (test set); | [42] |
| <i>Biomedical endpoints</i> | | | |
| QSAR 1 | Anti-HIV-1 activity TIBO and HEPT derivatives | n = 37, r ² = 0.8688, s = 0.557, F = 232 (training set) n = 20, r ² = 0.8759, s = 0.588, F = 127 (test set) | [43] |
| QSAR 2 | Toxicity, <i>V. fischeri</i> , log(1/IGC ₅₀), valence shells has been used as local graph invariant | n = 45, r ² = 0.8299, s = 0.402, F = 210 (training set) n = 21, r ² = 0.8902, s = 0.339, F = 154 (test set) | [44] |
| QSAR 3 | Toxicity to <i>Tetrahymena pyriformis</i> of heterogeneous set of benzene derivatives | n = 157, r ² = 0.883, s = 0.27, F = 1170 (training set); n = 60, r ² = 0.863, s = 0.28, F = 372 (test set); | [45] |
| QSAR 4 | The mutagenic activities of 95 heteroaromatic compounds in <i>S. typhimurium</i> TA98 S9, graph of atomic orbitals | n = 47, r ² = 0.7637, s = 1.05, F = 145 (training set); n = 48, r ² = 0.7569, s = 0.86, F = 144 (test set) | [34] |
| QSAR 5 | Aquatic toxicity, <i>Pimephales promelas</i> , log(1/LC50), Morgan extended connectivity is used as local graph invariant | n = 44, r ² = 0.8982, s = 0.251, F = 371 (training set) n = 25, r ² = 0.9181, s = 0.234, F = 258 (test set) | [46] |
| QSAR 6 | Acute toxicity LC ₅₀ -96h to rainbow trout (<i>Oncorhynchus mykiss</i>) of 274 organic pesticides | n = 233, r ² = 0.7689, s = 0.75, F = 769 (training set) n = 41, r ² = 0.6421, s = 1.14, F = 70 (test set) | [47] |
| QSAR 7 | Carcinogenic activity of methylated polycyclic aromatic hydrocarbons | n=30, r ² =0.8909, s=0.689 (training set) n=16, r ² =0.9247, s=0.594 (test set) | [51] |
| QSAR 8 | Lipophilicity (logP) of 76 industrial chemicals | n=36, r ² =0.8857, s=0.500, F=279 (training set) n=36, r ² =0.9251, s=0.382, F=414 (test set) | [52] |
| QSAR 9 | The mutagenic activities of these compounds in <i>S. typhimurium</i> TA100 + S9 microsomal preparation are expressed in log of revertant per nonamole, ln R. | n=36, r ² =0.6446, s=0.861, F=62 (training set) n=37, r ² =0.7843, s=0.616, F=142 (test set) | [53] |

Discussion

The model for normal boiling points [48] (the same substances as for QSAR 7, Table 3.2) is characterized by $n=134$, $r^2=0.9886$, $s=2.86$ °C, $F=3770$. Thus, the statistical characteristics for the QSAR 7 are better. QSAR analysis based on the quantum chemical descriptors for 57 anti-HIV-1 agents of Tetrahydro-imidazo[4,5,1-jk][1,4]-benzodiazepin-2 (1H)-one (TIBO) derivatives together with 1-[(2-hydroxyethoxy)methyl]-6-(phenylthio)thymine (HEPT) derivatives is represented. QSAR model for TIBO derivatives is characterized [49] by $r^2=0.8649$, $s=0.597$. In the above-mentioned work the statistical quality of the QSAR model for HEPT derivatives is the following: $r^2=0.9063$, $s=0.371$. The QSAR 1 is related to both TIBO and HEPT derivatives. The model is checked up with the external test set. Thus, the statistical quality of QSAR 1 and statistical quality of models which are described in Ref. 49 should be estimated as similar. Statistical characteristics of best model of the acute aquatic toxicity reported in Ref. 50 are the following: $n=69$; $r^2=0.863$; $s=0.30$. Statistical characteristics for QSAR 5 are better. Thus, these comparisons with models for the same endpoints from the literature indicate that the optimal OCWLGI-descriptors can be useful for the QSPR/QSAR analyses.

It is to be noted, there are a few topological indices, which can be translated into their OCWLGI-versions [50-61].

Finally, there are alternatives of the molecular graph for representation of the molecular structure. These are SMILES [62], SMART [63], and InChI [64]. In principle, these representations of the molecular structure also can be involved in the building up of the QSPR/QSAR [65-70].

Conclusions

Rigid and flexible topological indices are two different approaches of the QSPR/QSAR analyses. The OCWLGI descriptors (a sub-set of the flexible indices) calculated with molecular graph are a robust approach for predictive modelling of physicochemical and biomedicine parameters. Our experiences with OCWLGI descriptors have lead to formulation the following heuristics: (i) High individualization of molecular structure (e.g. with Morgan extended connectivity, valence shells, etc) can give overtraining; (ii) QSPR/QSAR based on the OCWLGI-descriptors should be

validated with several (as more as possible) splits into the training set and test set; and (iii) Rare invariants of molecular graph should be removed from the process of building up a model. There are possibility for search for novel versions of flexible topological indices which are based on the various representations of the molecular structure such as SMILES, InChI, and SMART.

Abbreviations

OCWLGI - optimization of correlation weights of local and global invariants of graph

QSPR/QSAR - quantitative structure-property/activity relationships

HFG - hydrogen-filled graph

HSG - hydrogen-suppressed graph

GAO - graph of atomic orbitals

References

1. Wiener, H. Structural Determination of Paraffin Boiling Points J. Am. Chem. Soc., 1947, 69(1), 17-20.
2. Wiener, H. Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, Among the Paraffin Hydrocarbons J. Am. Chem. Soc., 1947, 69(11), 2636-2638.
3. Wiener, H. Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure. Surface Tension, Specific Dispersion, and Critical Solution Temperature in Aniline J. Phys. Chem., 1948, 52(6), 1082-1089.

4. Wiener, H. Vapor Pressure-Temperature Relationships Among the Branched Paraffin Hydrocarbons *J. Phys. Chem.*, 1948, 52(2), 425-430.
5. Hosoya, H. Topological Index as a Sorting Device for Coding Chemical Structures.
6. Amidon, G.L.; Anik, S.T. Comparison of several molecular topological indexes with molecular surface area in aqueous solubility estimation. *J. Pharm. Sci.*, 1976, 65, 801-808.
7. Bonchev, D.; Balaban, A.T.; Mekenyan, O. Generalization of the Graph Center Concept, and Derived Topological Centric Indexes. *J. Chem. Inf. Comput. Sci.*, 1980, 20, 106-113.
8. King, R.B. In: Chemical applications of topology and graph theory. Proceedings of a symposium held at the University of Georgia, Athens, Georgia, USA, 1983
9. Randić, M. Graph Valence Shells as Molecular Descriptors. *J. Chem. Inf. Comput. Sci.*, 2001, 41(3), 627-630.
10. Randić, M.; Basak, S. C. On Use of the Variable Connectivity Index 1 f in QSAR: Toxicity of Aliphatic Ethers. *J. Chem. Inf. Comput. Sci.*, 2001, 41, 614-618.
11. Randić, M.; Plavsic, D.; Lers, N. Variable Connectivity Index for Cycle-Containing Structures. *J. Chem. Inf. Comput. Sci.*, 2001, 41(3), 657-662.
12. Roy, K.; Leonard, T.J. QSAR modelling of HIV-1 reverse transcriptase inhibitor 2-amino-6-arylsulfonylbenzotrioles and congeners using molecular connectivity and E-state parameters. *Bioorg. Med. Chem.*, 2004, 12, 745-754.
13. Jalbout, A.F.; Li, X. Anti-HIV-1 inhibitors of various molecules using principles of connectivity. *J. Mol. Struct. (Theochem)*, 2003, 663, 19-23.
14. Visco, Jr., D.P.; Poppale, R. S.; Rintoul, M.D.; Faulon, J.-L. Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *J. Mol. Graph. Mod.*, 2002, 20, 429-438.
15. Accelrys, http://www.esi.umontreal.ca/accelrys/life/cerius46/qsar/theory_descriptors.html (Accessed March 1, 2010)

16. Hu, Q.; Yi-Zeng Liang, Y.; Fang, K. The Matrix Expression, Topological Index and Atomic Attribute of Molecular Topological Structure. *J. Data Sci.*, 2003, 1, 361-389.
17. Amić, D.; Beslo, D.; Lucić, B.; Nikolić, S.; Trinajstić, N.; The Vertex-Connectivity Index Revisited. *J. Chem. Inf. Comput. Sci.*, 1998, 38, 819-822.
18. Randić, M.; Basak, S. C.; Optimal Molecular Descriptors Based on Weighted Path Numbers. *J. Chem. Inf. Comput. Sci.*, 1999, 39, 261-266.
19. Randić, M.; Pompe, M. The Variable Connectivity Index 1^f versus the Traditional Molecular Descriptors: A Comparative Study of 1^f Against Descriptors of CODESSA.
20. Benfenati, E.; Carbó-Dorca, R.; Gini, G.; Grauel, A. Multiple toxicity prediction using different molecular descriptors, Proceeding of the V Girona Seminar on Molecular Similarity, Girona, Spain, July 12-20, 2001.
21. Randić, M.; Basak, S. C.; New Descriptor for Structure-Property and Structure-Activity Correlations. *J. Chem. Inf. Comput. Sci.*, 2001, 41, 650-656.
22. Randić, M.; Pompe, M. The Variable Molecular Descriptors Based on Distance Related Matrices. *J. Chem. Inf. Comput. Sci.*, 2001, 41, 575-585.
23. Randić, M. Novel graph theoretical approach to heteroatoms in quantitative structure–activity relationships. *Chemometr. Intel. Lab.*, 1991, 10, 213-227.
24. Randić, M. On computation of optimal parameters for multivariate analysis of structure-property relationships. *J. Chem. Inf. Comput. Sci.*, 1991, 12, 970-980.
25. Amić, D.; Basak, S.C.; Lucić, B.; Nikolić, S.; Trinajstić, N. Structure – Water solubility modelling of aliphatic alcohols using the weighted path number. *SAR and QSAR in Environ. Res.*, 2002, 13, 281-295.
26. Randić, M.; Dobrowolski, J. Cz. Optimal molecular connectivity descriptors for nitrogen-containing molecules. *Int. J. Quant. Chem.*, 1998, 70, 1209-1215.
27. Randić, M.; Basak, S.C. Construction of high-quality Structure-Property-Activity Regerssions: The boiling points of Sulfides. *J. Chem. Inf. Comput. Sci.*, 2000, 40,899-905.

28. Toropov, A. A.; Toropova, A.P. Optimization of Correlation Weights of the Local Graph Invariants: Use of the Enthalpies of Formation of Complex Compounds for the QSPR Modelling. *Russ. J. Coord. Chem.*, 1998, 24, 81-85.
29. Toropov, A. A.; Toropova, A.P.; Voropaeva, N.L.; Ruban, I.N.; Rashidova, S.Sh.; Two Concepts of Weighing Molecular Graph Local Invariants in QSPR Modelling of the Enthalpies of Complexes: Sampling of Increments and Optimization of Correlation Weights. *Russ. J. Coord. Chem.* 1999, 25, 618-623.
30. Gutman, I.; Toropov, A.A.; Toropova, A.P. The graph of atomic orbitals and its basic properties. 1. Wiener index. *MATCH Commun. Math. Comput. Chem.*, 2005, 53, 215-224.
31. Toropov, A.A.; Toropova A.P. QSPR modelling of alkane properties based on graph of atomic orbitals. *J. Mol. Struct. (Theochem)*, 2003, 637, 1-10.
32. Krenkel, G.; Castro, E.A.; Toropov, A.A. Improved molecular descriptors to calculate boiling points based on the optimization of correlation weights of local graph invariants. *J. Mol. Struct. (Theochem)*, 2001, 542, 107-113.
33. Toropov, A.A.; Toropova, A.P. QSPR Modelling of the Formation Constants for Complexes Using Atomic Orbital Graphs. *Russ. J. Coord. Chem.* 2000, 26, 398-405.
34. Toropov, A. A.; Toropova A.P. Prediction of heteroaromatic amine mutagenicity by means of correlation weighting of atomic orbital graphs of local invariants. *J. Mol. Struct. (Theochem)*, 2001, 538, 287-293.
35. Toropov, A.A; Toropova A.P. Nearest neighboring code and hydrogen bond index in labeled hydrogen-filled graph and graph of atomic orbitals: application to model of normal boiling points of haloalkanes. *J.Mol. Struct. (Theochem)*, 2004, 711, 173-183.
36. Toropov, A.A.; Gutman, I.; Furtula B. Graph of atomic orbitals and the molecular structure descriptors based on it. *J. Serb. Chem. Soc.*, 2005, 70, 669-674.
37. Gutman, I.; Furtula, B.; Toropov, A.A.; Toropova A.P. The graph of atomic orbitals and its basic properties. 2. Zagreb index. *MATCH Commun. Math. Comput. Chem.*, 2005, 53, 225-230.

38. Toropov, A.A.; Toropova, A.P.; Gutman, I. Comparison of QSPR models based on hydrogen-filled graphs and on graphs of atomic orbitals. *Croat. Chem. Act.*, 2005, 78, 503-509.
39. Toropov, A.A.; Toropova, A.P. QSPR modelling of complex stability by optimization of correlation weights of the hydrogen bond index and the local graph invariants. *Russ. J. Coord. Chem.*, 2002, 28 (12), 877-880.
40. Toropov, A.A.; Toropova, A.P. Modelling of acyclic carbonyl compounds normal boiling points by correlation weighting of nearest neighboring codes. *J. Mol. Struct. (THEOCHEM)*, 2002, 581, 11-15.
41. González, M.P.; Toropov, A.A.; Duchowicz, P.R.; Castro, E.A. QSPR calculation of normal boiling points of organic molecules based on the use of correlation weighting of atomic orbitals with extended connectivity of zero- and first-order graphs of atomic orbitals. *Molecules*, 2004, 9 (12), 1019-1033.
42. Toropov, A.A.; Voropaeva, N.L.; Ruban, I.N.; Rashidova, S.Sh. Quantitative structure-property relationships for binary polymer-solvent systems: Correlation weighing of the local invariants of molecular graphs. *Polymer Science - Series A*, 1999, 41(9), 975-985.
43. Marino, D.J.G.; Castro, E.A.; Toropov, A. Improved QSAR modelling of anti-HIV-1 activities by means of the optimized correlation weights of local graph invariants. *Cent. Eur. J. Chem.*, 2006, 4(1), 135-148.
44. Toropov, A.A.; Duchowicz, P.; Castro, E.A. Structure-toxicity relationships for aliphatic compounds based on Correlation Weighting of Local Graph Invariants. *Inter. J. Mol. Sci.*, 2003, 4(5), 272-283.
45. Toropov, A.A.; Schultz, T.W. Prediction of aquatic toxicity: Use of optimization of correlation weights of local graph invariants. *J. Chem. Inf. Comput. Sci.*, 2003, 43(2), 560-567.
46. Toropov, A.A.; Toropova, A.P. QSAR modelling of toxicity on optimization of correlation weights of Morgan extended connectivity. *J. Mol. Struct. (THEOCHEM)*, 2002, 578, 129-134.
47. Toropov, A.A.; Benfenati, E. Correlation weighting of valence shells in QSAR analysis of toxicity. *Bioorg. Med. Chem.*, 2006, 14(11) 3923-3928.

48. Ivanciuc, O.; Ivanciuc, T.; Balaban, A.T.; The complementary distance matrix, a new molecular graph metric. *ACH–Model. Chem.* 2000, 137, 57–82.
49. González, O.G.; Murray, J.S.; Peralta-Inga, Z.; Politzer, P. Computed molecular surface electrostatic potentials of two groups of reverse transcriptase inhibitors: Relationships to anti-HIV-1 activities. *Int. J. Quantum Chem.* 2001, 83, 115–124.
50. Gute, B. D.; Basak, S.C. Predicting Acute Toxicity (LC50) of Benzene Derivatives Using Theoretical Molecular Descriptors: A Hierarchical QSAR Approach. *SAR QSAR Environ. Res.* 1997, 7, 117-131.
51. Marino, D.J.G.; Peruzzo, P.J.; Castro, E.A.; Toropov, A.A. QSAR carcinogenic study of methylated polycyclic aromatic hydrocarbons based on topological descriptors derived from distance matrices and correlation weights of local graph invariants. *Internet Electron. J. Mol. Des.*, 2002, 1, 115-133.
52. Peruzzo, P.J.; Marino, D.J.G.; Castro, E.A.; Toropov, A.A. QSPR modelling of lipophilicity by means of correlation weights of local graph invariants. *Internet Electron. J. Mol. Des.*, 2003, 2, 334-347.
53. Toropov, A.A.; Toropova, A.P. QSAR modelling of mutagenicity based on graphs of atomic orbitals. *Internet Electron. J. Mol. Des.*, 2002, 1, 108-114.
54. Bonchev, D. My life-long journey in mathematical chemistry. *Internet Electron. J. Mol. Des.*, 2005, 4, 434-490.
55. Costescu, A.; Diudea, M.V. QSTR study on aquatic toxicity against *Poecilia reticulata* and *Tetrahymena pyriformis* using topological indices. *Internet Electron. J. Mol. Des.*, 2006, 5, 116-134.
56. González, M.P.; Helguera, A.M.; Rodríguez, Y.M. TOPS-MODE and DRAGON descriptors in QSAR. 1. Skin permeation. *Internet Electron. J. Mol. Des.*, 2004, 3, 750-758.
57. Hosoya, H. The topological index Z before and after 1971. *Internet Electron. J. Mol. Des.*, 2002, 1, 428-442.
58. Kier, L.B. My journey through structure: The structure of my journey. *Internet Electron. J. Mol. Des.*, 2006, 5, 181-191.

59. Lukovits, I.; Miličević, A.; Nikolić, S.; Trinajstić, N. On walk counts and complexity of general graphs. *Internet Electron. J. Mol. Des.*, 2002, 1, 388-400.
60. Trinajstić, N. A life in science. *Internet Electron. J. Mol. Des.*, 2003, 2, 413-434.
61. García-Domenech, R.; Catalá-Gregori, A.; Calabuig, C.; Antón-Fos, G.; del Castillo, L.; Gálvez, J. Predicting antifungal activity: A computational screening using topological descriptors. *Internet Electron. J. Mol. Des.*, 2002, 1, 339-350.
62. <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>
63. http://www.daylight.com/dayhtml_tutorials/languages/smarts/index.html
64. <http://www.iupac.org/inchi/>
65. Toropov, A.A., Toropova, A.P., Mukhamedzhanova, D.V., Gutman, I. Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR) (2005) *Indian Journal of Chemistry - Section A Inorganic, Physical, Theoretical and Analytical Chemistry*, 44 (8), pp. 1545-1552.
66. Toropov, A.A., Rasulev, B.F., Leszczynski, J. QSAR modeling of acute toxicity by balance of correlations (2008) *Bioorganic and Medicinal Chemistry*, 16 (11), pp. 5999-6008.
67. Toropov, A.A., Benfenati, E. SMILES in QSPR/QSAR modeling: Results and perspectives (2007) *Current Drug Discovery Technologies*, 4 (2), pp. 77-116.
68. Toropov, A.A., Toropova, A.P., Benfenati, E., Leszczynska, D., Leszczynski, J. InChI-based optimal descriptors: QSAR analysis of fullereneC60.-based HIV-1 PR inhibitors by correlation balance (2010) *European Journal of Medicinal Chemistry*, 45 (4), pp. 1387-1394.
69. Toropov, A.A., Toropova, A.P., Benfenati, E., Leszczynska, D., Leszczynski, J. Additive InChI-based optimal descriptors: QSPR modeling of fullerene C 60 solubility in organic solvents (2009) *Journal of Mathematical Chemistry*, 46 (4), pp. 1232-1251.
70. Toropov, A.A., Toropova, A.P., Benfenati, E., Manganaro, A. Rapid communication QSPR modeling of enthalpies of formation for organometallic compounds by SMART-based optimal descriptors (2009) *Journal of Computational Chemistry*, 30 (15), pp. 2576-2582

Annex 4

A free and open source informatics library for chemistry: Chemistry Development Kit (CDK)

The Chemistry Development Kit (*CDK*) is a Java library for structural chemo- and bioinformatics, used in at least 10 different academic and industrial projects worldwide. As a successor of Christoph Steinbeck's CompChem libraries, the CDK library evolved into a full chemo-informatics package with code reaching from QSAR descriptor calculations to 2D and 3D model building. It is maintained as a SourceForge project under <http://www.sourceforge.net/projects/cdk>. SourceForge offers bug tracking, mailing lists, support manager and CVS access³.

The CDK library contains a huge number of various classes and it is impossible to describe all of them here. Instead, some of the basic classes needed for most of the calculations, including the calculations of molecular descriptors, will be explained in this section.

The class Atom represents the idea of a chemical atom. The following constructors are available for creating an Atom object:

- **public Atom()** – constructs a completely unset Atom.
- **public Atom(String elementSymbol)** - Constructs an Atom from a String containing an element symbol.
- **public Atom(String elementSymbol, javax.vecmath.Point3d point3d)** - Constructs an Atom from a String containing an element symbol and an additional Point3d object representing the 3D coordinates of the atom.
- **public Atom(String elementSymbol, javax.vecmath.Point2d point2d)** - Constructs an Atom from a String containing an element symbol and an additional Point2d object representing the 2D coordinates of the atom.

There are also better methods for partial charge, hydrogen count, stereo parity and location in a 2D and 3D space of the Atom object.

The class Bond implements the concept of a covalent bond between two atoms. A bond is considered to be the number of electrons connecting two atoms. The following constructors are available for creating a Bond object:

- **public Bond()** - Constructs an empty bond.
- **public Bond(IAtom atom1, IAtom atom2)** - Constructs a bond with a single bond order between the two atoms given as input parameters.
- **public Bond(IAtom atom1, IAtom atom2, double order)** - Constructs a bond with a given bond order between the two atoms given as input parameters.
- **public Bond(IAtom atom1, IAtom atom2, double order, int stereo)** - Constructs a bond with a given order and stereo orientation from an array of atoms.

Some of the more important methods of this class are the following:

- **public void setAtoms(IAtom[] atoms)** - Sets the array of atoms making up this bond.
- **public int getAtomCount()** - Returns the number of Atoms in this Bond.

- **public IAtom getAtom(int position)** - Returns the Atom from this bond at the position given as an input parameter.
- **public IAtom getConnectedAtom(IAtom atom)** - Returns the atom connected to the atom given as an input parameter.
- **public void setAtom(IAtom atom, int position)** - Sets an Atom in this bond at the position given as an input parameter.

There are also setter and getter methods for the order of the bond, the stereo descriptor, the geometric 2D center, and the geometric 3D center.

The class Molecule represents the concept of a chemical molecule, an object composed of atoms connected by bonds. The following constructors are available for creating a Molecule object:

- **public Molecule()** - Creates a Molecule object without Atoms and Bonds.
- **public Molecule(int atomCount, int bondCount, int lonePairCount, int singleElectronCount)** - Constructor a Molecule object where the parameters define the initial capacity of the arrays.
- **public Molecule(IAtomContainer container)** - Constructs a Molecule with a shallow copy of the atoms and bonds of an AtomContainer.

The class AtomContainer is a base class for all chemical objects that maintain a list of Atoms and ElectronContainers. The following constructors are available for creating an AtomContainer object:

- **public AtomContainer()** - Constructs an empty AtomContainer.
- **public AtomContainer(IAtomContainer container)** - Constructs an AtomContainer with a copy of the atoms and electronContainers of another AtomContainer.
- **public AtomContainer(int atomCount, int bondCount, int lpCount, int seCount)** - Constructs an empty AtomContainer that will contain a certain number of atoms, bonds, lone pairs and single electrons. It will set the starting array lengths to the defined values, but will not create any of these objects.

Some of the more important methods of this class are the following:

- **public void setAtoms(*IAtom[] atoms*)** - Sets the array of atoms of this AtomContainer.
- **public void setAtom(*int number, IAtom atom*)** - Set the atom at the position given as an input parameter.
- **public IAtom getAtom(*int number*)** - Gets the atom at the position given as an input parameter.
- **public java.util.Iterator atoms()** - Returns an Iterator for looping over all atoms in this container.
- **public int getAtomNumber(*IAtom atom*)** - Returns the position of a given atom in the atoms array. It returns -1 if the atom does not exist.
- **public int getAtomCount()** - Returns the number of Atoms in this Container.
- **public List getConnectedAtomsList(*IAtom atom*)** - Returns an ArrayList of all atoms connected to the given atom.
- **public int getConnectedAtomsCount(*IAtom atom*)** - Returns the number of atoms connected to the given atom.
- **public void addAtom(*IAtom atom*)** - Adds an atom to this container.
- **public void removeAtom(*int position*)** - Removes the atom at the given position from the AtomContainer.
- **public void removeAtom(*IAtom atom*)** - Removes the given atom from the AtomContainer.
- **public void removeAllElements()** - Removes all atoms and bond from this container.

Analogue methods exist for manipulating the Bond, LonePair and SingleElectron objects.

The class AtomContainerManipulator is a class with convenient methods for manipulating AtomContainer objects. Some of the more important methods of this class which are useful for implementing the descriptor classes involve hydrogen manipulation and are listed below:

- **public static int getTotalHydrogenCount(IAtomContainer atomContainer)** - Returns the summed implicit hydrogens of all atoms in this AtomContainer.
- **public static int countExplicitHydrogens(IAtomContainer atomContainer, IAtom atom)** - Returns the number of explicit hydrogens on the given IAtom.
- **public static int countHydrogens(IAtomContainer atomContainer, IAtom atom)** - Returns the summed implicit and explicit hydrogens of the given IAtom.
- **public static IAtomContainer removeHydrogens(IAtomContainer atomContainer)** - Produces an AtomContainer without explicit hydrogens but with hydrogen count from one with hydrogens. The new molecule without hydrogens is a deep copy.

The SmilesParser class parses a SMILES string and an AtomContainer. It does not parse stereochemical information, but the following features are supported: reaction smiles, partitioned structures, charged atoms, implicit hydrogen count and isotope information. It contains a number of methods but the method of our interest is the one for parsing a SMILES string:

- **public IMolecule parseSmiles(String smiles)** - Parses a SMILES string and returns a Molecule object representing the constitution given in the SMILES string.

In order to get a better understanding of the above mentioned classes, the following simple code segments illustrate their possible use. The first one creates an AtomContainer object and an Atom object from a String containing the element symbol 'C'. It sets the hydrogen count of this atom to 4 and it adds it to the AtomContainer. The second one creates an AtomContainer object by parsing the SMILES string "NC(CO)C(=O)O".

```
IAtomContainer methan= new AtomContainer();

Atom c=new Atom("C");

c.setHydrogenCount(4);

methan.addAtom(c);

SmilesParser parser=new SmilesParser();

IAtomContainer molecule=parser.parseSmiles("NC(CO)C(=O)O");
```

With the addition of the `cdk.qsar` module, CDK has been extended to allow for the calculation of molecular descriptors. Currently 33 descriptors are present covering topological, geometric and electronic descriptor classes. The rest of this section is dedicated to the use of CDK for calculating molecular descriptors.

In order to get a better insight into this matter, it is important to understand the structure of the descriptor classes and the way of implementing individual descriptors. All descriptor classes implement the `IMolecularDescriptor` interface and as such must implement all its methods, namely:

- **DescriptorSpecification** `getSpecification()` - returns an object containing the descriptor specification.
- **void** `setParameters(Object params[])` - sets the parameters attribute of the Descriptor object.
- **Object[]** `getParameters()` - gets the parameters attribute of the Descriptor object.
- **DescriptorValue** `calculate(IAtomContainer atomContainer)` - calculates the value of the descriptor for the atom/molecule given as an input parameter and returns a `DescriptorValue` object which contains this value.
- **IDescriptorResult** `getDescriptorResultType()` - returns an object that implements the `IDescriptorResult` interface indicating the actual type of values returned by the descriptor in the `DescriptorValue` object. Depending on the type of the descriptor, the `IDescriptorResult` can be of one of the following types: `BooleanResult`, `DoubleResult`, `DoubleArrayResult`, `IntegerResult` or `IntegerArrayResult`.
- **String[]** `getParameterNames()` - gets the `parameterNames` attribute of the Descriptor object.
- **Object** `getParameterType(String name)` - gets the `parameterType` attribute of the Descriptor object.

References

1. <http://apps.sourceforge.net/mediawiki/cdk>

Giuseppina Gini^a, Luigi Cardamone^a,
Magdalena Gocieva^a, Marina Mancusi^b,
Rima Padovani^b, Lorenzo Tamellini^a

a. Politecnico di Milano, Piazza L. da Vinci 32, 20133,
Milano, Italy

b. Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129,
Torino, Italy

Annex 5

REACH

REACH is a new European Community Regulation on chemicals and their safe use. It entered into force on June 1st 2007 and introduced an integrated system for Registration, Evaluation, Authorisation and Restriction of Chemical substances [1].

REACH replaces about 40 pieces of legislation with a streamlined and improved regulation, aiming at filling the gaps and solving some problems linked to the current system.

The aims of REACH are to:

- improve the protection of human health and the environment through the better and earlier identification of the intrinsic properties of chemical substances;
- maintain and enhance innovative capability and competitiveness of the EU chemicals industry (*the 10 kg threshold for registration discouraged research and invention on new substances and favoured the development and use of existing substances over new ones*);
- prevent fragmentation and ensure the free circulation of substances on the internal market of the European Union;

- promote alternative methods for the assessment of hazards of substances;
- facilitate data sharing in order to reduce tests on vertebrate animals and to reduce costs to industry. In fact, new tests are only required when it is not possible to provide information in any other permitted way and data gained by vertebrate animal testing are to be shared, in exchange for payment. Information not involving tests on vertebrate animals (*e.g. in vitro studies or QSARs*) must be shared on the request of a potential registrant.

The new law imposes the general obligation for manufacturers and importers of substances to submit a registration to the ECHA for each substance manufactured or imported to the European Countries in quantities of 1 tonne or above per year. ECHA is the European Chemicals Agency in Helsinki and it will manage and in some cases carry out the technical, scientific and administrative aspects of the REACH system at Community level, aiming to ensure that REACH functions well and has credibility with all stakeholders.

REACH covers all substances whether manufactured, imported, used as intermediates or placed on the market, either on their own, in preparations or in articles, unless they are radioactive, subject to customs supervision, or are non-isolated intermediates. Waste is specifically excepted. Food is not subject to REACH as it is not a substance, preparation or article. Member States may exempt substances used in the interest of defence. Other substances are exempted from parts of REACH, where other equivalent legislation applies.

A single regulatory system will be created that divides substances into two different categories: non-phase-in substances, i.e. those not produced or marketed prior to the entry into force of REACH, and phase-in substances that are those substances listed in the EINECS, or those that have been manufactured in the Community, but not placed on the Community market, in the last 15 years or the so-called “no longer polymers” of Directive 67/548.

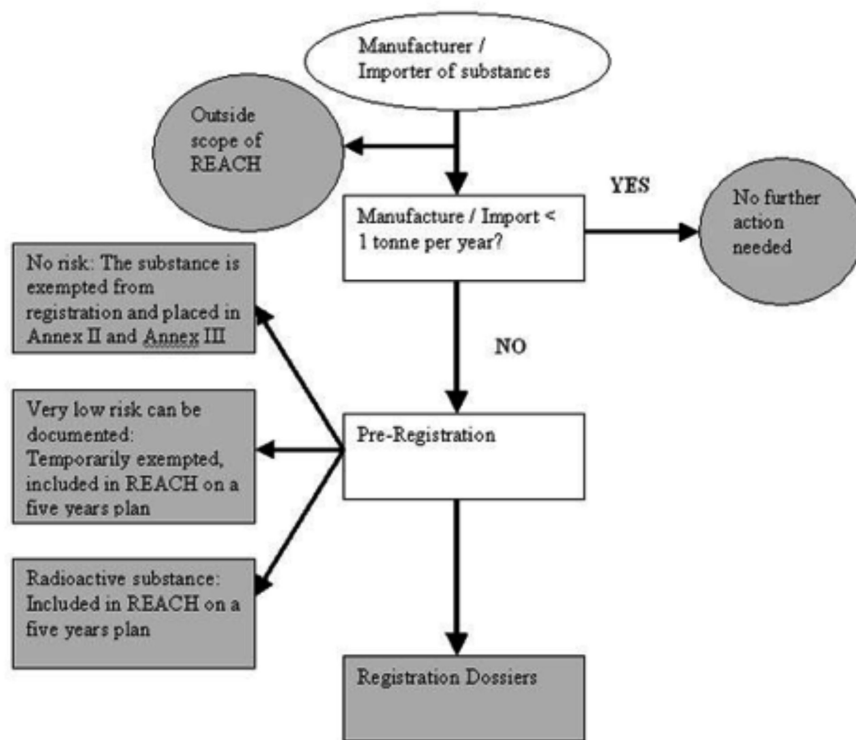
Novelties introduced by REACH

As the acronym REACH indicates, the four basic elements of the new regulation are: Registration, Evaluation, Authorisation and Restriction of Chemical substances.

Registration

As mentioned above, there is the general obligation for manufacturers and importers to submit a registration to the European Chemicals Agency for each substance manufactured or imported in quantities of 1 tonne or more per year. Failure to register means that the substance is not allowed to be manufactured or imported. Registrants have to submit a technical dossier, which contains some general information about both the substance (*i.e. identity, information about the manufacture and the uses, classification and labelling, etc.*) and the manufacturer or importer. In addition, registrants have to submit a chemical safety report (CSR) for the registration of substances that are produced or imported in quantities of 10 or more tonnes per year, where risks measures are defined. This report contains information on the different exposure scenarios linked to the different uses of the substance and it needs to point out the adequate measures for the risk assessment focused on the substance. The full registration process is summarized in *Figure 5.1*.

Figure 5.1. General scheme for the REACH registration of a chemical substance.



Evaluation

The Agency is responsible for performing the evaluation procedure. There are two types of evaluation with different aims: the dossier evaluation, on the one hand, and the substance evaluation, on the other hand. In the first case, the Agency checks the compliance of the registration dossier with the registration requirements and evaluate testing proposals made by industry, in order to prevent unnecessary animal testing, i.e. the repetition of existing tests and poor quality tests. In the second case, substances will be evaluated on the basis of considerations about risks, exposure and tonnage.

The Agency in co-ordination with the Competent Authorities of Member States may clarify suspicions of risks to human health or the environment by requesting further information from industry.

Evaluation may lead authorities to the conclusion that action needs to be taken under the restrictions or authorisation procedures in REACH, or that information needs to be passed on to other authorities responsible for relevant legislation. The evaluation process will ensure that reliable and useful data are provided and made available to the relevant bodies by the Agency.

Authorization

For substances of very high concern, an authorization is required for their use and their release on the market. This procedure aims at substituting the most dangerous substances and better managing risks arising from or linked to specific uses.

The substances required to be authorised are CMR substances (Carcinogenic, Mutagenic or toxic to Reproduction), PBT substances (*Persistent, Bioaccumulative and Toxic*), vPvBs (*very Persistent, very Bioaccumulative substance*), and substances identified by scientific evidence as causing probable serious and normally irreversible effects to humans or the environment.

The authorization application is to be submitted to the Agency, by manufacturers, importers and/or downstream users of a specific substance.

The Commission is responsible for the granting and the rejection of the authorization. Authorization is granted if the risks for human health and the environment coming from the use of a specific

substance are adequately controlled. If the risks cannot be controlled, the authorization would be granted if the socio-economic benefits of their use outweigh the risks for human health and the environment and if there are no safer suitable alternative substances or technologies. If there are, the applicants must prepare substitution plans; if not, they should provide information on research and development activities, if appropriate. The Commission may amend or withdraw any authorisation on review if suitable substitutes become available.

Restrictions

The restriction provisions act as the safety net for the system because they are applied to any substance on its own, in a preparation or in an article where there is an acceptable risk to health or the environment. This procedure regulates Community conditions for the manufacture, placing on the market or use of such substances and eventually forbids any of these activities if necessary.

Proposals for restrictions will be prepared by Member States or by the Agency on behalf of the Commission in the form of a structured dossier.

Cost/ benefit analysis of REACH suggests that the introduction of the new regulation will have some relevant benefits, such as:

- positive public health impact: a deeper knowledge about chemicals, hazards and more controls will help better implementation on existing legislation. According to World Bank estimates, diseases caused by chemicals were assumed to account for some 1% of the overall burden of all types of disease in the EU. Assuming a 10% reduction in these diseases as a result of REACH would result in a 0.1% reduction in the overall burden of disease in the EU. This would be equivalent to around 4,500 deaths due to cancer being avoided every year;
- positive environment impact: thanks to REACH, current chemical releases to the environment and exposure of humans via the environment can be reduced. A recent study commissioned by DG Environment illustrated that the long-term benefits of REACH would be significant, because its introduction will contribute to reduce pollution of air, water and soil as well as to reduce pressure on biodiversity.

However, the new regulation will also introduce additional costs, as explained in the Extended Impact Assessment of the Commission's proposal. The direct costs of REACH to the chemicals industry were estimated at a total of € 2.3 billion over the first 11 years after the entry into force of the Regulation.

Assuming that the market behaves as expected with only 1-2% of substances withdrawn because their continued production would not be profitable, the additional costs to downstream users of chemicals were estimated at €0.5 - 1.3 billion in a "normal expectation" case and €1.7 - 2.9 billion in a scenario with higher substitution costs assumed.

Combining the estimates of the direct and indirect costs, the overall costs were estimated to fall in the range of €2.8 - 5.2 billion. These costs will be incurred over a period of 11 to 15 years. Therefore, from a macroeconomic perspective, the overall impact in terms of the reduction in the EU's Gross Domestic Product (*GDP*) is expected to be limited.

Finally, a further work on the REACH Impact Assessment together with industry and monitored by all stakeholders was conducted by the Commission and some relevant conclusions have been drawn:

- There is limited evidence that higher volume substances are vulnerable to withdrawal following the REACH registration requirements. However, lower volume substances under 100 tonnes are most vulnerable to being made less profitable or unprofitable by the REACH requirements.
- There is limited evidence that downstream users will be faced with a withdrawal of substances of greatest technical importance to them.
- SMEs can be particularly affected by REACH having regard to their more limited financial capacity and lower market power in terms of passing on costs.
- Companies have recognised some business benefits from REACH [9].

Overview of interesting websites on REACH

Many different websites deal with the main aspects of REACH legislation. They can be divided into two main categories: those completely dedicated to the new regulation system and those which deal with REACH only within one or few sections.

Websites totally dedicated to REACH

Some websites that are totally dedicated to REACH are briefly described below. Among these websites some provide general information while others are posted by consulting agencies.

General information websites

- <http://eur-lex.europa.eu/JOHtml.do?uri=OJ:L:2007:136:SOM:EN:HTML>

The official Reach Regulation published on the Official Journal of the European Union.

- <http://echa.europa.eu/>

The ECHA website provides access to technical guidance, frequently asked questions (*FAQs*), software tools and helpdesks on REACH. Here the latest updates on guidance, tools, data on chemicals and the Regulation can be found.

- http://echa.europa.eu/reach/helpdesk/nationalhelp_contact_en.asp

Every European Country has its own helpdesk. The link mentioned refers to web page in ECHA website, where the list of national REACH helpdesks is provided with their specific links.

- <http://reach.mi.camcom.it>

Summaries about REACH in general are available: people involved (“*chi*”), field of action (“*dove*”), actions (“*come*”), time scheduling (“*quando*”), aims and basic principles (“*perchè*”). The website is in Italian.

Consulting services’ websites

- <http://www.reach-cdrom.eu/>

Detailed information and several documents about REACH in general are available on the website.

- <http://www.denehurst.co.uk/index.html>

Detailed information and several documents about REACH in general are available on the website. Specialised consultancy services are provided.

- <http://reach.itertech.it/index.php>

Detailed information and several documents about REACH in general are available on the website. Some consultancy services are provided too. The website is in Italian.

- <http://www.regolamentoreach.it>

Detailed information and several documents about REACH in general are available on the website. Some consultancy services are provided too. The website is in Italian.

- <http://www.reachcolours.it>

A lot of different solutions and services are provided in order to help firms to pre-register and register the different substances (*CAS, EINECS, ELINCS research, classification and labelling of substances, management of a SIEF etc.*).

Websites partly dedicated to REACH

Some interesting websites have only one or some sections related to the new regulation for chemical substances. A brief list is provided below.

- <http://ecb.jrc.ec.europa.eu/> - <http://ecb.jrc.ec.europa.eu/reach/> - <http://ecb.jrc.ec.europa.eu/qsar/>

This website belongs to The Consumer Products Safety & Quality (CPS&Q) Unit, formerly known as European Chemicals Bureau (ECB). Here several documents and guidance about REACH can be found. In addition, there is a specific section which is dedicated to QSAR and where some software tools are freely accessible, such as JRC QSAR Model Database, Toxtree, DART, etc. General information and a pdf file that describes the current version of the QSAR Prediction Reporting Format (QPRF) are also provided. http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm The European Commission's Environment Directorate-General (DG) has developed within an interim strategy a number of REACH Implementation Projects (RIPs) that foresee the development of guidance documents and IT-tools for the European Chemicals Agency, for industry and the authorities of the Member States including 5 central areas:

RIP 1 - REACH Process description

RIP 2- REACH-IT: Development of the IT system to support the REACH implementation

RIP 3 - Guidance Documents: Development of guidance documents for industry

RIP 4 - Guidance Documents: Development of guidance documents for authorities

RIP 5/6 - Setting up the Agency.

- http://ec.europa.eu/enterprise/reach/index_en.htm

General information and several documents about REACH and GHS (*Globally Harmonised System of Classification and Labelling of Chemicals*) are available on the website of European Commission, Directorate General for Enterprise and Industry.

- <http://www.hse.gov.uk/reach/index.htm>

General information are available on the website of UK Health and Safety Executive (*brief overview of REACH, pre-registration, case studies, etc.*).

- <http://www.env-health.org/a/3022>

General brief information about REACH are available on the website belonging to Health Environment Alliance (*brief overview of REACH, pre-registration, FAQ about REACH, etc.*).

- <http://www.rohs-international.com>

A suite of simplified guidance notes for REACH is available on the website and a lot of services are provided in order to help companies outside the EU to comply with REACH. This website belongs to by RoHS International.

- <http://www.iom-world.org/consulting/reach.php>

IOM (*Institute of Occupational Medicine*) Consulting offers several services and information to firms seeking to respond to these complex regulations. A few examples of the activities offered are providing guidance on how to ensure compliance that is specifically tailored for SMEs as well as providing services to major suppliers or users of chemicals, helping to identify and characterise relevant exposure scenarios, undertaking exposure modelling and/or measurement to inform the risk assessment process, advising on data gaps and helping firms fill them.

References

1. http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm

Orazio Nicolotti, Andrea Gissi,
Antonellina Introcaso, Angelo Carotti

*Dipartimento di Farmacia, Università degli Studi di Bari
"Aldo Moro", Via Edoardo Orabona 4, 70125, Bari*

Annex 6

Genetic Algorithms in QSAR for REACH

Introduction

Genetic algorithms (GAs) are a powerful optimization technique based on the Darwinian theory of the survival of the fittest. The application of GAs is strongly encouraged in all those complex situations requiring the unfeasible enumeration of a combinatorial number of hypothetical solutions. For instance, this is a typical problem encountered in the selection of the optimal pool of physicochemical properties when modelling a given REACH endpoint. This chapter briefly illustrates the strength and potential of GAs with an open eye to their use of QSAR in REACH; it is of course far from being a detailed and exhaustive description of the application of genetic algorithms to chemical problems.

What is a GA?

Since their development by Holland in 1975 [1], GAs have attracted great interest. The GA goal was that of miming the natural evolution of living species based on the Darwinian evolution principles from which descend unusual computer jargon terms such as ‘population’, ‘gene’, ‘chromosome’, ‘mutation’ and ‘crossover’ [2]. In other words, GAs are an optimization technique that takes inspiration from the evolution of the living species which ultimately can be considered by itself a sort of optimization, in which the response to be optimized is the adaptation to the environment.

Initially and for almost 20 years, limited computer availability restricted the real application of GAs; then, the advent of more powerful small computers in the early nineties overcame this major drawback and thus fostered application of GAs to both easy and moderate as well as complex scale problems.

The widespread use of GA was boosted by two milestone publications in Science [3] and Nature [4] on 1993 and 1995, respectively. The former showed a general view of GAs by illustrating how the algorithm works and its best use through some real-life applications; the latter described a problem of molecular dynamics successfully solved by a GA where standard search approaches had failed.

What is a GA? At a very general level, a GA can be thought of as an attempt to transfer to computer programs the principles of evolution theory [5,6,7]. As for living species, the individuals (*i.e.*, QSAR models) with a greater adaptation to the environment (*i.e.*, top predictive QSAR models) have a greater chance of surviving and, thus, of earning the right for mating (*i.e.*, generation of newer and better QSAR models).

It goes without saying that the improvement of the individuals (*i.e.*, QSAR models) takes place over generations because the selection of the genetic material (*i.e.*, selection of molecular descriptors) progressively improves. Briefly, the success of the evolution (*i.e.*, GA run) benefits from the death of worse individuals in favour of the survival of the best individuals so that they can spread their genetic material to the new generations.

Based on these concepts, several scheme and diverse settings can be developed to implement Gas [8]. However, all of them share three fundamental steps:

- 1. Random creation of the initial population:** Initially, the genetic material is randomly assigned to each individual with single bit set to 0 or 1 (*i.e., off/on molecular descriptors in QSAR model*). The population size keeps constant over the entire GA run with the number of individuals normally in the range from 20 to 100.
- 2. Fitness-based reproduction:** After the first generation, the individuals enter the reproductive phase in which they start mating to give birth to healthy offspring. In doing this, the chance of the fittest chromosomes (*i.e., best QSAR models*) of generating offspring is higher than that of the worst chromosomes. Thus, the offspring results by the recombination of the best parents. As in nature, the smart selection of survival models represents the intimate secret of the large success of GAs application. It should be considered that if such selection were simply random, then each chromosome would have the same chance of entering into future generations and the average fitness (*i.e., the model quality*) would be constant over generations. To avoid this pitfall, the chance of individuals of being selected was dependent on their fitness. In this respect, several selection methods exist. For instance, best individuals can be stochastically selected by drawing them from a roulette wheel having different sized slots on the basis of fitness, as shown in *Figure 6.1*.

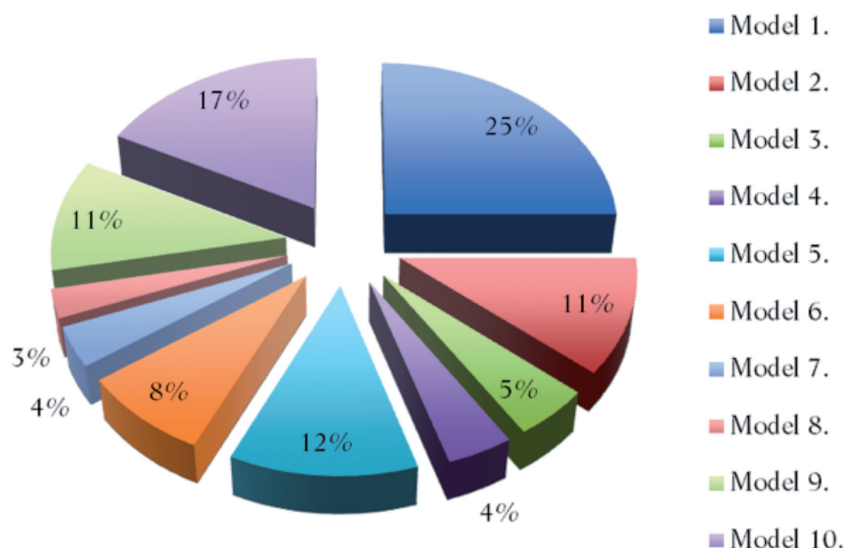


Figure 6.1: Roulette wheel approach based on fitness. Fitter and weaker models (i.e., individuals) have larger and smaller slots of the wheel.

Among others, the simplest way to assess the chance of survival of the individual p_i having a given fitness $resp_i$ is determined as follows:

In the competition of the survival, the best individuals will be more likely to propagate to a newer generation with a progressive increase of the average fitness through two basic mechanisms of recombination. One is the mutation that causes a random change of separate elements within a chromosome. A mutation is generally considered to be a background operator as it ensures that the probability of searching a particular subspace is low and never zero. Mutations can generally result in pathological conditions; however, such an irregular change can also determine good results contributing to the evolution of the population through searches in new zones of the parametric space, thus avoiding deadlock situations, typically local minima. The other and more frequently occurring genetic operator is crossover, in which a portion of genetic material diverse is taken from each parent and recombined to create new child chromosomes. Note that different recombination schemes can be adopted; one widely used in feature selection is the multipoint crossover that is intended to encourage the exploration of the search space rather than favour convergence to fit individuals early in the search.

$$p_i = \frac{resp_i}{\sum_{i=1}^N resp_i}$$

Figure 6.2: Example of crossover: on/off bits indicate the presence/absence of molecular descriptors.

- 3. Termination:** New generations will evolve until a given termination criterion is achieved. Normally, the three main criteria are: the completion of a prior defined number of generations; the evolution over a reasonable simulation time; the determination of a given threshold of the fitness value.

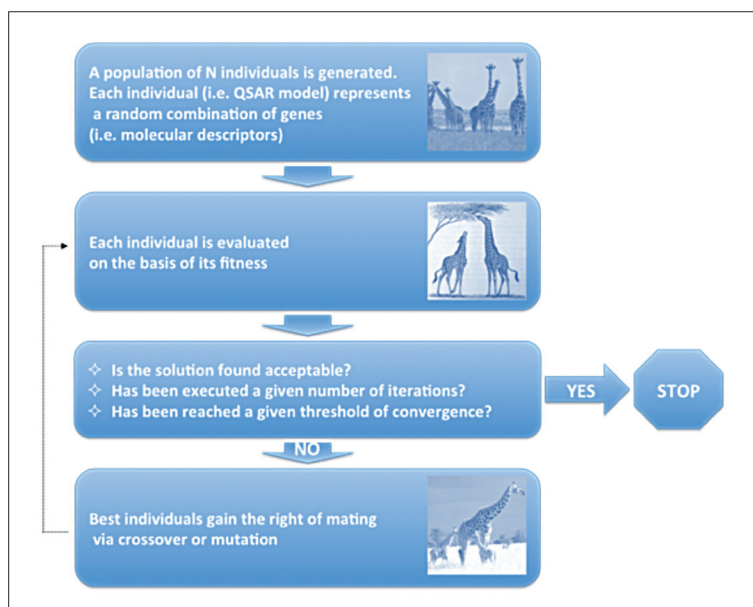


Figure 6.3. Zoomed in view of a GA. The population is randomly initialized. Each individual is evaluated on the basis of fitness. Finally, individuals are selected and promoted to spread genes via crossover and mutation.

Why use GAs in QSAR?

To better understand the potential of GAs, it is worth focusing on their application to QSAR whose ultimate goal is that of relating a given response (Y) (*i.e.*, one of the endpoints relevant to REACH) of a series of chemicals with some appropriate descriptors (X). More specifically, the aim of a QSAR study in REACH is the extraction of the most significant signals of the Y variance from the blend of information and noise contained in the parametric space of the physicochemical descriptors. Despite years of continuous efforts, the development of newer methods in QSAR is still the object of a never ending research. Actually, the number and nature of molecular descriptors represent key aspects in QSAR studies. In fact, it has to be considered that the molecular descriptors represent the knowledge platform on which QSAR models are placed. As a result, one might expect that the higher the number of molecular descriptors, the better. Unfortunately, this would be to reduce QSAR to a simple matter of data modelling while instead even greater attention has to be paid to its validation, coherence and applicability [9]. While validation can be assessed through conventional or *ad hoc* procedures, applicability is still sometimes questionable, since there could not be univocal definition and thus it could pertain to the experience and feeling of the REACH expert. For decision making, valuable clues come from read-across techniques. In fact, the relationships between molecular descriptors of similar chemicals biasing the same endpoint can help the smart user to predict the uncertain value of a target chemical whose endpoint value is missing.

In this report, we will discuss the potential of GAs in QSAR paying attention to REACH purposes. Since Hansch's first approach [10], it is today possible to have in hand several thousand molecular descriptors in a few seconds [11]. As a result, it is easy to generate huge data matrices in which rows represent the chemicals and columns the descriptors. However, discovering the right direction in an overwhelmed physicochemical space is far from being a trivial task and, thus, there is the need of selecting an optimal subset containing the most meaningful descriptors. In this view, it should be kept in mind that a combinatorial number of 2^{N-1} models can theoretically result when N descriptors are available, irrespective of the degrees of freedom.

For instance, the small sized Selwood collection [12] made by 31 chemicals with 53 properties involves a huge number of combinations of descriptors equal to 9×10^{15} QSAR models! Supposing the obtaining of one QSAR solution per second, the complete enumeration [13] would take as long as 285420921 years! Indeed, the need for parsimony (*i.e. Occam's Razor*) would recommend some rule of thumb (*i.e. a ratio equal to 5:1 considering the number of descriptors per chemical*) to constrain the real number of QSAR models and thus avoid a full systematic search. However, it should be taken into account that 23426 different QSAR analyses need to be carried even if the search is limited to only three-term models [*i.e.*, $\{N(N-1)(N-2)\}/\{3!\}$ where N represents the number of molecular descriptors]. It goes without saying that the variable selection is even more difficult when the parametric space is not restricted to linear terms only. In this respect, the interested reader can have a look at the general solvation equation of Abraham [14], which contains linear, quadratic and cross-product terms.

In this scenario, GAs have represented one of the right answers and deserve a special mention of merit as they avoid the mission impossible of enumerating all the possible QSAR models existing in a large parametric space but still allow quick retrieval of the optimal subset of variables to derive predictive models.

Stay fit stay well: learn by doing!

At this point, it should be clear that the successful use of QSAR in REACH requires the smart selection of the optimal pool of variables [15] for a given endpoint, a task that can be carried out by using GAs.

With this in mind, let us focus on how a GAs can be constructed for performing variable selection. First, individuals (*i.e.*, *QSAR models*) are represented as a 1-dimensional binary string of bits, each one indicating the presence/absence of a given variable. The initial population is created by randomly assigning bits. Each individual is, thus, assessed on the basis of a fitness function, which serves to rank them according to the chance of entering future generations.

In this scheme, better individuals (*i.e.*, *best QSAR models*) will more likely be chosen to propagate their genetic material (*i.e.*, *molecular descriptors*) to offspring via mutation and crossover. Establishing a given number of mating steps, the average fitness of the individuals in a population steadily increases until good combinations of genes are discovered and propagated to future generations. In this respect, GAs demonstrated to be the right approach for searching highly structured spaces. Importantly, another valuable benefit in using GAs is that they can be flexibly customized by using unconventional schemes so that several authors on the basis of their own feeling have published papers about feature selection GA-based, each of them using different settings. For instance, genetics storms can be enforced to maintain an acceptable degree of diversity within the population; in other words, when the average fitness of a given evolving population is nearly constant, the replicated and worst performing chromosomes are destroyed by sudden death, favouring the birth of new individuals. In addition, the GA can be structured as a *migration* or *island* model, which is based on multiple independent evolving populations ensuring better results compared to the single-population approach. However, implementing genetic competition requires the definition of a fitness function establishing the degree of adaptation to the environment and thus the chance of surviving of each chromosome within their population. In the case of a QSAR model for REACH, the fitness function to be optimized could be, for instance, the cross-validated variance explained by the selected set of variables (*in case of multivariate calibration*) or the percentage of cross-validated correct classifications (*in case of a classification problem*). However, variable selection is a difficult task which needs careful validation to avoid the pitfall of overfitting, basically due to the occurrence of chance correlations, that can often lead to overoptimistic results (*i.e.*, *high apparent predictive power*) [16]. Concerning regression problems, the fitness function should take into consideration two basic aspects. The first is related to the need of selecting few variables to ensure a quick and easy interpretation of the model output; this would be very helpful in reinforcing read-across assessment. The second deals with the model accuracy and, more importantly, its predictivity out of the training set; this would be very much relevant and represents the ultimate aim in the use of non-testing methods for REACH.

In this view, the reader will be then accompanied through some of the most important meaningful examples of fitness function applied to assess the goodness of a regression. Note that most of these

functions have been challenged using the Selwood dataset that, to date, constitutes a gold standard for QSAR practitioners for validating new methods and approaches.

Based on the law of parsimony (*i.e. Occam's Razor*), the Akaike Information Criterion penalty function [17] was used to identify an appropriate model structure when choosing between models by looking for an appropriate balance between the residual variance and the number of descriptors of a model. Rogers and Hopfinger analysed the Selwood dataset by using the genetic function approximation method [18], in which feature selection is performed using a GA and QSAR models are obtained via regression. Models are scored using the Friedman's lack of fit (*LOF*) function, which is based on the least-squares errors combined with a user-definable smoothing parameter that penalizes the inclusion of a high number of terms in a model. The mutation and selection uncover models (*MUSEUM*) algorithm proposed by Kubinyi [19] is based on an evolutionary algorithm that uses only the mutation operator to generate offspring. The FIT value is used as the fitness criterion that, to some extent, is an improvement on the Fischer significance value in that FIT is better calibrated toward the change in number of independent variables selected in each model. A related method, known as the evolutionary programming method, developed by Luke [20] has also been applied to the Selwood data set with a fitness function based on three terms. The first accounts for the root mean square between predicted and measured values. The second biases the solution toward a given number of descriptors and the third penalize the occurrence of non-linear models. Cho and Hermsmeier developed a novel GA-guided selection method (*GAS*) [21] to simultaneously optimize a set of encoded variables that include both descriptors and compounds and to construct QSAR models that are comparable to those obtained with other methods. More recently, a new approach called multi-objective QSAR (*MoQSAR*) was elaborated upon on the basis of genetic programming. This represents an extension of GA, to create a population of potential model solutions and on the basis of a multi-objective fitness function to handle multiple objectives independently without summation or weights [22]. By adopting Pareto ranking, *MoQSAR* [23] allowed the discovery of families of trade-off QSAR models, rather than a single solution, in only one run. According to this approach, the REACH expert was, therefore, given the option of choosing the most appropriate model from a pool of equivalent QSAR solutions. A more recent paper reports a novel method based on a new implemented GA for QSAR variable selection that explicitly consider even nonlinear- and cross-descriptors; the CVFIT [24] function allowed the occurrence of terms, other than those linear, only if the resulting QSAR models showed significantly higher internal predictive power as measured by the squared correlation coefficient of predictions (q^2).

Conclusions

The increasing number of papers describing the application of GAs shows their effectiveness and strength. It is now clear that the advantage of GAs over the classical techniques becomes greater with the complexity of the problem. Basically, the potential of the GAs is in the joint use of two search strategies that are exploration and exploitation. The former is typical of the random searches aiming at sampling different regions of the parametric space irrespective of what happens around them. The latter is typical of the local searches (*e.g.*, *simplex approach*) that try to reach the local maximum closer to the starting point irrespective of what happens in different regions of the parametric space. In this view, the secret of the success of GAs is in getting a good balance between exploration and exploitation. To this end, it has been observed that the results obtained by a GA can be significantly improved after hybridization with a standard technique that typically has a very poor exploration power but a very high exploitation potential. For instance, a number of hybrid algorithms that combine GAs and neural networks appear in the literature as promising methods for strengthening the impact of computational methods in drug design. A striking example is given by So and Karplus [25] who proposed a hybrid method that combines a GA for selecting descriptors and an artificial neural network for deriving models. Concerned with REACH, a relevant example is the development of the model CAESAR [26], which combines a GA with a heuristic approach for predicting the logBCF of chemicals.

Acknowledgements

The authors wish to thank Mr Domenico Gadaleta for helpful discussion.

References

1. Holland J.H., Genetic Algorithms, in: Leigh Tesfatsion homepage, Department of Economics, Iowa State University, 2007, <http://www.econ.iastate.edu/tesfatsi/holland.GAIntro.htm>
2. Goldberg D.E., Genetic Algorithms. In: Search, Optimization and Machine Learning, Addison-Wesley, Berkeley, 1989
3. Forrest S., Genetic algorithms: principles of natural selection applied to computation, Science. 1993 Aug 13;261(5123):872-8.
4. Maddox J., Genetics helping molecular dynamics, Nature 1995; 376:209.
5. Leardi R., Genetic algorithms in chemistry, J Chromatogr A. 2007 Jul 27;1158(1-2):226-33. Epub 2007 Apr 19.
6. Leardi R., Genetic algorithms in chemometrics and chemistry: a review, J Chemom, 15(7), 559-569, 2001
7. Leardi R., Boggia R. and Terrile M., Genetic algorithms as a strategy for feature selection, J Chemom, 6(5), 267-281, 1992
8. Niculescu S.P., Artificial neural networks and genetic algorithms in QSAR, J Mol Struct, 622:71-83, 2003
9. Tropsha A.E., Gramatica P., Gombar V.K., The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models, QSAR Comb. Sci. 2003; 22, 69-77.
10. Hansch C., Maloney P.P., Fujita T., Muir R.M., Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients, Nature 194, 178 - 180 (14 April 1962).
11. Todeschini R., Consonni V., Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany, 2000.

12. Selwood D.L., Livingstone D.J., Comley J.C., O'Dowd A.B., Hudson A.T., Jackson P., Jandu K.S., Rose V.S. and Stables J.N., Structure-activity relationships of antifilarial antimycin analogues: a multivariate pattern recognition study, *J Med Chem.* 1990 Jan;33(1):136-42.
13. McFarland J.W. and Gans D.J., On identifying likely determinants of biological activity in high dimensional QSAR problems, *Quant Struct-Act Relat* 1994, 13, 11-17.
14. Abraham M.H. and Le J., The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship, *J Pharm Sci.* 1999 Sep;88(9):868-80.
15. Waller C.L., Bradley M.P., Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional Quantitative Structure-Activity Relationship Studies, *J. Chem. Inf. Comput. Sci.* 1999, 39, 345- 355
16. Doweyko A.M., 3D-QSAR illusions, *J Comput Aided Mol Des.* 2004 Jul-Sep;18(7-9):587-96.
17. Rodriguez-Vazquez K. and Fleming P.J., Multi-objective genetic programming for nonlinear system identification, *Electron. Lett.* 1998, 34, 930-931
18. Rogers D. and Hopfinger A.J., Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships, *J. Chem. Inf. Comput. Sci.*, 1994, 34 (4), pp 854-866
19. Kubinyi H., Variable selection in QSAR studies. I. An Evolutionary Algorithm, *Quant. Struct.-Act. Relat.* 1994; 13, 285-294.
20. Luke B.T., Comparison of different data set screening methods for use in QSAR/QSPR generation studies, *J. Mol. Struct: Theochem* 2000, 507, 229-238.
21. Cho S.J. and Hermsmeier M.A., Genetic Algorithm guided Selection: variable selection and subset selection, *J Chem Inf Comput Sci.* 2002 Jul-Aug;42(4):927-36.
22. Nicolotti O., Giangreco I., Introcaso A., Leonetti F., Stefanachi A. and Carotti A., Strategies of multi-objective optimization in drug discovery and development, *Expert Opin Drug Discov.* 2011 Sep;6(9):871-84.

23. Nicolotti O., Gillet V.J., Fleming P.J. and Green D.V., Multiobjective optimization in quantitative structure-activity relationships: deriving accurate and interpretable QSARs, *J Med Chem.* 2002 Nov 7;45(23):5069-80.
24. Nicolotti O. and Carotti A., QSAR and QSPR studies of a highly structured physicochemical domain, *J Chem Inf Model.* 2006 Jan-Feb;46(1):264-76.
25. So S.S. and Karplus M., Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks, *J Med Chem.* 1996 Mar 29;39(7):1521-30.
26. Zhao C., Boriani E., Chana A., Roncaglioni A. and Benfenati E., A new hybrid system of QSAR models for predicting bioconcentration factors (BCF), *Chemosphere.* 2008 Dec;73(11):1701-7. Epub 2008 Oct 26.

Andrey A. Toropov^a, Alla P. Toropova^a,
Emilio Benfenati^a, Giuseppina Gini^b

*a. Istituto di Ricerche Farmacologiche Mario Negri,
Via La Masa 19, 20156, Milano, Italy*

*b. Department of Electronics and Information,
Politecnico di Milano, Via Ponzio 34/5,
20135 Milano, Italy*

Annex 7

The CORAL software: principles, results, perspectives

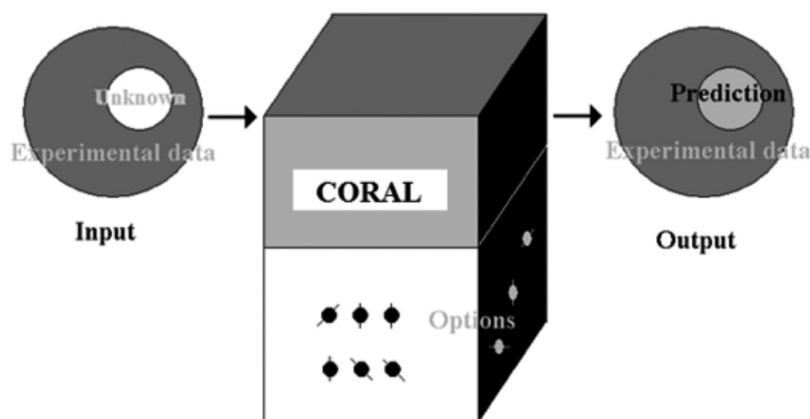
Introduction

The CORAL software has been used as a tool for building up the quantitative - structure property / activity relationships (QSPR/QSAR) of a number of various endpoints: inhibitors of human serine proteinases [1], anticonvulsant agents [2], anti-malaria activity [3], anticancer agents [4,6], toxicity toward *Daphnia magna* [7], mutagenicity [8], toxicity toward rat [9-11], bioconcentration factor [12], carcinogenicity [13], solubility [14] and anti-HIV-1 activity of fullerenes [15]. Our aim is to briefly inform about the CORAL software, in “the first approximation”. The above mentioned web site contains the reference manual where the CORAL software is described in detail.

Input for CORAL is molecular structures (*represented by SMILES*) together with available experimentally defined numerical values of an endpoint for part (not for all) of these substances.

Having this information one can insert this information by means of the above-mentioned operations and obtain predicted values. The quality of the prediction is depended on the selected options. *Figure 7.1* shows the general scheme of the building up of a CORAL model.

Figure 7.1: General scheme of the building up a model with the CORAL software



Thus CORAL provides a prediction that is probabilistic. Moreover, the prediction is defined by the method that is constructed by user.

However, one can estimate the quality of this prediction (*i.e., reliability of the selected method*) through three probabilistic criteria. The model is satisfactory if (*and only if*) there are:

- Good quality of the model for the external test set;
- Reproduction of the quality for the external test set in several runs of the Monte Carlo method optimization;
- Reproduction of the quality for external test set for several various splits into training and test sub-sets.

Principles

The CORAL (**COR**relation **A**nd **L**ogic) software has been developed with the following principles:

1. Molecular structure of the majority of substances can be represented by SMILES.
2. Often the SMILES can be translated into molecular graph [7,8].
3. SMILES is the source of molecular attributes representing *local* and *global* molecular features.
4. The building up of QSPR/QSAR model for an arbitrary split into the training and test sets should be qualified as a probabilistic event.
5. The statistical quality of each QSPR/QSAR model is a mathematical function of the split into the training and test sets.
6. The average statistical quality of QSPR/QSAR models that is obtained for several splits into training and test sets is a more robust criterion for the estimation of an approach than the statistical quality for only one split.
7. The average statistical quality of a model *for external test sets* is more significant data than the average statistical quality for training sets [10,13].
8. The correlation weights for molecular features (*which are extracted from graph and/or SMILES*) can be used for classification of the above-mentioned features according to their values for several models into three categories: features with stable positive values of correlation weights (*promoters of increase for the endpoint value*); features with stable negative values of correlation weights (*promoters of decrease of the endpoint value*); and undefined features which have positive values of correlation weights together with negative correlation weights values for series of runs of the Monte Carlo optimization.
9. Data on the correlation weights for molecular features calculated with graph and/or SMILES (*which are promoters of increase of an endpoint and promoters of its decrease*) give the possibility to define the applicability domain (*that is a set of compounds for which the model is reliable*): ideal applicability domain is a set of compounds which have not molecular features with undefined role (*which are not stable promoters of increase or decrease of endpoint*).

10. The simplest method as a rule gives models with highest reliability. But often there is the conflict between the reliability and accuracy.

Results

A robust CORAL model should be defined according two parameters: threshold (T) and the number of epochs of the Monte Carlo optimization (N). Substances can contain various molecular features. The features with wide prevalence apparently should be involved in the modelling process. But rare molecular features can lead to overtraining (where a model is very good for training but poor for test set). The threshold is a parameter for definition of rare (*noise*) features. The rare features are not taken into account in the CORAL models. The Monte Carlo computational experiments have yielded threshold values (T^*) and number of epochs of the optimization (N^*) in order to maximize the correlation coefficient of the CORAL model for the external test set (*Figure 7.2*).

In fact, the correlation coefficient between experimental and calculated values of an endpoint is a mathematical function

1. $R2(test) = F(Threshold, N_{epoch})$

The function calculated according to Eq. 1 is a surface with a number of maximums. One can carry out the computational experiment on some field Ω :

2. $Threshold \in (T_{\sigma}, T_X); N_{epoch} \in (N_{\sigma}, N_Y)$

The above-mentioned T^* and N^* can be the result of the computational experiment. However, satisfactory results can be due to a “lucky” split into the training and the test sub-sets. Consequently, the calculation should be carried out with several random splits, in order to estimate the predictability of the method used.

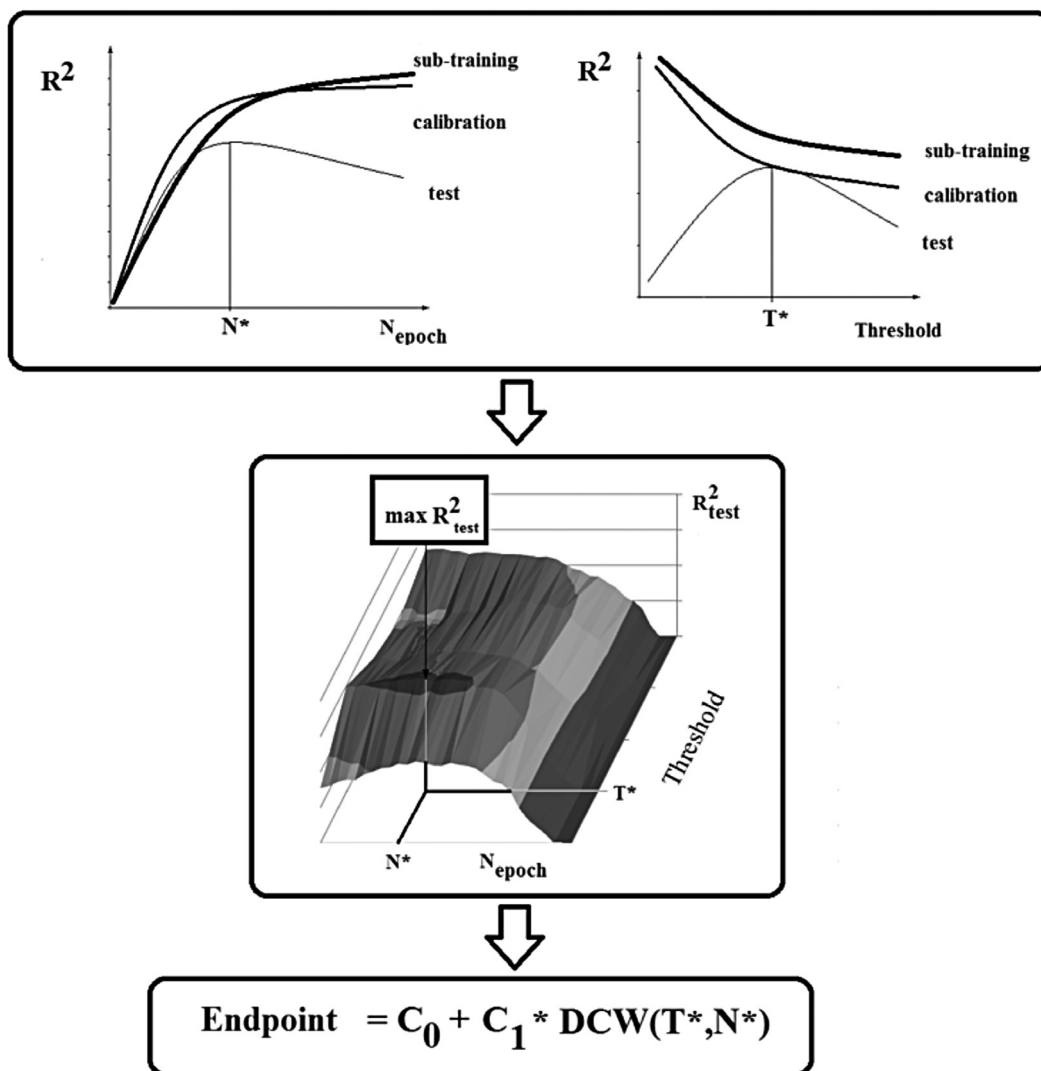


Figure 7.2: The scheme of the building up the CORAL model for an arbitrary endpoint. DCW is the descriptor of correlation weights of the molecular features involved in the model.

It is to be noted that the data on the correlation weights of molecular features can serve as basis for mechanistic interpretations of various endpoints, where various molecular features play the role of promoters of increase of the endpoint or vice versa promoters of decrease of this endpoint.

Perspectives

Possible ways to improve the CORAL software are:

- (1) extension of list of SMILES attributes which are involved in the modelling process; in particular, one can mark molecules by additional symbols, e.g. '{', '}', '£', etc. in order to take into account symmetry, H-bonds, van der Waals interactions, etc.;
- (2) extension of list of graph invariants; and
- (3) involving of InChI as additional representation of the molecular structure.

Conclusions

The CORAL software provides the user a dense web of possibilities to check various methods. Different methods can be obtained using a different list of the molecular features involved in the modelling process. As rule, a complex method (*with a large number of molecular features*) gives excellent model for the training set but a poor model for the test set. Consequently, the user must define a compromise that gives a satisfactory model for both training and test sets.

References

1. García J., Duchowicz P.R., Rozas M.F., Caram J.A., Mirífico M.V., Fernández F.M. and Castro E.A., A comparative QSAR on 1,2,5-thiadiazolidin-3-one 1,1-dioxide compounds as selective inhibitors of human serine proteinases, *J. Mol. Graph. Model.* 31 (2011) 10–19.
2. Garro Martinez J.C., Duchowicz P.R., Estrada M.R., Zamarbide G.N. and Castro E.A., QSAR Study and Molecular Design of Open-Chain Enaminones as Anticonvulsant Agents, *Int. J. Mol. Sci.* 2011, 12, 9354-9368.
3. Ibezim E., Duchowicz P.R., Ortiz E.V. and Castro E.A., QSAR on aryl-piperazine derivatives with activity on malaria, *Chemometr. Intell. Lab.* 110 (2012) 81–88.
4. Mullen L.M.A., Duchowicz P.R. and Castro E.A., QSAR treatment on a new class of triphenylmethyl-containing compounds as potent anticancer agents, *Chemometr. Intell. Lab.* 107 (2011) 269–275.
5. Toropov A.A., Toropova A.P., Benfenati E., Gini G., Leszczynska D. and Leszczynski J., SMILES-based QSAR approaches for carcinogenicity and anticancer activity: Comparison of correlation weights for identical SMILES attributes, *Anti-Cancer Agents in Medicinal Chemistry*, 11 (10), 2011, pp. 974-982.
6. Benfenati E., Toropov A.A., Toropova A.P., Manganaro A. and Gonella Diaza R., CORAL software: QSAR for anticancer agents, *Chem Biol Drug Des.* 2011 Jun;77(6):471-6.
7. Toropova, A.P., Toropov, A.A., Martyanov, S.E., Benfenati, E., Gini, G., Leszczynska, D., Leszczynski, J. CORAL: QSAR modeling of toxicity of organic chemicals towards *Daphnia magna* (2012) *Chemometrics and Intelligent Laboratory Systems*, 110 (1), pp. 177-181.
8. Toropov A.A., Toropova A.P., Martyanov S.E., Benfenati E., Gini G., Leszczynska D. and Leszczynski, J., Comparison of SMILES and molecular graphs as the representation of the molecular structure for QSAR analysis for mutagenic potential of polyaromatic amines (2011) *Chemometrics and Intelligent Laboratory Systems*, 109 (1), pp. 94-100.

9. Toropova A.P., Toropov A.A., Benfenati E., Gini G., Leszczynska D. and Leszczynski J., CORAL: Quantitative structure-activity relationship models for estimating toxicity of organic compounds in rats (2011) *Journal of Computational Chemistry*, 32 (12), pp. 2727-2733.
10. Toropova A.P., Toropov A.A., Benfenati E. and Gini G. Co-evolutions of correlations for QSAR of toxicity of organometallic and inorganic substances: An unexpected good prediction based on a model that seems untrustworthy (2011) *Chemometrics and Intelligent Laboratory Systems*, 105 (2), pp. 215-219.
11. Toropova, A.P., Toropov, A.A., Benfenati, E. and Gini, G., QSAR modelling toxicity toward rats of inorganic substances by means of CORAL (2011) *Central European Journal of Chemistry*, 9 (1), pp. 75-85.
12. Toropov A.A., Toropova A.P., Lombardo A., Roncaglioni A., Benfenati E. and Gini G. (2011) CORAL: Building up the model for bioconcentration factor and defining its applicability domain, *European Journal of Medicinal Chemistry*, 46 (4), pp. 1400-1403.
13. Toropova A.P., Toropov A.A., Gonella Diaza R., Benfenati E. and Gini G. (2011) Analysis of the co-evolutions of correlations as a tool for QSAR-modeling of carcinogenicity: An unexpected good prediction based on a model that seems untrustworthy, *Central European Journal of Chemistry*, 9 (1), pp. 165-174.
14. Toropova A.P., Toropov A.A., Benfenati E., Gini G., Leszczynska D., Leszczynski J. (2011) CORAL: QSPR models for solubility of [C 60] and [C 70] fullerene derivatives, *Molecular Diversity*, 15 (1), pp. 249-256.
15. Toropova A.P., Toropov A.A., Benfenati E., Leszczynska D. and Leszczynski J. (2010) QSAR modeling of measured binding affinity for fullerene-based HIV-1 PR inhibitors by CORAL (2010) *Journal of Mathematical Chemistry*, 48 (4), pp. 959-987.

