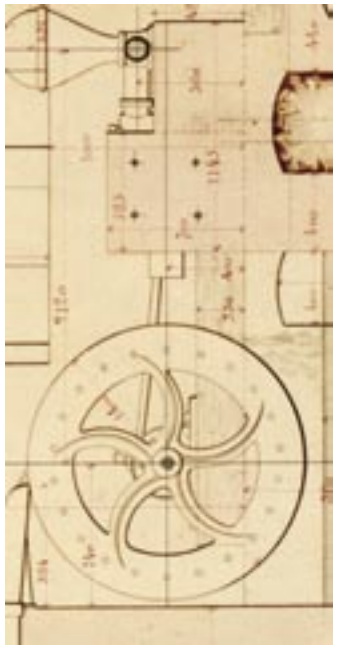
ALTA SCUOLA POLITECNICA

POLITECNICO DI MILANO
POLITECNICO DI TORINO



CHEMISTRY, TOXICOLOGY, and QSAR: an introduction

Luigi Cardamone, Computer Engineering, Politecnico di Milano
Magdalena Gocieva, Computer Engineering, Politecnico di Milano
Marina Mancusi, Biomedical Engineering, Politecnico di Torino
Rima Padovani, Biomedical Engineering, Politecnico di Torino
Lorenzo Tamellini, Mathematical Engineering, Politecnico di Milano

Principal Academic Tutor:
Giuseppina Gini, Politecnico di Milano

Other Academic Tutors:
Bartolomeo Montrucchio, Politecnico di Torino

External Tutor:
Emilio Benfenati, Istituto di Ricerche Farmacologiche ‘Mario Negri’

1 Summary

1.1 Outline

In our everyday life we have to deal with an exponentially increasing number of different chemical compounds: the number of registered chemicals is about 28 millions, including food colouring and preservatives, drugs, varnishes and paints for wearing and objects, pesticides and many others. It is well recognised that chemicals may pose high risk to environment and to humans, and hence their toxic activity has to be assessed.

Biological active substances interact with bio-molecules, triggering specific mechanisms like activation of an enzyme cascade or opening of an ion channel, which finally leads to a biological response. These mechanisms, determined by the chemical composition of the considered substances, are unfortunately largely unknown, and thus one has to study toxicity experimentally.

It is possible to conduct at least three different kinds of experiments to assess the biological activity of a molecule: "in vivo" experiments, that is animal testing, "in vitro" experiments, that is using tissue culture cells, and "in silico" experiments, which refers to computer simulations.

Figure 1: the three experimental ways

Both animal testing and “in vitro” experiments are of course time consuming and expensive, and especially animal testing is considered ethically unacceptable by a growing majority of people. For these reasons, and also thanks to the increased power of computers, the scientific community and the industrial world focused on “in silico” experiments, which represent a more mathematical and computational approach, developing a number of models and strategies to predict physical and chemical properties of compounds.

Computational chemistry has changed the classical way to make experimental science. We are more and more 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 molecular descriptors, ranging from simple properties to complex molecular fingerprints. Descriptors can help in transforming the study of interactions of molecules with living organisms in a kind of data mining problem. Data mining could find relevant correlations between descriptors and the response of interest. It is not enough; if we want to make predictive models we need to assess also the predictive power of our relations.

In this project we will focus on a particular kind of data mining applications, called QSARs (Quantitative Structure Activity Relation). These 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, cancinogenicity, skin sensitization, that is called "endpoint of interest".

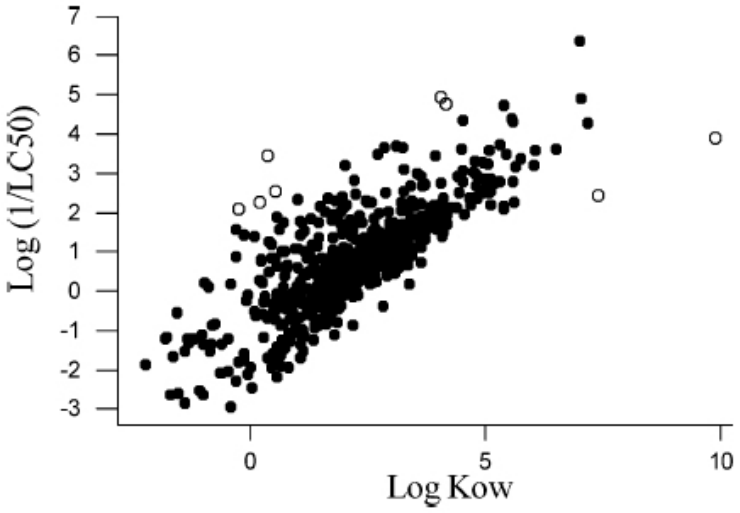


Figure 2: an example of QSAR analysis. The descriptor is Log Kow; the endpoint is death and it is measured as a lethal concentration (LC 50)

In other words, once the properties of a compound are quantified in a set of variables, called "descriptors", these algorithms model the toxicity of a compound with respect to a given endpoint as a function of some descriptors, and the coefficients of the function are estimated through statistical analyses applied on a sufficiently large database.

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 forecasting tool in drug design, environment protection, hazard analysis, for all those compounds whose structure is similar to the structure of the ones used to tune the model.

The usage 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 fixes the rules for chemical production in E.U., and one of its key points is that it requires a risk analysis for each chemical placed in European market in amount greater than 1 ton/year. To further underline the broadness of this law, suffice it to say that it is 849 pages long, and international mass-media defined it 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).

1.2 How this report is organized

In Section 2 we illustrate the three basics areas that constitute the necessary technical background needed to develop our work: toxicity mechanisms, QSAR building and calculation of descriptors. Section 3 is dedicated to a short presentation of the REACH legislation.

In Section 4 we analyze more in details the needs of stakeholders involved in the problem. For what concerns the researchers we will explore the issues caused by tight copyright policies on software, focussing especially on the difficulties in getting reliable values for descriptors due to the calculations performed with non Open Source code; we will also consider the lack of standardization in descriptors, and how to deal with unknown toxicity mechanisms. From the industrial point of view, we will discuss the problems due to REACH legislation and in particular to the QSAR usability for risk assessment.

In Section 5 we give a summary of the State of The Art: examples of effective QSAR models, existing software and other help desks for companies.

2 Toxicology and prediction

2.1 Toxicology principles: modes of action

Toxicology (from the Greek words *toxicos* 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 kilo, while others may be quite toxic even if their concentrations are much higher (some grams per kilo). 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 LD_{50} (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 *Aspergillus flavus*)	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)

Figure 3: LD_{50} 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 known 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. The cellular death is the direct consequence of the damage induced by chemical agents and it may lead to serious effects on the tissue in exam. In fact a lot of toxic responses come from the cellular death and the loss of efficiency of crucial organs, that affect the functionality of the whole organism. However different responses exist, which are not due to the cellular death, but they 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 (usually of an extracellular origin) 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 little changes of the 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.

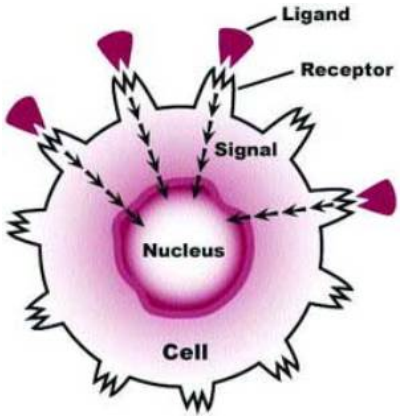


Figure 4: Functional scheme of the ligand-receptor interaction. This interaction is very selective and it influences the inner cell mechanisms.

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

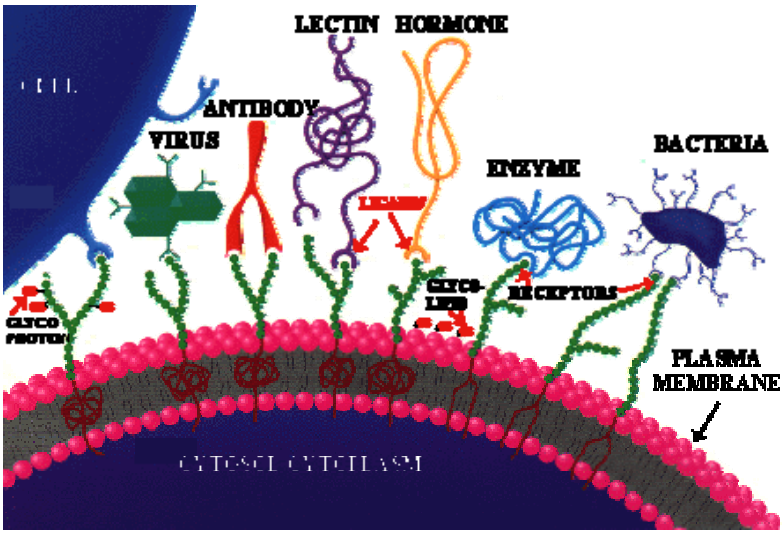


Figura 5: Different types of ligand-receptor interaction: different cell membrane receptors interact with their specific extracellular ligands.

In other cases chemicals become toxic interfering in the normal oxidation process of carbohydrates, that lead to the production of Adenosine Tri-Phosphate (ATP). Some of them interfere in the oxygen supply to tissues, 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 lot 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 example is the carbon monoxide, which binds iron in the hemoglobin with high affinity, causing the reduction of the 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 the 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.

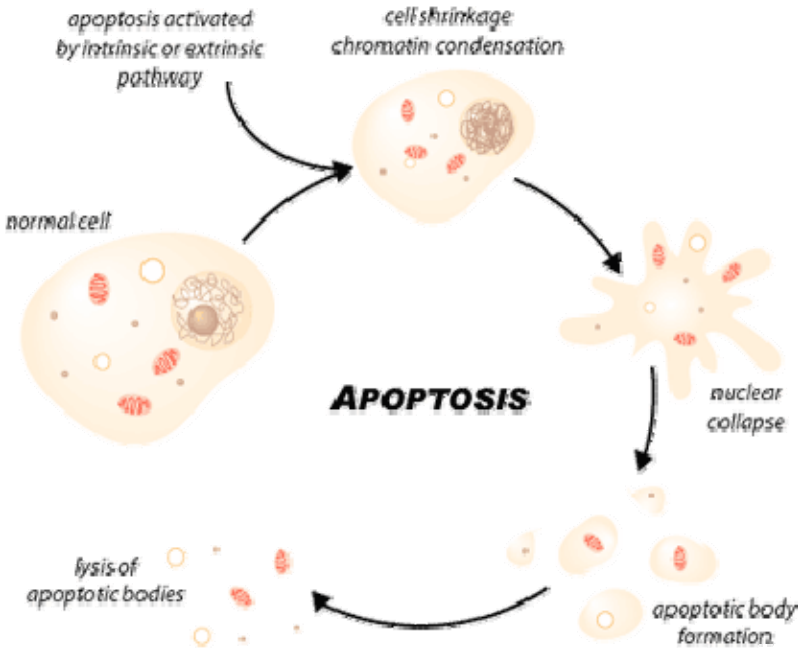


Figure 6: Description of the apoptosis process. Although many pathways and signals lead to apoptosis, there is only one mechanism that actually causes the death of the cell in this process; after the appropriate stimulus has been received by the cell, that cell will undergo the organized degradation 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 (know as blebs), cell breaking apart into several vesicles called apoptotic bodies, which are then phagocytosed.

Another kind of toxic effect is the selective death of cells within an organ or a tissue, that 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.

At last 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 would 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 the DNA damage (cancer promoters).

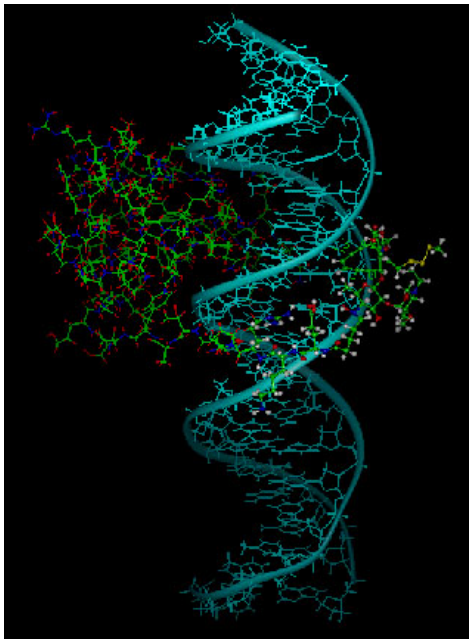


Figure 7: The DNA cleavage made by Calicheamicin. This is a highly toxic substance but has found pharmacological use as a conjugate in antibody cancer therapeutics (e.g. Gemtuzumab), because it binds DNA, chops it into pieces and thus kills the tumor cells..

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 be made 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 other ones are not easily attributable to a specific biochemical mode or a particular enzyme [12].

Finally it is important to emphasize that the mode and the site of action of the majority of substances is not well known yet. This knowledge gap is actually the main problem that has to be faced assessing the toxicity of a chemical by alternative methods, such as QSARs. Nowadays several steps 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.

2.2 Theory of QSAR models

2.2.1 What are molecular descriptors

Nowadays, the concept of molecular structure is one of the most important concepts in the scientific development. The reasoning based on the molecular structure has played a major role in the development of physical chemistry, molecular physics, organic chemistry, quantum chemistry, chemical synthesis, medicinal chemistry, etc.

A system is complex by definition when its behaviour as a whole can not be derived from the properties of its parts. It is evident that a molecule, with its embedded concept of molecular structure, exactly fulfils these conditions. The properties of a molecule do not depend only on the properties of the component atoms but also on their mutual connections. Therefore, they are inherent to the whole molecular organisation and stability and cannot be derived as the sum of the properties of the component atoms. In principle, we have a holistic system.

Due to its complexity, molecular structure cannot be represented by a unique formal model. Depending on the level of the underlying theoretical approach, several molecular representations can represent the same molecule. These representations however are often not derivable from each other.

Different molecular representations have been proposed, for example the 3-dimensional Euclidean representation, the 2-dimensional representation based on the graph theory, or the vector representations called fingerprints where the frequencies of several molecular fragments are stored. Each of these representations constitutes a different conceptual model of the molecule and by each model different sources of chemical information become available.

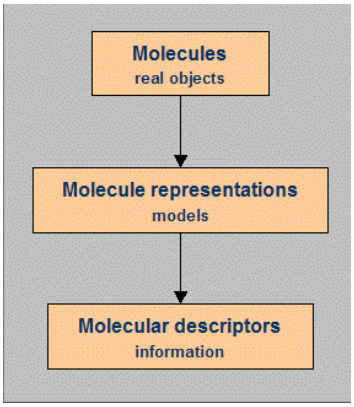


Figure 8: From molecules to molecular descriptors

The molecule implicitly contains all the chemical information, however, only a part of this information can be extracted by means of experimental measurements. This is where the concept of molecular descriptors comes into use: they are numbers able to extract small pieces of chemical information from the different molecular representations.

So what exactly is a molecular descriptor? It is any molecular property that characterizes the molecule, or more specifically: "The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment." The term “useful” in this previous definition has a double meaning: it means that the number can give more insight into the interpretation of the molecular properties and/or is able to take part in a model for the prediction of some interesting property of other molecules.

In the last decades, many researches have been focused on studying ways to capture and convert the information encoded in the molecular structure into one or more molecular descriptors. In such way, quantitative relationships can be established between structures and properties on the one hand, and biological activities and other experimental properties on the other.

Therefore, molecular descriptors are now playing a key role in scientific research. Evidence of the interest of the scientific community in the molecular descriptors is provided by the huge number of descriptors proposed until today: more than 3000 descriptors are actually defined and computable by using dedicated software tools. Each molecular descriptor represents a small part of the whole chemical information contained into the real molecule and, as a consequence, the large number of descriptors is continuously increasing with the increasing of the complexity of the investigated chemical systems.

The molecular descriptors range from simple counts of atoms, functional groups, or molecular weight, to properties based on 3D structure. They may require measurement of a physicochemical property, which is not recommended for a predictive technique, or may be calculated from knowledge of the contributions of functional groups, molecular topology or electron distribution. Despite their diversity, they are considered to describe only three aspects of a molecule – its hydrophobic, electronic and steric properties, which are held responsible for the biological activity in an organism.

Physicochemical properties tend to describe fundamental molecular effects, such as the partition coefficient being a measure of hydrophobicity. It is believed that these properties may be related to the mechanism of action of a molecule and are therefore less susceptible to spurious correlation. That is why many modellers claim that predictive techniques should be based on fundamental physicochemical properties, such as the partition coefficient, water solubility, ionization or dissociation constant, melting point, boiling point, etc. While the properties of a pure chemical substance may be measured, the computation of physicochemical properties has many advantages. These include the speed and low cost of calculation and, more importantly, the fact that calculation may be performed for chemicals that are not available.

A subgroup of descriptors, named topological indices (for example, Wiener Index, Harary Index, Zagreb Indices, etc.), has become a point of interest in recent years. The topological indices are derived from knowledge of the connections within a molecule and are formulated from graph theory. They account for many molecular properties, such as molecular size, branching and to a certain degree shape.

Molecular orbital theory has also been widely applied to the prediction of toxicity and fate, for example metabolism, persistence, or biochemical reactivity. The vast increase in the speed of computers has enabled the rapid calculation of numerous atomic and molecular orbital descriptors, such as charges, dipole moment, energy levels, etc.

The field of molecular descriptors is interdisciplinary and encompasses many different theories and disciplines. Knowledge of algebra, graph theory, information theory, computational chemistry, physical chemistry, quantum chemistry as well as theories of organic reactivity is needed for the definition of molecular descriptors. For their use, knowledge of statistics and chemometrics in addition to the specific knowledge of the problem is needed, as well as the principles of the QSAR approaches which will be explained in more detail later on. In addition to all this, sophisticated software and good programming skills are often a requirement. Molecular descriptors constitute a field where the most diverse strategies for scientific discovery can be found.

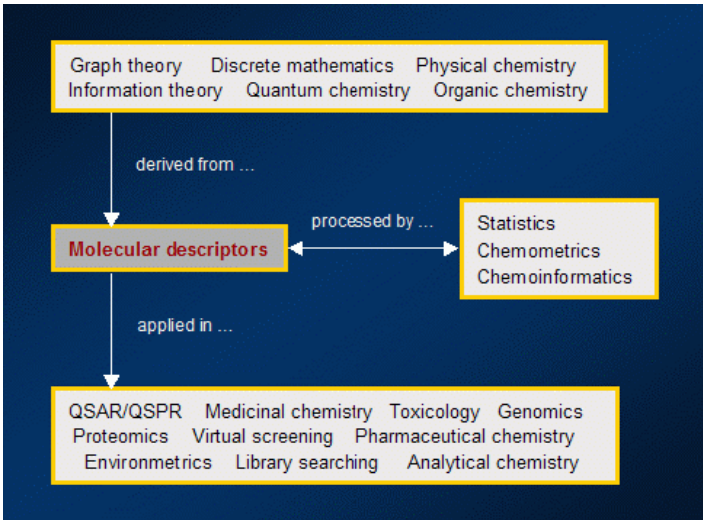


Figure 9: Theories and disciplines encompassed by the filed of molecular descriptors

2.2.2 Which are the endpoints of interest

Molecular descriptors have become part of the most important variables used in molecular modelling. Until about 30 years ago, molecular modelling was based on discovering mathematical relationships between experimentally measured quantities. Nowadays, it is performed modelling a measured property by the use of molecular descriptors which capture structural chemical information.

There is a remarkable number of diverse activities that have been successfully modelled. The activities of interest here are related to toxicology and drug design, for example the toxicity of a chemical to an environmental organism or human being, the fate of a pollutant in an ecosystem, the pharmacokinetic properties of a xenobiotic in humans, etc. The measures of these activities, which are quantities somehow related to the toxic activity of the molecules, are called endpoints and represent the output variables of the modelling.

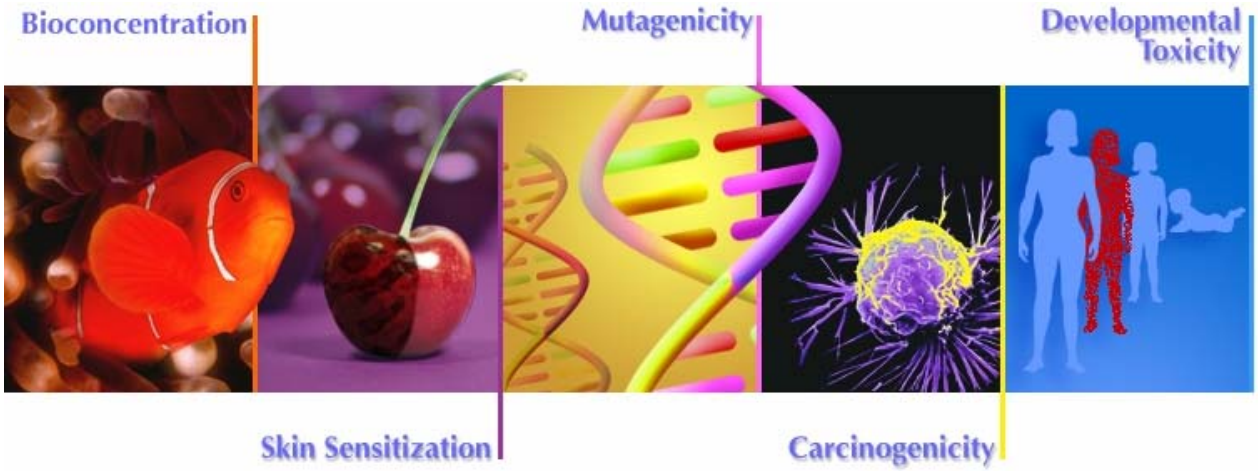


Figure 10: examples of endpoints. These are the endpoints of interest in CAESAR project.[1]

[1] For further information see: http://www.caesar-project.eu/

From the point of view of the type of biological activity that is being modelled, these toxicological endpoints can be divided in two major groups: human health endpoints and environmental toxicity endpoints.

Two very important human health endpoints are chemical carcinogenicity and mutagenicity. Carcinogens can be genotoxic, which interact directly with DNA and are thought to work by inducing mutations, or epigenetic, which act through mechanisms that do not involve direct DNA damage, however, no unifying theory exists for their mode of action. On the other hand, mutagens provoke heritable changes to the genetic material. The modelling of chemical carcinogenicity and mutagenicity is a very important goal in toxicology because of the huge impact they have on the quality of life and because of the enormous investment in time, money and animal lives needed to test chemicals adequately.

Other human health endpoints of great significance are chemical metabolism and biotransformation of chemicals within biological organisms, in order to determine their biological effects such as their effectiveness or toxicity as agents, for example pharmaceuticals or agrochemicals. Their modelling is motivated by the need for an efficient screening of large numbers of chemicals and the concern to reduce animal testing. Until recent years, a less intensely researched area has been the modelling of pharmacokinetic parameters. As the focus of the pharmaceutical or agrochemical industry has shifted towards faster and smarter screening in order to avoid failure of candidate chemicals, pharmacokinetic parameters such as absorption, distribution, metabolism and excretion (ADME) are becoming toxicological endpoints with increasing significance. When developing a new drug, it is important that the drug reaches its site of action in adequate concentration without accumulating in the body or producing side-effects. The pharmacokinetic properties must be optimized such that the drug will be readily absorbed, transported to the appropriate site and eliminated from the body in a timely manner. Approximately 40% of drug candidates fail during preclinical tests as a result of unacceptable pharmacokinetics. Due to the high economic cost of failure at this late stage, the benefits of the ability to screen out the likely failures at an earlier stage are evident.

In the area of environmental toxicity, some of the more important endpoints are persistence, bioaccumulation and soil sorption.

Persistence relates to the length of time that a predefined fraction of a substance remains in a particular environment before it is physically transported to another compartment and/or chemically or biologically transformed. Knowledge of persistence is crucial for evaluating the hazard and risk from chemicals released into the environment. In general, higher risk is assumed with increasing persistence of compounds. Persistence is not a simple endpoint, but a complex phenomenon comprising biodegradation/transformation, accumulation, distribution and transport processes.

Bioaccumulation is defined as uptake by an organism of a chemical from the environment by any possible pathway. On the other hand, soil sorption or sorption of a chemical (usually from water) by soil prevents the chemical from being transported further in the environment, although temporarily in most of the cases. In order to quantify the risks that environmental pollution poses, an understanding of the many factors that affect the distribution of a chemical in the environment is needed, two of which are precisely bioaccumulation and soil sorption.

From a statistical point of view two types of biological activity can be modelled: continuous and categorical.

Continuous data are numerical values that describe a concentration that provokes a particular effect. Usually this is a 50% effect lethal concentration, such as 96-h LC_{50}, defined as the concentration that kills half of an animal test population in 96 hours. A subgroup of continuous data for modelling is not necessarily biological in nature, and it includes physicochemical properties (partition

coefficient, melting point), fate descriptors (persistence, bioaccumulation), pharmacokinetic properties (bioavailability, metabolism) and many others.

Categorical data are typically yes/no data that classify a chemical as being active or inactive in a particular toxicological assay (for example, carcinogenic or non-carcinogenic), or may be descriptive such as high or low bioavailability. For some activities and endpoints a classification may be assigned on the basis of a particular number of criteria, originally developed from quantitative measures of toxicity.

2.2.3 Statistic theory of QSAR / How to construct models

Chemicals have various effects on organisms to which they are exposed, some of which are desirable and some are undesirable. Nowadays, the number of these chemicals is rapidly increasing, in pace with the fast industrial development. That is why it is extremely important to assess their potential toxicological effects on organisms and the environment.

Two main streams have been developed in order to explain the complex relationships between molecules and observed quantities, or endpoints. The first one is related to the search for relationships between molecular structures and physicochemical properties and is called QSPR (Quantitative Structure-Property Relationships). The second one, which is the focus of our work, is related to the search for relationships between molecular structures and biological activities and is called QSAR (Quantitative Structure-Activity Relationships).

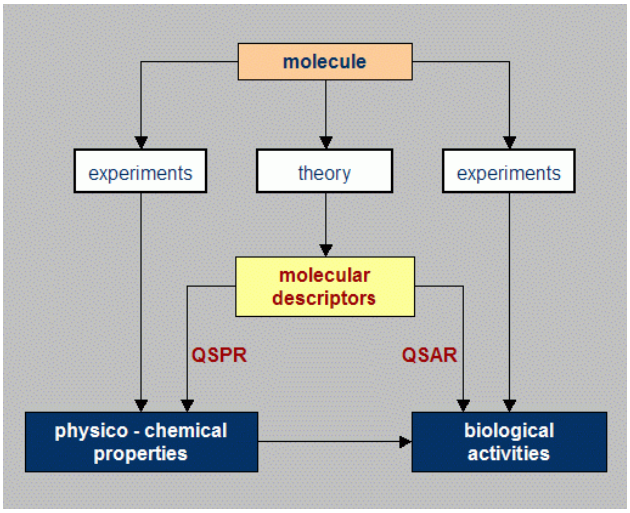


Figure 11: QSPRs/QSARs relating molecular structures with endpoints

Since chemical structure was elucidated, the relationship between chemical structure and biological activity has intrigued scientists. It has been recognized that the investigation of QSARs may provide useful tools for obtaining information regarding the effects of chemicals on man and the environment. Initially developed to assess the value of drugs, QSARs are now proposed as a method to assess general toxicity.

QSARs are based on the assumption that the structure of a molecule (its geometric, steric and electronic properties) contains the features responsible for its biological activity. For example, as already explained in the previous sections, biological activity can be expressed quantitatively as in

the concentration of a substance required to give a certain biological response. When the information encoded in the molecular structure is expressed by molecular descriptors in the form of numbers, one can form a quantitative structure-activity relationship between the two. By QSAR models, the biological activity of a new or untested chemical can be inferred from the molecular structure of similar compounds whose activities have already been assessed.

QSAR's most general mathematical form is:

Activity = f (physiochemical properties and/or structural properties)

It is therefore evident that the three key components required for the development of a QSAR model are:

- Some measure of the activity (in this case toxicity) for a group of chemicals in a biological or environmental system – toxicological endpoint
- A description of the physicochemical properties and/or structure for this group of chemicals – molecular descriptors
- A form of statistical relationship to link activity and descriptors

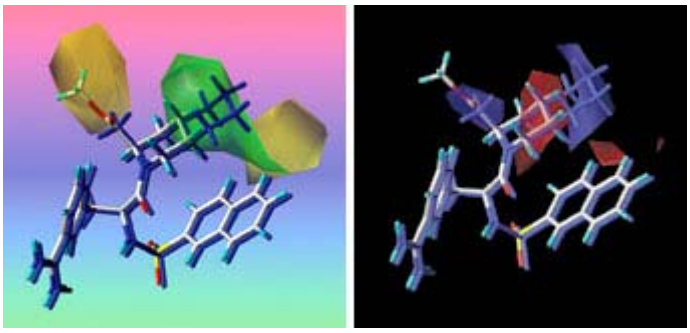


Figure 12: QSAR model visualization: building graphical models that relate biological activity of molecules to their structure

At first sight, the selection of compounds for developing QSAR models may appear to be self-evident. If we are interested in the biological effects of a certain group of chemicals we collect or measure all the compounds of that group that can be found. However, this strategy is not the best way to gather data and it may happen that too many results for the wrong compounds prevent the establishment of a good QSAR model. The successful construction of QSAR models requires experimental design, in which each compound included corresponds to a design point and the experimental factors that need to be varied in order to create the design are the physicochemical properties that characterize the compounds. The toxicological endpoint can also be an experimental factor and the goal is to develop a model that links the endpoint to the physicochemical descriptors. It is crucial that the design includes compounds that give both high and low values of the endpoint of interest, and if possible, a uniformly-spread range of intermediate values.

The response data, which are measures of the biological activity of compounds and represent the output variables in the QSAR models, can be measured directly by the investigators or collected from the literature. Knowledge of the precision and range of these data is of high importance. Some measurements have a natural range, but others may cover many orders of magnitude which may be

deceptive. It is therefore dangerous to take these data at face value. Examination of their distribution can be very useful because it can indicate where a certain type of processing is required. The precision is another property of interest because the model should have a standard error no better than the measurement errors. This is because of the fact that it should not be possible to calculate something more precisely that it can be measured. A standard error that is better (less) than the experimental one is a good indication that the model has been overfitted, which means that it fits the training data set well, but cannot generalize to other sets, which is the purpose for fitting a model.

The descriptor data, which capture information about the chemical structure of compounds and represent the input variables in the QSAR models, can be obtained from a variety of sources. In the early period of QSAR modelling, the choice of the descriptors was limited because they were generally tabulated physicochemical properties. Nowadays, there are over 3000 different molecular descriptors and it is common to use many more descriptor variables than there are compounds in the set when building a QSAR 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 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, whose aim 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 the forward stepping multiple regression.

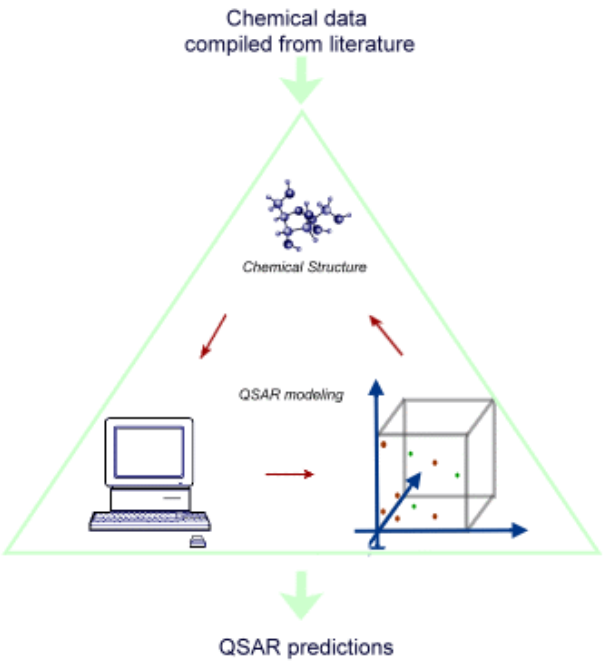


Figure 13: The process of QSAR modelling for predicting the biological activity of novel compounds

In this context, it must be mentioned that one of the major problems in QSAR modelling is the availability of high quality experimental data for building the models. The input data must be both accurate and precise in order to develop a meaningful model. Any developed QSAR model is statistically as valid as the data that led to its development.

In addition to this, a problem related to molecular descriptors is their reproducibility: experimental values can differ greatly even when referred to the same compound. As an illustration, several approaches have been developed for the theoretical calculation of the partition coefficient (log P), but in these calculations it is not uncommon to have differences of several orders of magnitude. In modern QSAR approaches, it is common to use a wide set of theoretical molecular descriptors of different kinds which take into account the various features of the chemical structure. There are many software packages that calculate wide sets of different theoretical descriptors. The greatest advantage of theoretical descriptors is the fact that they can be calculated homogeneously by defined software for all chemicals, including those not yet synthesized but represented by a hypothesized chemical structure, and therefore they are reproducible.

A variety of methods for building QSAR models exists. 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 are 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 failure to invert a matrix, while others may give a false, 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: methods that deal with classified responses (for example, mutagen / not mutagen, toxic / slightly toxic / non toxic) and methods 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 a result that is easy to understand. MRA is a powerful means for establishing a correlation between independent variables (molecular descriptors) and a dependent variable (biological activity). In addition, Artificial Neural Networks can be used for modelling both classified and continuous data.

After the model is developed, regardless of the type, it is of crucial importance to assess its performance by validating its predictive application. Most statistics packages generate a variety of statistical quantities for the common modelling approaches which will enable a judgement of significance and will give some guidance on whether the model may have arisen by chance. This is

based on the assumption that the data conform to some statistical distribution, usually multivariate normal. Unfortunately, this only indicates how well the model fits the data within the modelling assumptions and does not really give any information on how well the model might work. The best fit models are not the best ones for prediction. As a consequence, the only way to know how well a model may work is to try it out.

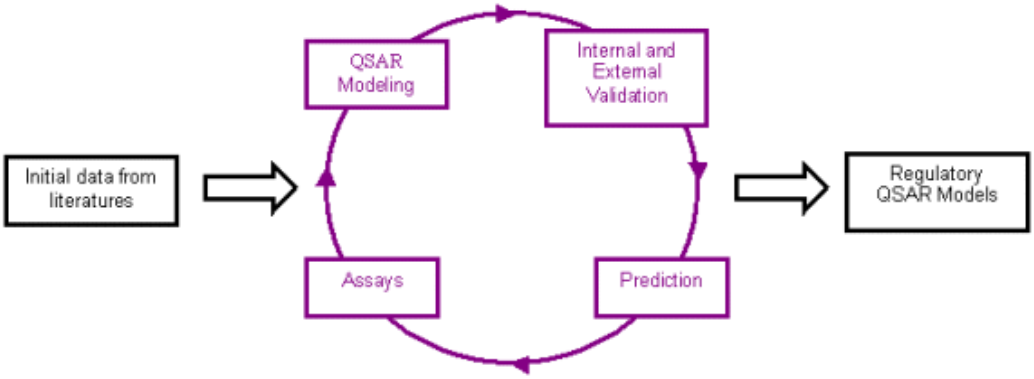


Figure 14: Depiction of the recursive process for developing QSAR models

One common approach is the Leave-One-Out Cross Validation (LOO or CV), which 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.

However, none of these procedures allows us to judge whether a relationship is real or it has happened by chance. One way to check for chance effects is to scramble the response values and then try to build models using the scrambled data. This can be repeated a number of times and some fit statistics, such as R_2, can be tabulated for the resulting models. If the R_2 value for the model of the unscrambled response is higher than the R_2 value for the scrambled sets, it is reasonable to assume that the model is not a chance fit.

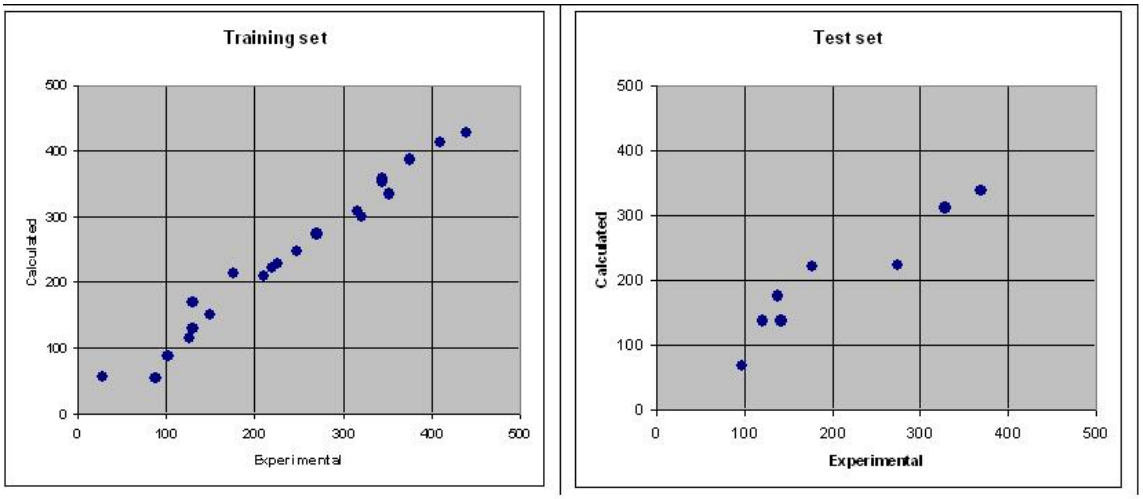


Figura 15: QSAR model validation by using a training set and a test set

Obviously, the best test of a model is to present it with unseen data, ether by holding back some of the original data to form a test set or by synthesizing or testing some more compounds once the model has been built. Only a stable and predictive model can be considered a reliable model and can be usefully interpreted for its mechanistic meaning.

There is an argument that, if the main aim of QSAR modelling is simply prediction, the attention should be focused on model quality and it is not necessary to try to interpret models. Another argument is that it is dangerous to attempt to interpret models, since correlation does not imply causality. Regarding the interpretability of QSAR models, Livingstone states: "The need for interpretability depends on the application, since a validated mathematical model relating a target property to chemical features may, in some cases, be all that is necessary, though it is obviously desirable to attempt some explanation of the "mechanism" in chemical terms, but it is often not necessary, per se". On this basis, we can differentiate predictive QSARs, where the focus is best prediction quality, from descriptive QSARs, where the focus is descriptor interpretability.

To summarize all that was previously mentioned, what makes a successful QSAR model? The ideal QSAR model should: (1) consider an adequate number of molecules for sufficient statistical representation, (2) have a wide range of quantified end-point potency (for example, several orders of magnitude) for regression models or adequate distribution of molecules in each class (for example, active and inactive) for classification models, (3) be applicable for reliable predictions of new chemicals (validation and applicability domain) and (4) allow to obtain mechanistic information about the modelled end-point.

2.2.4 Ethic and economic impacts of QSAR

It has been more than 40 years since QSAR modelling was first used in the practice of agrochemistry, drug design, toxicology, industrial and environmental chemistry. Its growing power in the following years may be attributed to the rapid and extensive development in of methodologies and computational techniques that have allowed to delineate and refine many variables and approaches used in this modelling approach. Initially developed to assess the value of drugs, QSARs are now proposed as a method to assess general toxicity. They initiated a radical change in the way of thinking and leading toxicological studies. They aim at going beyond the limits of the traditional approach and facing the complexity of the biological world through a deeper analysis of the intrinsic toxicity mechanisms of actions and their driving forces. QSAR modelling is

a challenging approach and the whole scientific community agrees on the numerous potential advantages that could come from the application of QSARs.

There are many reasons why one may wish to predict the toxicity of chemicals. It is fundamental that computer models allow for the effects of chemicals to be predicted and these predictions may be obtained from knowledge of chemical structure alone. For most methods, provided that the chemical structure can be described in two or three dimensions, the effects may be predicted. Information regarding the chemicals may be gained without chemical testing, or even the need to synthesize the chemical. QSARs are therefore often employed to establish a correlation between structural features of potential drug candidates and their binding affinity towards a macromolecular target in order to identify chemical structures that could have good inhibitory effects on specific targets and have low toxicity. In addition to designing in attractive features of molecules in drug and pesticide design, it is now possible to design out toxic features.

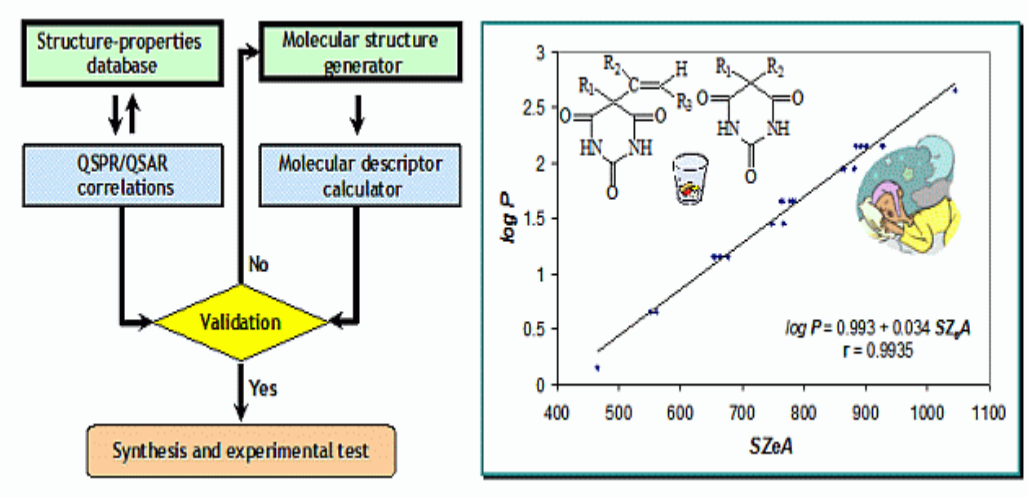


Figure 16: Predicting the toxicity of chemicals in drug design

Furthermore, approximately 100000 separate chemicals may be released into the environment annually and it is therefore frightening to consider that reliable toxicity data exist for only a tiny proportion of these chemicals, probably less than 5 percent. Computer-aided prediction of toxicity has the capability to assist in the prioritisation of chemicals for testing, and for predicting specific toxicities to allow for labelling.

For several decades there has been a growing public concern regarding the use of animals in testing, especially in toxicology and medical research. This has resulted in the boycotting of companies, organizations and individuals associated with animal testing. Campaigners for animal welfare cite a number of approaches to reduce and ultimately replace animal tests. There is clearly a role for predictive techniques in the replacement of animal tests, either as stand-alone methods, or more commonly as part of a tiered assessment strategy. The integration of computational methods in combination with the judicious use of physicochemical properties is a viable alternative to animal testing. Having in mind that in 2002 in Europe it is estimated that 10.7 million animals were used for experimental aims, QSARs would save a lot of animals and solve ethic problems about their use.

Moreover, animal testing takes about 1-2 years per compound, so companies sometimes prefer to continue using tried and tested substances rather than starting such a long testing procedure. A

single QSAR may also take 1-2 years to be developed, but once the model is ready to be applied it could drastically reduce the time requested for testing because a lot of compounds may be tested almost immediately (it depends on the type of descriptors and the complexity of the model, but these tests do not go over a pair of days).

In this context, it is also necessary to underline that recent studies have proved that animal testing is not as valid as it is believed to be. The main problem is in the difference between men and animals, thus often some results can not be directly applied to the latter in the same way as the former. QSARs, on the other hand, are based on a mechanistic interpretation of biological activities and so, if we are able to apply them with an adequate level of uncertainty, we could have a more general, and therefore applicable, analytical approach.

Unfortunately, the level of uncertainty that is associated to QSARs is still too high to proceed with the complete substitution of animal testing, even though the legislative framework has accepted their application (at least in theory). QSAR information will most often be used to supplement test data within chemical categories and endpoint-specific Integrated Testing Strategies (ITS). Uncertainty remains the only barrier for QSARs to fully replace animal testing.

Toxicological testing is costly financially as well as in terms of the animals used and the time taken. By using the methods to predict toxicity these costs are greatly reduced, which should allow for faster and less expensive product development, e.g. pharmaceuticals, as well as assessment of environmental effects. According to the study of Pedersen and colleagues, presented on the *Stakeholder Workshop on Impact Assessment of REACH* on 2 November 2003 in Brussels, the cost-saving potential of valid QSARs is estimated to be 700-940 million euros.

An often ignored spin-off from the development of QSARs is the increased understanding they can provide in both the biology and chemistry of active compounds. There are countless examples where knowledge of biology and chemistry has been advanced by modelling in the field of toxicological effects.

Thus, it is expected that huge efforts will be made in order to improve the current incomplete knowledge on toxicological mechanisms and to develop more reliable QSAR models with a lower degree of uncertainty.

3 The REACH regulation

3.1 REACH in general

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 [25].

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 current 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 vertebrates 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 Chemical 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.

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

3.1.1 Registration

As mentioned above, there is the general obligation for manufacturers and importers to submit a registration to the European Chemical Agency for each substance manufactured or imported in quantities of 1 tonne or above 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.

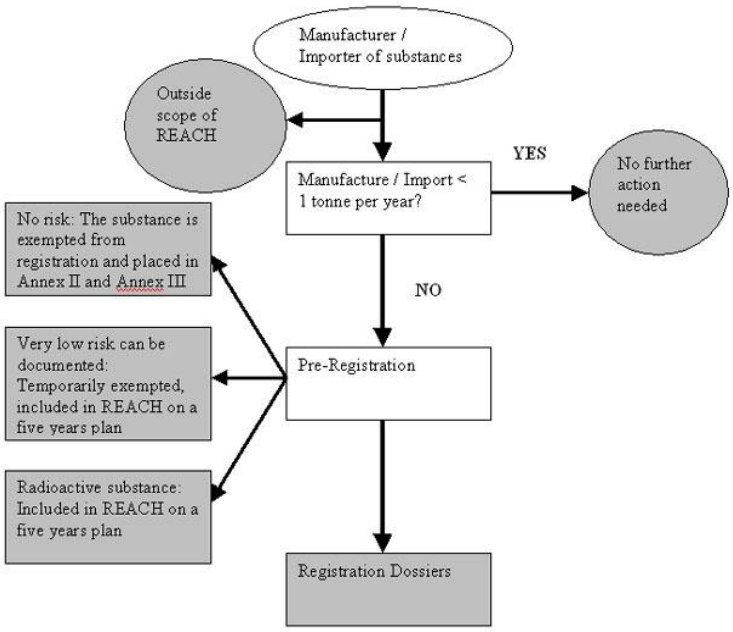


Figura 17: Scheme of the REACH registration process.

3.1.2 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 is provided and made available to the relevant bodies by the Agency.

3.1.3 Authorization

For substances of very high concern, an authorisation is required for their use and their placing on the market. This procedure aims at substituting the most dangerous substances and better managing risks coming 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 from scientific evidence as causing probable serious and normally irreversible effects to humans or the environment, equivalent to the previous ones, on a case-by-case basis, as endocrine disrupters.

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 is 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 not any 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.

3.1.4 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.

Reporting a brief cost and benefit analysis of REACH, it is possible to claim that the introduction of the new regulation would have some relevant benefits, such as:

- positive occupational impact;
- 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 introduce also 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 very 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 or non profitable 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].

3.2 REACH focusing on QSARs

In the ideal situation, QSAR results can be used on their own for regulatory purposes if they are considered relevant, reliable and adequate for the purpose, and if they are documented in an appropriate manner. In practice, there may be uncertainty in one or more of these aspects, but this does not preclude the use of QSAR estimate in the context of a Weight of Evidence approach, in which additional information compensates for uncertainties resulting from the lack of information on the QSAR [7].

A number of conditions need to be met in order for QSAR results to provide an acceptable alternative to experimental data.

There is widespread agreement that models should be scientifically valid if they are to be used in the regulatory assessment of chemicals; since the concept of validation is incorporated into legal texts and regulatory guidelines, it is important to clearly define what it means, and to describe what the validation process might entail.

For the purposes of REACH, an assessment of QSAR model validity should be performed by reference to the internationally agreed OECD principles for the validation of QSARs.

The validation exercise itself may be carried out by any person or organization, but it will be the industry registrant of the chemical who needs to argue the case for using the QSAR data in the context of the Registration process. This is consistent with a key principle of REACH that the responsibility for demonstrating the safe use of chemicals lies with industry.

The principles for QSAR validation identify the types of information that are considered useful for the assessment of QSARs for regulatory purposes; however, fixed criteria will be difficult, if not impossible, to define in a pragmatic way, given the highly context-dependent framework in which non-testing data will be used. Instead, experience and common understanding should be gained by learning-by-doing approach, and by documenting the learning.

Under REACH, there will be no formal adoption process for QSARs; the information generated on the characteristics of a QSAR model will be used as the basis for deciding whether the information on the substance, taken as a whole, is adequate for the regulatory purpose. This process will therefore involve an initial acceptance of the data (including non-testing data) by the industry registrant and the subsequent evaluation, on a case-by-case basis, by the authorities.

The OECD Principles for QSAR validation state that in order "to facilitate the consideration of a QSAR model for regulatory purposes, it should be associated with the following information:

1. A defined endpoint;
2. An unambiguous algorithm;
3. A defined domain of applicability;
4. Appropriate measures of goodness-of-fit robustness and predictivity;
5. A mechanistic interpretation, if possible."

3.2.1 Validity of QSAR model

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 the context of QSARs, this definition is rather abstract and difficult to interpret in relation to the OECD validation principles; thus, for the practical validation of QSAR models intended for use in the regulatory assessment of chemicals, the following operational definition has been proposed: "the validation of a QSAR is the process by which the performance and mechanistic interpretation of a model are assessed for a particular purpose."

In this definition, the performance of a model refers to its goodness-of-fit, 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 a 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.

3.2.2 Reliability of QSAR prediction

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 wonder if a specific QSAR is appropriate for the compound of interest. This means firstly to consider 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 it 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 checking its predictive capability for one or more analogous compounds that are similar to the one 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 implies an expert judgment, which should be clearly rationalized.

3.2.3 Adequacy of QSAR prediction

In order for a QSAR result to be adequate for a given regulatory purpose, the following conditions must be fulfilled:

- the estimate should be generated by a valid (relevant and reliable) model;
- the model should be applicable to the chemical of interest with the necessary level of reliability;
- the model endpoint should be relevant for the regulatory purpose.

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

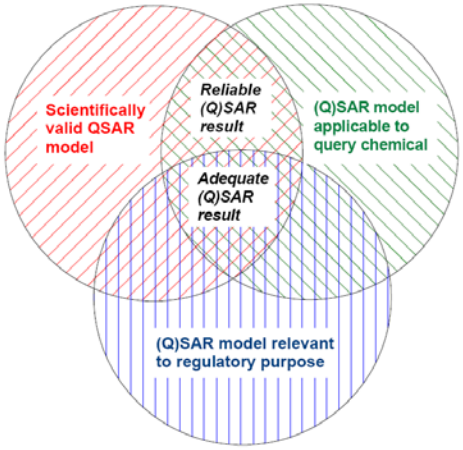


Figure 18: Interrelated concepts of QSAR validity, reliability, applicability, adequacy, regulatory relevance. The circles refer to (Q)SAR models whereas the intersections refer to (Q)SAR results with certain features. In order for a (Q)SAR result to be reliable for a given chemical, it should be generated 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.

Finally, if a registrant intend to use QSAR data instead of experimental data, the adequacy of the QSAR result should be documented by using the appropriate QSAR Reporting Formats. Different types of QSAR Reporting Formats (QRFs) are being developed to provide a standard framework for summarizing and structuring key information about QSAR models and their predictions. In the first one, the QSAR Model Reporting Format (QMRF), it 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.

4 Users' Requirements

4.1 Research Institutions' requirements

QSAR is a very attractive field that involves large part of research activity in chemistry and biology. Both academic and private research institution are interested in this area. The reasons are about the potentiality of QSAR models: to find a QSAR model means to acquire a deeper understanding of all the processes that are related with chemical compounds and of course get the access to an enormous quantity of information about properties and effects even for unknown molecules.

Unfortunately, first studies suggest that a real model that connects each molecular compound with its activity and with its physical and chemical properties is just a dream. For the enormous number of variables a complete QSAR requires a so complicated mathematical model that at the moment is out of the capacity of human knowledge.

Besides this strong statement, the practical use of the QSAR analysis is restricted to only some characteristics to predict and only for a specific class of molecules. From this point of view the problem seems much more affordable. In fact the research field is full of QSAR models that predict only some particular characteristics.

To build a QSAR model the starting point is to select a list of descriptors that describe the important properties of class of chemical compounds. Then several machine learning techniques or statistical tools can be applied to extract information and a deductive model that also work for unseen molecules. The first problem of research is at the beginning of this process: which descriptors to use?

In the literature there are thousands of descriptors that have been used during all the years of research in this field. But it is impossible to use all the descriptors available in literature to build a QSAR model. There are two principal reasons. The first one is that it is computational heavy to calculate all the possible descriptors for all the chemical compounds taken in exam for a particular model. Second, and most important, the machine learning techniques and the statistical tool have difficulties that increases exponentially with the dimensionality of the data and so with the number of descriptors. These difficulties are caused from the presence of redundant descriptors or in particular from those descriptors that do not give useful information to predict a specific activity.

So the first need of the research community is the reduction of the number of descriptors to use in the developing of QSAR models. This task resulted to be particularly difficult because a specific descriptor can be much useful to predict a certain activity and useless in predicting another. Other descriptors can be useful for a particular set of molecules while useless for another set.

The evidence suggests that such minimization of the number of descriptors is possible only after that the activity to predict and class of compounds have been chosen. This means that the first part of the process devoted to build a QSAR model must deal with the minimization of descriptors: this is not an easy process because requires very specific knowledge. And even when specific knowledge is available there can be some useful descriptors that can be cut of from the analysis. This can be the case of topological descriptors that can appear a priori uncorrelated with a certain activity but can also give useful information.

Another big problem of research in QSAR field is about sharing knowledge. In this field there is a large part of knowledge that is private. This is due to the economical importance that some QSAR models have for pharmacological companies, and in general for those companies that lead a private research looking for chemical compound with innovative properties.

The problem of sharing knowledge is a big one: it brakes the research. Many researches institutes can work on the same model but they can not benefit of the results of the others.

This lack of sharing knowledge can involve both the entire QSAR model or only the calculation of descriptors: the situation is complicated in both cases. Many papers show that good results have been obtained with QSAR approach but they do not reveal the model used: so both the descriptors used and machine learning techniques are unknowns. This kind of results has no utility for another research institute that is interested to performs a similar study.

Much often scientific papers describe accurately the QSAR model and the descriptors used but they do not publish the source code of descriptors. This is another big problem in research community because descriptors are not standardized. For each descriptor there are tens of different variations and each variation can be implemented with tens of small variations. Also less important variations such as changing the type of a variable can modify the final value of the descriptor. So, until descriptors source code is not published it is impossible to compare the results obtained between two different research institutions. For example some researchers can get bad result from the same model presented with good results from other researchers: the difference can be caused by different versions of the chosen descriptors.

Some researchers can not publish the source code because they used for the calculation some professional non-free programs like Codessa and Dragon. This kind of software can be very expensive and this means that institutions with less found are cut off from research in some areas of QSAR.

To summarize this part, the principal requirements of research institutions are two: minimization of the number of descriptors to use and the sharing of knowledge.

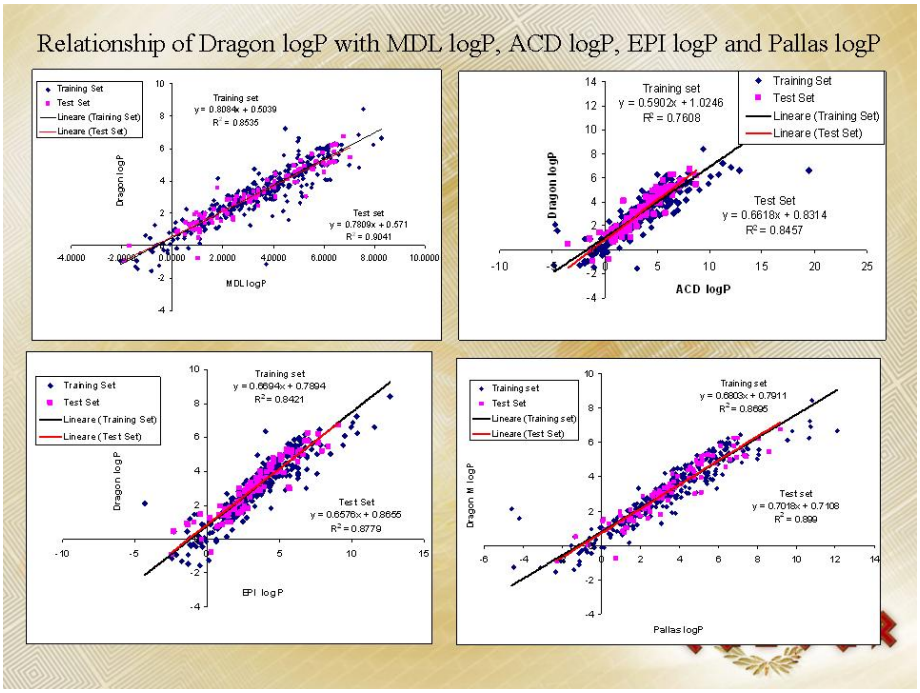


Figure19: how different software (Dragon, MDL, ACD, EPI, Pallas, see section 4.2) gets different results when calculating the same descriptor, in this case logP. From CAESAR project[2]

4.2 Firms' requirements

Since the REACH entered into force, manufacturers and importers of the chemical substances that are produced or imported in quantities of one tonne or above per year have to submit a registration dossier to the ECHA. Chemical firms have to demonstrate that risk coming from the substances they deal with are limited and can be controlled, because the registration is the necessary condition for the commercialization of products.

The innovative aspect of REACH, that is changing firms' approach to the risk assessment of their chemicals, is the acceptance of alternative methods (such as QSARs, in-vitro studies and chemical grouping) for regulatory purposes. In addition, reducing animal testing is one of the main objectives of REACH and this new regulation foresees the use of QSARs when testing does not appear necessary because the same information can be obtained by other means. As a consequence, interest in QSAR models is increasing more and more also among firms, so it is necessary to consider them as one of the most important stakeholder of Vichem project. Even if all chemical firms need to gain complete and clear information about REACH provisions, this necessity is more problematic for Small and Medium Enterprises (SMEs), which do not have the adequate know-how and economical resources to face the new regulatory framework.

The needs of these stakeholders may be ascribable to two different aspects. The first one is the need of obtaining knowledge about REACH provisions that is structured, comprehensive but not wasteful, and easy to be applied in practice. Currently a lot of information about REACH is obtainable by surfing the net, but it does not fit the features just mentioned and it may create confusion and misunderstandings. The second need is still to gain complete and well organized

[2] For further information see: http://www.caesar-project.eu/

information, but the focus is on those aspects present in REACH regulation, that refer to the use of QSARs for regulatory purposes. In this case there is the opposite problem, because this kind of information is poor and difficult to find out.

Both of these needs can be translated into similar firms' requirements, that are technical. The final requirements may be exhaustive guidance tools, which do not directly contain the overall knowledge related to REACH in general, on the one hand, and the regulatory provisions about QSARs, on the other hand, but which provide a simple and brief way to find and manage useful information.

In order to assess the users' requirements a questionnaire was chosen as the most effective method to be used. Its structure was divided into three main sections and each of them was dedicated to a specific topic: REACH, REACH focusing on QSARs, and QSARs models. All the questions had the same purpose to understand what kind of information firms would be interested in to satisfy their needs. In the REACH section, the questions mainly concerned with the usefulness of a guidance on the websites dealing with REACH, of an interactive tool to manage information about workshops and conferences, and of a glossary of recurring terms. The second section aimed at finding out if consulting services or specific summaries of the articles about QSAR regulatory uses may be interesting. The last section investigated the interest of firms on the availability of open source codes for descriptors calculation, of a guidance on free computational tools and specific websites, and of article repository. This questionnaire was send to Federchimica to be widely spread among Italian chemical firms, but unfortunately no replies were obtained.

Thus, the users' requirements were obtained joining the information coming from the analysis of the state of art about the different relevant topics and from the experience of Mario Negri institute's team.

5 State of the Art tools

In the Section 2 we illustrated the background knowledge needed to understand the issues presented in the Section 3. The subject of this section is how stakeholders tackle these issues. First we show a little more in detail some examples of real QSAR models, giving details on how these models deal in practice with the complexity of toxic mechanisms. Then we make an overview on the available software, and in particular we focus on CDK Java Libraries, Open Source Java librarie for Computational Chmistry that will be extremely useful in our solution. The section is concluded by a list of websites where it is possible for firms to collect information on REACH regulation.

5.1 QSAR models in practice

In this section we give some examples of QSAR models developed in the last 10 years. These examples will give a more detailed idea of how these model actually works, and of the range of possible endpoints.

The first model we present is a model for quail dietary toxicity, presented in [30]. The considered dataset refers to quail dietary exposure to different kinds of pesticides (chlorinated compounds, thiazines, organophosphates..), whose toxicity is measured as log(1/LC), where LC stands for LC50-96h, that is the concentration that kills half of the quail population in 96 hours. In other words, the examined endpoint is death.

Atomic orbital graphs are used to describe the chemical structure of the pesticides, instead of standard molecular graphs, to take into account atomic orbitals, such as $1s^1$,$2p^2$, $3d^{10}$; this further information is mathematically expressed through a descriptor called ${}^0X_{cw}$, a particular invariant for these graphs. A linear model is obtained through Monte Carlo optimization of the value of the descriptor and least squares method, leading to an expression of toxicity as:

$$log(1/LC) = C_0 + C_1 \cdot X_{CW}$$

Validation on a test set shows a reasonable agreement with experimental data ($R^2 \approx 0.65$), considered that a single model here is forced to predict activity of pesticides with a number of different (and unknown) toxic mechanisms: models like this one are called global models.

Other global models can be found in [34]. In this paper the endpoint is bioconcentration, that is the process of accumulation of chemicals by aquatic organisms (e.g. fishes) through non-dietary routes, such as cutaneous absorption. Bioconcentration is measured as a bioconcentration factor (BCF), that is the ratio between the concentration of a given chemical in the organism and the steady concentration of the same chemical in the environment. Authors use topological, constitutional and functional groups descriptors to develop a number of global models: linear regression models as the one in the previous example, and non linear models through radial basis function neural networks. We recall that for radial basis function neural networks the final model is:

$$log(BCF) = \sum_j w_i h_j(x) + b,$$

where j are the neurons in the hidden layer, whose activation function is:

$$h_j(x) = exp(-||x - c_j||^2 / r_j^2),$$

that is a Gaussian function centred in c_j with width r_j. All these models show a reasonable predictive power, as R^2 stays between 0.74 and 0.80 for all models.

Anyhow, global models in general cannot achieve really high predictive results, as it is unlikely that such complex phenomena as toxic mechanisms are well described by a unique function of some descriptors, regardless of its complexity. To further improve predictive power, one has to build multi step models, that start with local descriptions of the data space and then put together these local information in a suitable way. This is the idea of the so called hybrid models, or expert systems.

A first example of hybrid models is found once again in [34]. In this approach to expert systems, the underlying hypothesis is that every single global model catches different pieces of information stored in the data; therefore a winning strategy to estimate the endpoint could be to run all the different models, and then to perform an "ensambling"/ "averaging" of the different results.

In the article the authors follow these steps:

1. run the global models;
2. average the results;
3. divide the range of outcomes in some areas, say three areas;
4. build in each area the final predictive model, that will be:
   $$log(BCF) = C_0 + C_1 \cdot operator(y_1, .., y_k)$$
   where y_k is the endpoint estimate by the k-th model, and operator can be min, max, mean.

Another possible strategy is to build expert systems that first divide the information space according to the biological rationale of the studied phenomenon, then build a proper model for each sub-domain and finally link every chemical to the most suitable sub-domain, and hence to the best model.

An example of this procedure can be found in [26], in QSAR applied to acute toxicity for fishes. First, the authors individuate eight different modes of action: base line narcosis or narcosis I, polar narcosis or narcosis II, ester narcosis or narcosis III, oxidative phosphorylation uncoupling, respiratory inhibition, electrophile/proelectrophile reactivity, AChE inhibition and CNS seizure response.

As a second step, every chemical of the considered data set (617 chemicals) is assigned to a single mode of action, through both an experimental procedure and a review of the existing literature. We remember that this implies killing a statistically significant number of fishes for each chemical, measuring until death occurs a number of variables such as heart rate, blood pH, hematocrit and symptoms such as convulsions, spasms, reaction to stimuli, body coloration.

After this splitting procedure, a QSAR linear regression is performed for each class and finally an heuristic expert system that assigns every new chemical considered to its (predicted) mode of action is built. This expert system looks for those molecular fragments that are specific for each mode of action; thus it is implemented as a set of conditional statements that will be used to classify chemicals: for example *"if a chemical has an oxygen attached to an aromatic ring and the number of halogens is less than or equal to 2, then the associated mode of action is narcosis II"*. Specific fragments for each mode of action were identified "a priori" by human knowledge, but they could

have been computationally assessed as well. Validation on a test set gives R^2 =0.91, that is to say an excellent predictive power.

It is important to underline that each mode of action is not related to a single toxic molecular mechanism. For example, CNS seizure agents and respiratory inhibitors can act through a variety of receptors (see [8] and [1]). Moreover, the experimental assessment procedure shows that standard chemical classification is not suited for modes of action. For example, ethers are usually associated to narcosis I mode of action, but the experimental assessment shows that ethers may also act as electrophiles; conversely, compounds from classes that are not supposed to be narcosis I type act in this way.

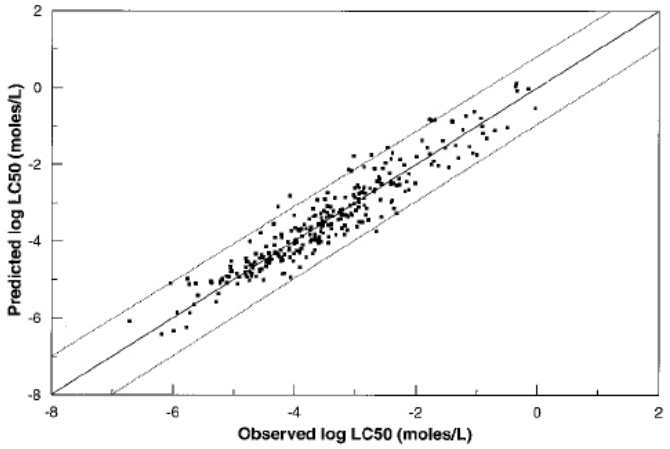


Figure 19: predicted LC50 vs. observed LC50 in [26]. The ideal plot would be with all points on the unity line (central line). The upper and lower bound represent plus or minus one log unit from unity.

Classification based on mode of action analyses therefore seems to be the most appropriated choice, but even this approach is arguable. The first issue is that modes of action are not uniquely defined. The second issue is that modes of action should be considered as continuum, as they are the visible outcome of several cooperating or competitive molecular mechanisms rather than linked to a single one, as we already pointed out.

As a consequence, also chemical classification based clustering may be successfully used in expert system QSAR building. An example is [18]. In this paper, that refers to the same data set of [26], local linear models are built on 13 subsets obtained splitting chemicals according to the E.P.A. classification: hydrocarbons, ethers, alcohols, aldehydes, ketones, acids, nitriles and sulphur compounds, amines, benzenes, phenols, heterocyclics, carbmates and other pesticides, and the remaining chemicals were merged in a residual class.

In this case, the classifier does not choose the most likely mode of action, but tries to identify the most fruitful model, on the basis of an automatic selection of nominal chemical classes. The appropriated class is not selected through standard definition (“human knowledge”) nor heuristics conditional statement, but once again using constitutional descriptors.

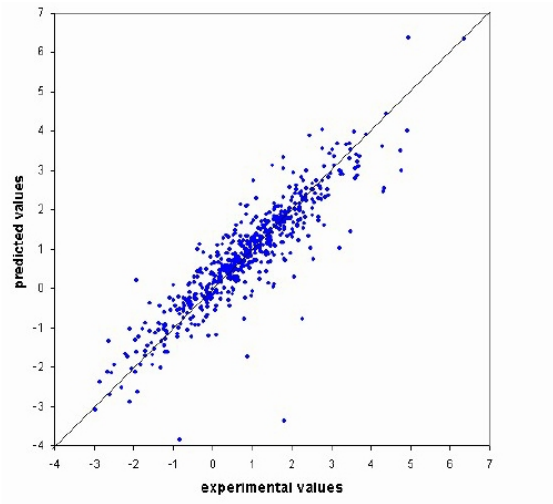


Figure 21: Predicted values vs. experimental values from expert system in [18]

In both [26] and [18], the expert systems show a remarkable improvement of predicting power with respect to the single global models. In addition to this noticeable feature, expert systems are attractive also because they lead to a reduction of the needed computational time, as the time consuming algorithms that tune the parameters of the models are applied to small set of data (the sub-domains) rather than to the whole dataset.

5.2 Available software

5.2.1 Computational tools for applying QSARs

A wide variety of publicly available and commercial computational tools have been developed that are suitable for the development and application of QSARs. Such tools include methods for range of QSAR-related tasks, including data management and data mining, descriptor generation, molecular similarity analysis, analogue searching and hazard assessment.

Among these tools, QSAR-based expert systems enable predictions of chemical toxicity to be obtained directly from chemical structure. All are built upon some experimental toxicity data with rules derived from the data[2,3].

A list of the most used tools/databases publicly available will be shown below. In annex I, a list of reference website is also provided.

Ambit is freely available software for data management and QSAR applications, including searchable databases and tools for grouping and applicability domain assessment. The *AMBIT* database stores chemical structures, their identifiers such as CAS, INChI numbers, attributes such as molecular descriptors, experimental data together with test descriptions, and literature references. The database can also store QSAR models. In addition, the software can generate a suite of 2D and 3D molecular descriptors.

Toxtree, developed by Ideaconsult Ltd under contract to ECB, is a freely available application which is able to estimate different types of toxic hazard by applying structural rules. Currently, plug-ins are available for applying the following rulebases: a) the Cramer classification scheme for TTC (Threshold of Toxicological Concern) estimation; b) the Verhaar scheme for predicting the mode of toxic action in aquatic species; c) decision trees for estimating skin and eye irritation and corrosion potential, based on the BfR rules, and d) the Benigni-Bossa rulebase for mutagenicity and carcinogenicity.

The *JRC QSAR Model Database*, which is currently under development will be a searchable tool for linking chemicals of interest to a collection of robust summaries of (Q)SAR models. The summaries are being compiled by using a standard (Q)SAR Model Reporting Format (QMRF). A database with a web-based interface will be implemented to allow on-line access to the JRC QSAR Model Database.

The *EPI (Estimation Program Interface) Suite* program integrates a number of estimation models for the prediction of environmental and physical/chemical properties in one convenient interface.

These models include KowWin (for estimating log Kow), AopWin (for predicting gas-phase reaction rates), HenryWin (for Henry's Law constant), MPBPVP (for predicting melting point, boiling point, and vapour pressure), WsKow (for estimating water solubility and log Kow), Hydro (for estimating hydrolysis rate constants for specific organic classes), DermWin (for estimating the dermal permeability coefficient (Kp)), ECOSAR (described above) and BCFWin (for estimating the bioconcentration factor). *EPI Suite* is freely available from the US-EPA website.

Referring to commercially available tools other two examples can be mentioned.

TOPKAT is a statistical system developed by Accelrys, Inc consisting of a suite of QSAR models for a range of different endpoints. There are currently 16 modules for the following endpoints: aerobic biodegradability, Ames mutagenicity, Daphnia magna EC50, developmental toxicity, fathead minnow LC50, FDA rodent carcinogenicity, NTP rodent carcinogenicity ocular irritancy, logKow, rabbit skin irritancy, rat chronic LOAEL, rat inhalation toxicity LC50, rat Maximum Tolerated Dose (MTD), rat oral LD50, skin sensitisation, and Weight of Evidence rodent carcinogenicity.

TOPKAT models are typically based on the analysis of large datasets of toxicological information derived from the literature. The molecular descriptors used include structural (e.g. molecular bulk, shape, symmetry), topological and electrotopological indices. The QSARs are developed by regression analysis for continuous endpoints and by discriminant analysis for categorical data. It estimates the confidence in the prediction by applying the patented Optimal Predictive Space (OPS) validation method.

TerraQSAR™ is a collection of computation programs for the prediction of biological effects and physico-chemical properties of organic compounds. The available models developed using a probabilistic neural network (PNN) methodology include: DM 24 hr EC50 for Daphnia magna, E2-RBA estrogen receptor binding affinity (RBA), FHM 96-h LC50 for fathead minnow (Pimephales promelas), log P octanol/water partition coefficient, etc.

A number of commercial software programmes have been developed for the calculation of the molecular descriptor. Some of them are provided below. The list of reference website is in Annex I.

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.

Last release: 3.0.

OASIS has been developed for the calculation of steric, electronic and hydrophobic descriptors.

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.

5.2.2 Chemistry Development Kit (CDK)

The Chemistry Development Kit (CDK) is a Java library for structural chemo- and bioinformatics, used in over 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 access[3].

The CDK library contains a huge number of various classes and it is impossible to describe all of them. 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.

[3] See: http://apps.sourceforge.net/mediawiki/cdk

- **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 setter and getter methods for the partial charge, the hydrogen count, the stereo parity and the 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.

5.3 Overview of interesting websites on REACH

A lot of different websites deal with the main aspects of REACH legislation. They can be divided into two main categories, such as the websites that are completely dedicated to the new regulation system, on the one hand, and the websites which deal with REACH only within one or few sections.

5.3.1 Websites totally dedicated to REACH

There are a lot of websites that are totally dedicated to REACH and some of them will briefly described below. Among these websites there are ones where general information are provided and some others that are made 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, available from the link above.

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.).

5.3.2 Websites partly dedicated to REACH

Some interesting websites have only one or some sections related to the new regulation for chemical substances and 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 is made by 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.

Here detailed information, documents and manuals on the REACH regulation are provided.

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 made by 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 made by 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 above all that companies outside the EU will need to comply with REACH. This website is made by RoHS International.

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

IOM (Institute of Occupational Medicine) Consulting offers several services and information to get firms to find a way through these complex regulations. 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.

6 Conclusion

We shortly reviewed the area of computational chemistry, and in particular QSAR. A major event in the area has been the activation of the new REACH legislation, which foresees the exploitation of such mathematical tools for the risk assessment of produced chemicals.

This is further enhancing the importance of these techniques, but also generating a number of issues, that we stated as a "non-free/non-accessible knowledge" problem. We illustrated this phenomenon in Section 4, and we pointed out that it involves both the research world and the chemical companies.

Going back to a scientific point of view, it remains a scientific challenge to evaluate the biological effects of compounds using only in silico tools, and more research is needed. No public model under REACH still exists; we also hope to be an impulse in this direction.

7 Bibliography

[1] D.C. Anthony and D.G.Graham, Toxic responses of the nervous system, in *Casarett &Doull's Toxicology: The Basic Science of Poisons* 4th ed., by M.O.Amdur, J. Doull and C. D. Klaasen, Mc Graw-Hill New York, NY U.S.A. 407-429, 1993

[2] R.S. Boethling and D. Mackay, Eds., *Handbook of Property Estimation Methods for Chemicals: Environmental and Health Sciences*, Lewis Publishers, Boca Raton FL, 2000.

[3] Casarett, Doull, Klaasen, Casarett & Doull's *Tossicologia – I fondamenti dell'azione delle sostanze tossiche*, EMSI, Roma, 2000.

[4] Crane M, Newman MC. 2000. *What level of effect is a no observed effect?* Environ Toxicol Chem 19:516–519.

[5] J. C. Dearden, M. D. Barratt, R. Benigni, D. W. Bristol, R. D. Combes, M. T. D. Cronin, P. M. Judson, M. P. Payne, A. M. Richard , M. Tichy, A. P. Worth, J. J. Yourick, The development and validation of expert systems for predicting toxicity. *The report and recommendations of an ECVAM/ECB workshop (ECVAM workshop 24)*, *ATLA* **25**, 223-252, 1997.

[6] J. Dearden, In silico prediction of drug toxicity. *Journal of Computer-Aided Molecular Design* 17, 119-127, 2003.

[7] ECHA, *Guidance on information requirements and chemical safety assessment*, Chapter R.6 QSARs and grouping of chemicals, 9-66, 2008.

[8] D.J. Ecobichon, Toxic effects of pesticides, in *Casarett &Doull's Toxicology: The Basic Science of Poisons* 4th ed., by M.O.Amdur, J. Doull and C. D. Klaasen, Mc Graw-Hill New York, NY U.S.A. 565-572, 1993

[9] Enterprise & Industry Directorate General, Environment Directorate General, *Reach in brief*, *European Commission*, December 2006

[10] G. Gini, and A. Katritzky, (Eds.) *Predictive Toxicology of Chemicals: Experiences and Impact of Artificial Intelligence Tools*, AAAI Press, Menlo Park, California, 1999.

[11] L. Goldberg (Eds.), *Structure–Activity Correlation as a Predictive Tool in Toxicology*, Hemisphere, New York, 1983.

[12] A. Goldestein, L. Aronow, S. M. Kalman, *Principles of Drug Action*, John Wiley & Sons. Inc., New York, 1974.

[13] C. Hansch and A. Leo, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington D.C., 1995.

[14] Hansch, C.; Malony, P. P.; Fujita, T. and Muir, R. M., *Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants with partition coefficents. Nature,* 194, 178–180. 1962

[15] Helma,C. and Kramer,S. (2003) A survey of the predictive toxicology challenge. *Bioinformatics*, 19, 1179–1182.

[16] W. Karcher, and J. Devillers, Eds., *Practical Applications of Quantitative Structure-Activity Relationships (QSARs) in Environmental Chemistry and Toxicology*, Kluwer, Dordrecht, 1990.

[17] Kazius, J., McGuire, R., and Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem. 48*, 312-320. 2005

[18] C. König, G. Gini, M.Craciun and E.Benfenati, Multi-class classifier from a combination of local experts: toward distributed computation for real problem classifiers, *International Journal of Pattern Recognitionand Artificial Intelligence*, 18:5 , 801-817, 2004

[19] D.J. Livingstone, *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*, Oxford University Press, Oxford, 1995.

[20] D.J. Livingstone, The characterization of chemical structures using molecular properties: a survey, *J. Chem. Info. Comput. Sci.*, 2000.

[21] W.J. Lyman, W.F. Reehl and D.H. Rosenblatt (Eds.), *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*, American Chemical Society, Washington, DC, 1982.

[22] www.moleculardescriptors.eu

[23] www.qsar.it

[24] Richard, A. M., Gold, L. S., and Nicklaus, M. C. (2006) Chemical structure indexing of toxicity data on the internet: Moving towards a flat world. *Curr. Opin. Drug Disco*V*ery De*V. *9* (3), 314-325.

[25] Regulation (EC) No 1907/2006 of the European Parliament and the Council of 18 December 2006.

[26] C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister and R. A. Drummond, Predicting modes of action from chemical structure: acute toxicity in the fathead minnow (Pimephales promelas), *Environmental Toxicology and Chemistry*, 16:5 948-967, 1997

[27] T.W. Schultz, M.T.D. Cronin and J.D. Walker, Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective, *J. Mol. Struct*. (Theochem), 2003.

[28] Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* 2003, *43*, 493–500.

[29] R. Todeschini and Vivianna Consonni, *Handbook of Molecular Descriptors*, Wiley-VCH, 2000

[30] A.A.Toropov and E.Benfenati, QSAR model of quail dietary toxicity based on the graph of atomic orbitals, *Bioorganic &Medicinal Chemistry Letters*, 16, 1941-1943, 2006.

[31] Van der Hoeven N. 1998. Power analysis for the NOEC: What is the probability of detecting small toxic effects on three different species using the appropriate standardized test protocols? *Ecotoxicology* 7:355–361.

[32] Weininger, D.; Weininger, A.; Weininger, J. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* 1989, *29*, 97-101.

[33] Yang, C., Richard, A. M., and Cross, K. P. (2006) The art of data mining the minefields of toxicity databases to link chemistry to biology. *Curr. Comput.-Aided Drug Des. 2* (2), 135-150.

[34] C. Zhao, E. Boriani, A. Chana, A.Roncaglioni and E.Benfenati, A New Hybrid QSAR Model for the prediction of bioconcentration factor, *Chemosphere*, 73, 1701-1707, 2008

ANNEX

list of software, databases, websites on QSARs and REACH

Here it will be presented a list of currently available software with their reference websites.

Accord for Excel (Accelrys Inc., San Diego, CA, USA)

http://www.accelrys.com/products/accord

ADAPT (Prof. P.C. Jurs, PennState University, University Park, PA 16802, USA)

http://research.chem.psu.edu/pcjgroup/adapt.html

CODESSA (Semichem Inc. – 7204 Mullen,Shawnee, KS 66216, USA)

http://www.semichem.com

DRAGON (Talete srl, via Pisani 13, 20124 Milano, Italy)

http://www.talete.mi.it

GRIN/GRID (Molecular Discovery Ltd. – West Way House,Elms Parade, Oxford OX2 9LL,UK)

No indication are available about the reference website.

HYBOT-PLUS (Prof. O. Raevsky – Russian Academy of science, IPAC)

http://molpro.ipac.ac.ru/hybot.html

MOLCONN-Z (Prof. L.H. Hall – 2 Davis Street, Quincy, MA 02170, USA)

http://molpro.ipac.ac.ru/hybot.html
http://www.eslc.vabiotech.com/molconn/manuals/310s/preface1.html

OASIS (Laboratory of Mathematical Chemistry. Prof. O. Mekenyan – Bourgas University, 8010 Bourgas, Bulgaria)

http://www.oasis-lmc.org

POLLY (Prof. S. Basak – University of Minnesota, 5013 Miller Trunk Highway, Duluth, MN 55811, USA)

No indication are available about the reference website.

SYBYL/QSAR (Tripos Inc. – 1699 South Hanley Rd.,St. Louis, MO 63144-2913, USA)

http://www.serc.iisc.ernet.in/broadcast_messages/msg02517.html

TSAR (Accelrys Inc., San Diego, CA, USA (formerly Oxford Molecular Ltd, UK))

http://www.accelrys.com

A list of current tools/databases, either publicly or commercially available, will be shown below with their reference websites.

FREE AVAILABLE TOOLS

Ambit

http://ambit.acad.bg

Danish QSAR database

http://ecbqsar.jrc.it

Toxtree

http://ecb.jrc.it/qsar/qsar-tools

Toxmatch

http://ecb.jrc.it/qsar/qsar-tools

JRC QSAR Model Database

http://qsardb.jrc.it

DART

http://ecb.jrc.it/qsar/qsar-tools

OECD QSAR Application Toolbox

http://oecd-toolbox-march-2008.software.informer.com/

AIM

http://www.epa.gov/oppt/sf/tools/methods.htm

OncoLogic®

http://www.epa.gov/oppt/newchems/tools/oncologic.htm

ECOSAR

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

EPI Suite

http://www.epa.gov/oppt/exposure/pubs/episuite.htm

COMMERCIALLY AVAILABLE TOOLS

Leadscope®

http://www.leadscope.com

Derek

http://www.lhasalimited.org

HazardExpert

http://www.compudrug.com

TOPKAT

http://www.accelrys.com

The CASE family of methods

http://www.multicase.com

TIMES

http://www.multicase.com

TerraQSAR™

http://www.terrabase-inc.com