International Journal on
ARTIFICIAL INTELLIGENCE TOOLS
Architectures, Languages, Algorithms

Volume 16 • Number 2 • April 2007

E-Modelling: Foundations and Cases for Applying AI to Life Sciences

G. Gini and E. Benfenati

World Scientific
NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

International Journal on Artificial Intelligence Tools
Vol. 16, No. 2 (2007) 243–268

World Scientific
www.worldscientific.com

E-MODELLING: FOUNDATIONS AND CASES FOR APPLYING AI TO LIFE SCIENCES

GIUSEPPINA GINI

Dipartimento di Elettronica e Informazione (DEI), Politecnico di Milano
Piazza Leonardo da Vinci, Milano, Italy

EMILIO BENFENATI

Istituto di Ricerche Farmacologiche "Mario Negri", Milano, Italy

Life sciences, and biology in particular, are heavily impacted by the development of methods for data collection and data analysis. Taking advantage of the availability of data, modeling biological effects is becoming more and more popular and relevant in life sciences, due to the diffuse and wide use of information technology (IT) tools. IT is increasing the availability of models, the horizon and complexity of modeling activities, reaching new targets, and boosting "virtuality". Modeling, as any inductive activity, originates from two needs: to predict future outcomes using previous experience, and to explain observations, or in other terms to infer knowledge.

In this paper we analyze criteria, problems, possibilities and advancements which indicate that the time for e-modeling in biology and other life sciences is ripe. To do this, we take as example a particular field of biological sciences, the computational toxicity (CT) of chemicals, and its usual QSAR (Quantitative Structure Activity Relationships) approach. The open possibilities are evaluated, including a reshaping of the interface between toxicology, chemistry and computer science. The epistemological problem about "what models are" is approached in a pragmatic sense.

Keywords: Modeling; biological data; prediction.

1. Introduction

Intelligent systems, both automatic or for decision support, become more and more the product of machine intelligence and data mining technologies. They are broadening traditional computer science to include some aspects of mathematics, statistics, and social sciences.

Consider that in the last several years dramatic advancements in life sciences have been possible due to the massive use of computing technologies, allowing fast computation and achieving potentialities which appeared remote in the not too distant past (such as in the 1980s). Thus, the basic in-vivo and in-vitro experiments are more and more substituted by the in-silico experiments, where virtual entities are manipulated.

Moreover "virtuality" is allowing a new integration of different scientific domains, as the case of bioinformatics demonstrates. We argue that an important step in this direction is played by the changing mind from purely statistics usage of data to the data mining and machine learning view of the recent years.[1,2]

Pioneering work of artificial intelligence applied to chemistry, biology, medical diagnosis, has produced important experiences, such as DENDRAL[3, 4] to name the first system aimed at discovering knowledge in chemistry. More recently, artificial neural networks (ANN), fuzzy logic (FL), machine learning (ML), allowed new possibilities in modeling from data and from knowledge, even taking care of the ignorance and the approximation.

Thus, changes have impacted the research in life sciences in three directions:

- It is easier and faster to make more complex models;
- More researchers have possibilities to perform studies using these tools;
- The introduction of a virtual level of the research is reshaping the research, allowing "in silico" studies.

These changes require a broad discussion on the requirements, criteria, limits and possibilities of these new avenues of the research. The topic is complex, also because the "in silico" research is like a clone of the reality, and thus it changes according to the real field it is applied to. Thus, different aspects are involved for the different domains. We will discuss here this problem, taking as example the new area of computational toxicology (CT),[5] an area of grooving interest in chemometrics.[a]

Chemometrics, the information aspects of chemistry, encompasses the basic steps of:

- extracting information from chemical data = data analysis;
- making chemical data have information = experimental design;
- investigating complicated relationships = modelling.

Basic Chemometrics strategies evolved from statistical experimental design, which gives the ways to generate a set of examples, reduce the range of attribute dimensions, and transform data to simplify the response function by linearizing, stabilizing the variance, and making the distribution more normal. See Fig. 1 for a general view.

One of the most active areas in chemometrics is QSAR (Quantitative Structure-Activity Relationships),[6] developed in the last 40 years to assess the value of drugs, and now proposed as a way to assess general toxicity, and to obtain new knowledge from data. For drug activity and toxicity to a given target, most of the QSAR models are regressions, mainly referring to the dose with the toxic effect in 50% of the animals. Classification systems for QSAR or SAR (Structure-Activity Relationships) refer to regulatory bodies (as National Toxicology Program - NTP, United States Environmental Protection Agency - EPA), that aim to use predictive methods for priority setting and for risk assessment.

The basis of QSAR is in a set of "postulates" as defined from evidence and theory, expressed as follows:

- The molecular structure is responsible of all the activities shown
- Similar compounds have similar biological and physico-chemical properties
- Congenericity: QSAR is applicable only to similar compounds

From this definition of QSAR it is evident that the localness of the model must be preserved, and generalization requires attention.

We are increasingly aware of the need to understand and predict the consequences of chemicals on human health and the environment; this is now studied in ad hoc

[a]http://www.chemometrics.se/

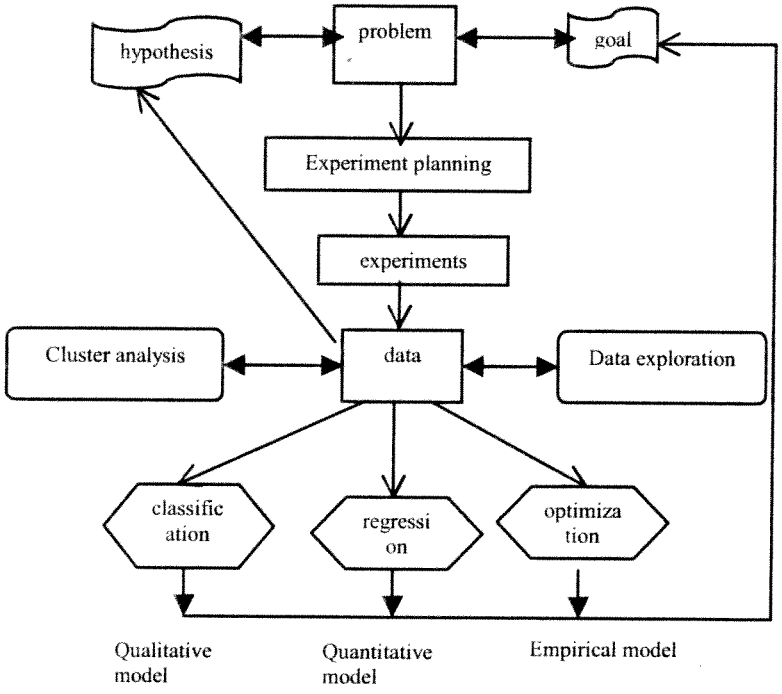

Fig. 1. The chemometrics process.

experiments, which are very expensive, years long, and involve animals. The huge number of compounds makes this especially challenging: there are more than 23 million listed in the Chemical Abstract Service.[b] However, data on toxicity is available only for a small percentage of the industrially produced chemicals. Of the 3,000 chemicals that the US imports or produces at more than one million lbs/yr, an EPA analysis found that 43% of these chemicals have no testing data on basic toxicity and only seven percent have a full set of basic test data.[7]

Many factors and parameters are involved in the toxicity mechanism. There is also a representation problem, namely what is a molecule? Many molecular representations have been proposed, as quantum similarity, spectral properties, topological descriptors, etc., but no unique description can be reached. Every molecular representation sees the molecule as an object, while many properties may be better represented seeing atoms and their position in the molecule. Another important distinction is between 2D representations (molecules are graphs) or 3D representations, obtained after complex physico-chemical calculation.

The final target of predicting toxicology is to work "in silico, not in vivo", so to assess toxicity in a virtual laboratory. This trend is not new in chemistry: computer models and computation are present in any area of analysis and synthesis.

Open challenges now are to collect and organize information on different toxicological data to create a free access data base, as in the OECD priorities This may be useful for the regulatory assessment of pesticides and drugs, and for the labelling of chemicals according to safety regulation. Remember that the new REACH[8] legislation in Europe will limit the use of animals in testing. Thus, predictive models have to be evaluated on the basis of the number of mistakes, of the sign of the wrong classification,

[b]http://www.cas.org/EO/regsys.html

of the extent of the wrong classification, to substitute experiments. This decision will open the issue of the value of models in life sciences.

Modeling, as any inductive activity, originates from two needs: to predict future outcomes using previous experience, and to explain observations, or in other terms to infer knowledge.

Good and acceptable models can (1) explain patterns in data; (2) correctly predict the results of new experiments or observations; and (3) are consistent with other ideas (models, beliefs, and metaphysical commitments).

In the following sections we will explore the modelling problems of QSAR and discuss their relevance in the scenario of intelligent systems.

2. Modelling

Modelling is usually a part of the knowledge process, a broader area which is of interest in philosophy of science as well as in AI.

2.1. *The knowledge process*

In most of the cases the knowledge process follows some phases as below described.[9]

1) Formation of taxonomies. In many cases the first step in the knowledge process is the formation of taxonomies. At this level categories and basic concepts are established. From these "consolidated" concepts new disciplines started in the past. The name taxonomies recalls the categories widely used in biology. But this process is continuously applied. For instance, the *mode of action* concept introduced in aquatic toxicology, referring to mechanisms as narcosis, oxydative phosphorylation uncoupling, etc, is an example of this attempt to categorize the phenomenon.
2) Qualitative laws. Qualitative laws identify relationships between the identities above defined. They can have the form of *if ... then* rules. For instance, if a *nitroso group* is present in the molecule, then carcinogenicity is likely.
3) Quantitative laws. The following step involves definition of quantitative laws. An example is the equation which links aquatic toxicity with *log P*,[c] which is the partition coefficient between octanol and water.

    1) Structural models. Science often goes beyond empirical summaries, and produces structural models, which involves unobserved entities. For instance, in the case of CT, the chemical descriptors are often used to build up models.
    2) Process models. A further step is a model which explains phenomena in terms of mechanisms. This step in not always required or evaluated. Actually, in the case of modern CT most of the efforts have been put on the predictive capability of the model.[10, 11] But this is not the only important aspect, since a model can also be useful to organize knowledge. A good mechanistic model is able to extract

[c]The octanol-water partition coefficient logP is a laboratory-measured property of a substance. It provides a thermodynamic measure of the tendency of the substance to prefer a non-aqueous or oily milieu rather than water (i.e. its hydrophilic/lipophilic balance). The octanol-water partition coefficient is defined as the ratio of the concentration of a chemical in octanol and in water at equilibrium and at a specified temperature. Octanol is an organic solvent that is used as a surrogate for natural organic matter.

If a compound contains ionizable groups, it may exist in solution as a mixture of different ionic forms. The distribution coefficient logD is the ratio of the sum of the concentrations of all species of the compound in octanol to the sum of the concentrations of all species of the compound in water. It is a combination of logP and pKa (the ionization constant) and produces a logP for any pH value.

relevant features which apply for a given domain of compounds (eventually heterogeneous on a chemical point of view). Such a model contributes to knowledge advancements, since it better defines different mechanisms involved in the toxic process. However, such a model is not necessarily capable to be predictive in a strict sense. The common way to imagine a predictive model is a model which predicts toxicity for a new compound; actually, a model can also predict a mechanism, it means that it explores the features and not the instances of a given domain.

2.2. *Models versus reality*

The model is not the true world. The model is no pure theory. The model is something between these two entities. The model is devoted to a specific area of the real world. Thus, it does not include all the real world situations. Linked to this specificity for a local real issue, is the fact that the "theory" itself of the model is not universal.

Often the model for a scientific theory is a deductive system, which uses notions already known in a different science, which has, for some aspects, the same formal structure of the first theory. This gives a psychological advantage to use something accepted in another field.

In his work on "Method" Archimedes describes the 'mechanical method', which is different by the real demonstration. This method allows investigating, but not proving, he said. The ideal purity of the methods was clearly defined by Aristoteles[d] (*Analytica posteriora*, 1, 7): it is not possible to demonstrate a fact going from a gender to another.

Similarly, the model is an approach, which asks for help to a paradigm to show the truth. Actually, a good model is a way of discovery; it cannot give demonstrations, but it is usually illuminating.

Many fundamental theories started from models. Archimedes used models from weight theory to explain hydrostatic principles. Ibn al-Haytam used the model of a bouncing ball to study light reflection. Mechanical models were introduced in the XVII century and used in many fields: light (Huygens, Newton), electricity (Franklin, Volta, Faraday, Maxwell) heath (Boerhaave, Carnot), embriology (Buffon), cellular structure (Schleiden), geology (Hutton), chemistry (Boscovich, Dalton), gas theory (Bernoulli, Herapath), as the only commonly accepted methods. Mathematical models have been used to study moving current (Maxwell) or to predict positron (Dirac), while biology suggested some of the most recent mathematical techniques, like neural networks and genetic algorithms.

The case of CT can be put within the *formal* or *mathematical* models, where a system of equations is taken as a model for a given system. For instance, we can say that fish toxicity is a function of a parameter like the partition of a chemical between octanol and water; then we can speculate about the meaning of such a mathematical equation, reasoning about a biophysical process involved in the phenomenon.

Causal models have been employed to study causal relationship. For instance, electrophylic or nucleophylic behavior in generic chemical reactions of a given chemical has been suggested to play a role in aquatic toxicity.[12]

Functional and *structural models* have been classically used in biology, using inferential processes. They allow extending features from one organism to another, and there are examples in comparative anatomy and in evolutionary studies. In the case of toxicology, they can be used to extrapolate results from one organism to another.[12]

[d]http://plato.stanford.edu/entries/demonstration-medieval/

Other models can be classified as *qualitative* (or *property*) models. For instance, the use of plastic bags filled with triacetin has been proposed as a model for aquatic toxicity: the accumulation of a chemical within the bag can be put in relation with the toxic effect in fish.[12]

The introduction of more powerful IT tools resulted in the introduction of more flexible approaches, integrating different perspectives. The so called "hybrid systems", which integrate different computer programs, allowed a deep use of information of different sources and kind, both numerical and symbolical.

2.3. *The model hypothesis*

The model is an epistemological procedure. Epistemology is that branch of philosophy aimed to study knowledge and discovery processes. The epistemological problems referred within modeling are the following:

- knowledge representation,
- knowledge elaboration,
- knowledge consolidation and check.

The overall process starts from true world; it implicates definition and abstraction of relevant aspects (knowledge representation); then on these aspects we apply theories (knowledge elaboration); these theories generate results, which are evaluated in knowledge consolidation and verification, to come back to the real world.

At the origin of the model there is a core hypothesis, from which the developer has to define variables or predicates used to describe the data or phenomena to be explained, and output representation.

This process is fertile, and has to be done to get deeper insight into our model. However, the basic existing knowledge of the established disciplines involved in the modeling have to be considered, because they will give necessary elements on one hand and will provide requirements, criteria, and constraints on the other. This does not mean that all existing rules have to be rigidly satisfied. Indeed, new theories can break or overtake existing rules. But we have to understand and recognize these rules.

In the case of CT, we assume that the *toxicity is related to the chemical*. This seems an obvious statement, but it deserves a better analysis.

What we assume is that

$$\text{Tox} = f(\text{Chem.}) \tag{1}$$

where Tox is toxicity, f is a mathematical function and Chem. represents the chemical compound.

Such a simplified expression is at the basis of the work of Hansch,[6] for instance, a pioneer in the field of the quantitative structure-activity relationship (QSAR). However, we have to better understand the philosophical implications and limits of the Eq. 1.

1. From the old Paracelsus' assumption we know that the dose makes a toxic compound. Indeed, the effect we observe depends on the amount of the assumed chemical. To simplify the phenomenon description, toxicologists have defined a kind of standardized effect, such as the dose, which produces a given effect (e.g. death in 50% of the cells). For instance, chemical A will give the same toxic effect of chemical B using a dose double of that of chemical B; what changes is the dose, not the effect – note that chemical B in this example is more toxic, since it requires a

lower dose for the same effect. Thus we can compare different chemicals only on the basis of their chemical nature, because we have defined a standard effect.

2. From the previous point, we understand that different chemical will require different toxic doses to produce the same effect. Thus, Eq. 1 has a meaning as a comparison between different compounds.
3. But the toxic effect refers to a cell or organism. Does this have an influence? If we keep into consideration as toxicity the lethal toxic effect, such as the so called LD_{50}, which is the dose which kills 50 animals in an experiment of 100 animals, immediately we see that the same dose on 50 animals produces an effect which is opposite to that on the other 50. The difference is as black and white, because 50 die, and 50 stay alive. This fact has a major meaning: the toxic effect is also dependent on something else, not only the chemical. It depends on the organism, or the cell.
4. From point 3 we see that the eq. 1, which appeared as a deterministic one, can be better considered on a stochastic point of view. Thus, we can continue to use the eq. 1, keeping into account the effect which a chemical can generally produce in a population.
5. It is well known that the chemical effect is mediated by processes, which can be very complicated, in many cases unknown. Thus, what we called Tox in eq. 1 can be described as:

$$\text{Tox} = \text{f}(tox1; tox2; \ldots; toxn) \tag{2}$$

each of these factors *toxn* describing different processes related to toxicity.

6. The chemical part, called Chem. in Eq. 1, is actually something which again can be much more complicated; for instance, biochemical processes can easily transform the original compound in a new compound, more or less toxic than the original one.

These examples show that we should carefully try to "open the box" of our model, recognize its parts, assumptions and domains. Thus, it may happen that a model has inside other "models". In the case of CT, our problem is toxicology (for instance for humans, or for a given ecotoxicological system - aquatic or terrestrial). This is the real world. But we use experimental models (selected animals and conditions), to mimic our real universe to be studied. Then we refer to a chemical model: we hypothesize that chemistry can explain our experimental model. Furthermore, we want to use mathematical rules to describe the chemical processes and properties involved in the toxicity model.

Some of the points discussed above are relative to the chemical part, other to the toxicological part. Indeed, we have first to characterize the chemical and toxicological parameters of the model, and then how they interact. In other words, we have to define semantic and syntactic aspects.

In the present practice, CT models are the results of experiments, knowledge, intuition, demonstration. While pure mathematical models are the result of data analysis. So in principle they can give better results in terms of predictivity (they do not suffer for missing connections) but they can be more obscure to the observer (they do not follow any mechanicistic path). On this point opinions of researchers in life science are still diverse. The practice of the so called mechanicistic approach is usually advocated as a way to understand the model.

Let us discuss more this point. A central message in Ref. 13 was that higher level models are not compared directly with data, but with models of data which are lower down in a hierarchy of models. It is not data but phenomena that theories explain and that are used to test theories. And phenomena are first constructed from the data, using for

instance statistical techniques. On this view, when testing the fit of a model with the world, one does not compare that model with data but with another model, a model of the data. Thus one reasons from a high level model not to make predictions about data, but to make predictions about a model of possible data: the actual data are processed in various ways so as to fit into a model of the data. It is this latter model, and not the data itself, that is used to judge the similarity between the higher level model and the world. Consider here, for example, that the toxicity endpoint is not a datum but the result of fitting experimental data into a curve to get the LD_{50}, for instance.

When the research is guided by broad general principles, as in many areas of physics and engineering, the models often embody these principles. Where such principles are lacking, as in biology and toxicology, the models derive from mathematical techniques.

Galileo, in "*Discorsi e dimostrazioni matematiche intorno a due nuove scienze*" (1638) established that the validity of some of the notions he introduced did not result from empirical direct observations, but from the fact that they were able to act as premises of a deductive process, whose final consequences "seems to coincide with what natural experiments present to our senses". This deductive process was opposite to the inductive Aristotelian procedure to abstract from experiences. Similarly, virtual models can be useful tools, waiting for the experimental validation of the developed theory.

3. The Structure of Information

3.1. *Semantic and syntactic aspects*

The language, the terms, the main aspects of the used scientific domains should be fully understood, in order to warrantee the meaning, the robustness and the validity of the model. AI researchers have been the first to pay attention to this aspect. This has been because the interface between the computer and the human expert was within the field of interest of AI scientists.

In the case of CT of chemicals at least three scientific fields are involved: toxicology, chemistry, and computer science. Actually, the palette of skills is even more different. Indeed, we should distinguish between toxicology and ecotoxicology (and there are different experts for each field, such as carcinogenicity, mutagenicity, reproductive toxicity, terrestrial or aquatic ecotoxicology, etc.). Chemistry involves theoretical chemistry, when we apply molecular theory, or different expertise such as topology or chemical reactivity. For the algorithms, many different techniques can be used, and even though a good statistical basis is always necessary, nowadays we have to include other tools, such as knowledge engineering, data mining, expert systems, fuzzy logic, parallel computing, etc. The modeler should be familiar with the basic statements of each used scientific domain and with the laws he uses.

3.2. *Data processing in computational toxicology (CT)*

The developer of a model in CT has to prepare the data on which the model will operate. Data may be quite sparse, lack certain values, be noisy, or include outliers. Data can be improved manually or using techniques for interpolation, inference, or smoothing. Quite often data are processed before their use in mathematical algorithms. Thus, they can be scaled, centered, normalized. A very important point is that these and all other treatments and use of the data should be done within a more general strategy, involving the steps described in the following.

3.2.1. *Classification*

The values in a model can be expressed as continuous values or as a class label, represented by an integer, or a linguistic entity (for instance medium, high, low). As a matter of fact, they also correspond to the ways to develop mathematical models, using classifiers or regression methods.

The very first simple idea of toxicity is binary: a compound is toxic or not. The binary classification is often applied to carcinogenicity. Indeed, in this case a conservative theory states that even one single molecule can produce the DNA damage, which can generate the cancer. Thus, we are no more speaking about doses, but about the presence of a given effect: again there is a specific theory, behind the carcinogenic/not carcinogenic division.

The regulatory and scientific organisms instead classified chemicals according to their evidence of activity. Four classes are defined by IARC (International Agency for Research on Cancer), namely non carcinogenic, possibly carcinogenic, probably carcinogenic, carcinogenic. We notice here two fundamental points:

1) the target is to classify carcinogenicity to humans: however other models (experimental models on animals) are mainly used, because evidences on humans are indirect and difficult to collect.
2) In this classification there are two components: toxicity (toxic or not) and current human knowledge (known or not, or known at a different degree). The "pure" classes are the first and last one, referring to compounds which are or not carcinogenic to humans. Then, the other chemicals are moved towards these two "poles" according to the experimental studies (in a wide sense, including epidemiological studies). Thus, the other classes are "not well defined", and can be considered as characterized by an intermediate degree of toxicity. Actually, this interpretation may be misleading, if considered directly as a toxicity scale, especially for class three.

Completely different is the classification basis used for ecotoxicology. In this case the regulation defines, for simplicity's sake, three or four classes. In this case there is a real different degree of toxicity, because each ecotoxicological class includes chemicals which do have a certain toxic dose.

However, classification can be done in other ways. Two simple ways are

(1) to split the scale of the toxicity values of the data set in equal intervals, taken as classes;
(2) to split the chemicals in the ordered rank of toxicity, in a way to have balanced classes.

The last of these ways is more likely to produce robust models, and can be recommended in the case of a limited data sets, to avoid the problem of classes poorly represented.

3.2.2. *Continuous values*

Also for carcinogenicity the continuous potency dose has been defined as TD50, which is the dose which produces an increase of 50% of tumors. Also for continuous values we have to take attention about their meaning. For instance, some compounds do not present a TD50 value for different reasons: for some of them it has not been possible to reach the toxic dose, in other cases the experiment statistics were not acceptable, in other cases the chemical exerted an acute toxicity which covered the eventual carcinogenicity (which takes more time to appear).

Carcinogenicity is not the only case where an experimental model cannot be applied. For instance, in aquatic toxicology for some compounds a toxic dose could not be defined. This may happen if a compound is not sufficiently water-soluble, so, the amount of chemical, which is in solution, can be not enough to cause a toxic response.[12]

In the case of ecotoxicology, the use of regression is more common than classifications. This happens for an historical reason, because the pioneer studies in this field have used regression with quite simple chemical descriptors. A second reason is that for many ecotoxicological studies continuous values are more suitable to describe a toxic phenomenon, as in the case of narcosis, in which there is a continuum of increasing toxic doses.

3.3. *Units of measure*

The choice of units can require a careful consideration more than expected, since it can have consequences on the model.

In toxicology there is not a unique opinion about the units for the dose, in moles or in gram. Most of the studies use moles, while regulation refer to grams; to transform a dose in g into a dose in mol we have to divide by the molecular weight (MW). Actually, there are several models which use MW as a descriptor.

Let consider the following equations:

$$y = ax \tag{3}$$

Let's divide y by MW, to get

$$y/MW = y' \tag{4}$$

Now we can ask: is y' a function of MW? The answer is yes. Thus, it may be incorrect to search if the toxicity, expressed in mol, is a function of MW.

However, we could make a very similar argumentation starting from a toxicity value expressed in mol. Indeed, from eq. 4, the dose in mg is simply the dose in mol multiplied by the MW. In this case too toxicity, expressed in mg, is a function of MW (of course in one case the function is direct, in the other indirect).

The answer to this problem has to be searched in toxicology. If toxicity mechanism is due to some specific mechanisms, in which the molecule individually interacts with a specific macromolecule of the organisms, in this case it is correct to use moles. If vice versa we are looking at toxicity phenomena in which the mass of the toxic compound plays a role (as in the case of unspecific toxicity, such as narcosis, in which the toxic compound interacts with the cellular membrane), it can be preferable to use the dose in g.

We have also to mention that the MW in many cases is relatively constant along the series of chemical data sets, compared to the changes in the toxicity range.

Thus, we have seen with this example that terms, definition (dose as g or mol) are not only a matter of convention, but they involve hypothesis behind, which sometimes are not considered by the modeler. The unit definition is related to

(1) the term,
(2) the underlying hypothesis
(3) the correctness of the results.

We note that the example we used for toxicity prediction can be easily transferred to other situations, in which the activity is related to chemicals (i.e. pharmaceutical activity, or physiological behavior).

3.4. *The case of biological data*

There are serious problems for the availability of data both in quantity and in quality. Actually, there are quite large collections of data, however, the data quality is heterogeneous, and it is quite common to find high variability in the toxicity values for the same compound. To cope with the population and test procedure variability, toxicologists introduced defined protocols, defining for instance sex, age, administration scheme, etc. All these parameters define an "endpoint" and reduce variability of the results. Still some variability remains, due to natural variability of the population, to the poor experimental model, to noise. We should preserve the information on such a variability, also as an index to evaluate the quality of our model, which will be affected by such a variability.

At this point we face another problem: since the toxicity value is not crisp as the boiling point, and we can easily find in the literature several values for the same endpoint, we have to decide which one to choose. Consider that a natural variability for biological data is about a factor of 2. There are two major hypotheses:

(1) We can use the median, in order to avoid giving too much relevance to extreme values;
(2) we can use the lowest value, to maintain a conservative approach.

Of course, this decision is not simply a scientific one, and can be taken in different ways for different situations, and depending on the algorithms we will use. Thus, if we would like to build up a conservative ecotoxicological model, valid to protect the most sensitive species, we can prefer to use the lowest values. As an alternative, if we are using an algorithms which keeps into account uncertainty (some fuzzy models), we may prefer to keep all information, including low and median values.

A direct consequence of the variability and complexity of the data is the fact that the model will be better if it is based on a wider and representative data set. However, it is difficult to obtain numerous, reliable data sets of chemicals with their toxicity values, as already stated.

An interesting point is that the basic chemical information is devoted to find out toxicity, but its inverse is not definable: we only have lack of toxicity, but not a beneficial effect. It is as if the scale went from zero in one direction only.

Some residues can reduce toxicity because increase water solubility, or bulky groups can stops the toxic mechanism. But we will not have beneficial properties for this fact, or at least we are not considering them within this study.

3.5. *Explicit and implicit knowledge*

Sometimes the difference between explicit and implicit knowledge is reflected in the difference between expert systems (that infer the truth of facts from existing rules) and data mining systems (that extract relations by induction on data).

Thus, implicit knowledge cannot rely on theory, but has to use data and then try to extract information from them.

The explicit knowledge instead has been clearly explored and codified. It can take the form of rules or simple equations. It has the advantage that it is transparent and easily understandable, because in most of the cases it refers to mechanisms or causal connections. Its major disadvantage is that it is not complete.

Let us take the example of carcinogenicity prediction. Historically, rules from human experts have been used to build up programs to predict carcinogenicity.[14] Even if we

know that a given rule is true, it can be modified by other processes, such as metabolic processes which modify the chemical before it could exert its toxic activity. This problem can be faced adding rules that explain both the growing and the reduction of toxicity.

In the case of ecotoxicology the explicit knowledge is not so detailed as in the case of carcinogenicity. For aquatic toxicology a clear involvement of the partition coefficient, log P, has been found since the first ecotoxicological studies. However this parameter refers to a basic general toxicity, while other mechanisms are responsible for increased toxicity. Several studies addressed this point considering the mode of action (MOA), that are conceptually defined extrapolating from experimental data.[15]

Implicit knowledge is what contained within the data set used to build the model. Data will strongly affect the model. We have to consider both the quality and quantity of the data. The quantity is related to how representative is our data set of a population we want to model. If our target is a well focused, homogeneous population (for instance a well-defined homogeneous group of chemicals), we will need a much more limited number of objects to build up the model. The other point is the quality of the data, used as input of the model. Quality and quantity are also related, because if the quality is reduced by the presence of noise, we will need more objects, for statistical reasons.

4. Models from Chemistry

The model of toxicity makes the hypothesis that toxicity can be explained with the chemical rules. Thus we use a different discipline (chemistry) to explain toxicity, which is defined by toxicology.

4.1. *Representing chemicals*

The first question is: how to consider the chemical information? Our view of a chemical is usually through the eyes of chemists. Unfortunately, this is not quite correct. We should somehow use the point of view of the organism, which is affected by the chemical. This would imply a better knowledge of the mechanism of action of the chemical. As we said, some explicit knowledge of this kind is available, and thus chemical information can be given in a way to reproduce that knowledge. In this approach it is quite common to take into account not only the toxic compound, but also the interacting natural chemical. We can discuss chemical methods considering:

(1) methods which only evaluate the toxic chemical
  (1.1) methods which try to preserve the complex 3D description (for instance through quantum chemical orbitals);
    (1.1.1) single value descriptors: generally on the whole molecule
    (1.1.2) residues. This can be simply based on the presence or not, or there can be a toxicity value associated to them.
  (1.2) methods encoding the information in some salient conceptually simple feature. This includes:
(2) methods which consider the natural chemicals interacting with the toxic compound.
  (2.1) methods relative to binding to proteins;
  (2.2) methods describing chemical reaction.

Another difference is between methods, which describe the global molecule or only consider fragments.

The possible cases are listed in Table 1 and described more in detail below.

1) logP. logP is a classical parameter in ecotoxicology for aquatic toxicity. The basic mechanism it refers to is the adsorption of the chemical in the cell membrane. In this it represent a physico-chemical process, which mimic a biological process. Actually, in many cases logP is calculated, because experimental data are available only for limited number of compounds.
2) Other global descriptors. While logP refers to some knowledge on the toxicity process, for most of the other global descriptors there is a lack of knowledge on the eventual relationship with toxicity phenomena, even though with time some descriptors appear to be more frequently related to toxicity. Here, implicit knowledge can be used. In the last years many studies extended model possibilities. Thus thousands of chemical descriptors have been introduced even though most of them are minor modifications of others. It is quite obvious that using a high number of chemical descriptors the starting hypothesis is that we do not know which one(s) of them is a good one, and we need a valid method to find it out.
3) Pre-defined fragments. Besides global descriptors, another possibility is to use general chemical fragments. In the past quite simple fragments have been used to modulate a chemical behavior: in this case there is a common skeleton and simple substituent groups suppressing or enhancing the property. Hansch, as we said, introduced QSAR models, which apply to specific molecules, with the same skeleton. Outside the group of similar molecules (because sharing the common skeleton) the QSAR model is not applicable. Residues slightly modify the structure and as a consequence the activity. Nowadays there is more the tendency to search for general models, and to use automatic learning approaches, in which residues are seen in a similar ways as holistic chemical descriptors. In general fragments are used to search for their presence or not in the molecule, but it is possible to apply some of the tools already discussed for molecules, and with programs like CODESSA[16] to calculate for fragments the same chemical descriptors as for global molecules.
4) Generated fragments. A further possibility is to use programs, which split the molecules into fragments, created on the basis of the training set of molecules given by the user. In this case there is no previous knowledge of the active fragment, but the system itself will identify which specific fragment is involved in the toxicity phenomenon. Thus, such an approach can be seen as an automatic way to find residues responsible for toxicity (or for reduced toxicity). Compared to the method in the previous point, this approach is more directed to biologically defined fragments, while in the previous approach the residues were chemically defined.
5) Toxic fragments. In this case we use explicit chemical knowledge. They are based on residues (parts of the chemicals). These approaches are by definition incomplete, since our knowledge on mechanism is incomplete. Incomplete does not mean wrong: they contain part of the truth.
6) Quantum similarity. There is another way to extract the chemical information. Fragments refer to parts of the molecule. Chemical descriptors are somehow like the shadows of Plato: we project shadows in walls and from these shadows we derive our knowledge. Similarly, chemical descriptors are like different projections on different planes, but none of them constitute the molecule itself. Quantum chemistry introduced molecular orbitals, which should represent the authentic nature of the molecules. It has been claimed that in this way all information relative to the molecule is included. The question is then how to use this way to parameterize chemicals, since molecular orbitals are complex functions. To cope with this problem, quantum molecular similarity has been introduced. Molecular orbitals are superimposed, and a matrix of similarity is obtained. With this approach each

compound can be compared to another one. In this way the toxic property is not seen as an absolute element but within a scale of molecules. Toxicity itself is no more encoded in theoretical parameters, but is represented by real compounds. According to this philosophy the toxicity can be represented by a typical compound and the toxicity of another compound can be predicted using the overlapping of molecular orbitals. An intrinsic advantage of this approach is that it has the capability to distinguish steric isomers.

7) CoMFA. The CoMFA[17] methodology is a 3D technique where the molecules with known properties, the training set, are suitably aligned in 3D space according to various methodologies (to maximize the steric overlap, to use crystallographic data, employing a steric and electrostatic alignment algorithm) Having arrived at the alignment, charges are then calculated for each molecule to calculate steric and electrostatic fields are by interaction with a probe atom at a series of grid points. One then attempts to correlate these field energy terms with a property of interest. Even if the programs are different from quantum similarity, the overall strategy is similar, and refers to a rich chemical information.
8) Receptor-like interaction. In this case the explicit biochemical knowledge involves mechanisms with some macromolecules. Thus, these models go beyond the simple hypothesis that toxicity is a function of the chemical (eq. 1). In that equation the only information was on the "small" chemical, causing the effect on the organism. Models based on the receptor approach have introduced knowledge on the active site, present in the organism, indicating for instance the size of the receptor niche and the charge distribution. This is more similar to models for drugs.
9) Bioreactivity. This method involves changes of the toxic compounds and their evolution in a product compound, through a mechanism related to toxicity, in which the biomolecules are active as well.

Table 1. The representation of the chemical information in CT.

Chemical description	Explicit biochemical knowledge	high chemical information	Fragment or whole molecule
logP	yes	+	whole
Other molecular descriptors	no	+/++	whole
Predefined fragments	no	+	fragment
Generated fragments	no	+/++	fragment
Toxic fragments	yes	++	fragment
Quantum similarity	no	+++	whole
COMFA	no	+++	whole
Receptor-like interaction	yes	+++	whole
Bioreactivity	yes	+++	whole

We have seen multiple choices to describe chemicals, but all of them are somehow inside fixed boundaries, related to the hypothesis they refer to. In other words they make abstraction of the conceptual model: the toxicity can be found only if the compound has a given residue, or if the mechanism is linked to a given parameter, such as logP, or if the active site is involved or if mutagenicity is related to reactivity with a given nucleotide (in the bioreactivity approach).

A basic distinction between parameters and residues is that the latter is more linked to integers while the first to continuous values: for instance, the number of chlorine can be only integers (including zero). It implies that some algorithms could not be used, such as classical regressions. A great problem is the possible high number of parameters,

compared to the relatively small number of chemicals in the data sets. If this is not solved, over-fitting and wrong models can be obtained.

Another remark is that our efforts go in the sense to predict if a compound is toxic, and not vice versa: to predict a compound with a given toxic activity. In other cases, instead, both directions are interesting; for instance, drug companies want to produce drugs with the wished therapeutic properties. For toxicity prediction we do not require the reversibility of the process, and this makes easier the process. Most of the chemical factors used in CT do not allow identifying the chemical structure which generated a given descriptor, and this in principle should encourage companies in giving their toxicity and descriptors data without attempting their confidential data (the chemical structures).

At a same time a model for CT can be more difficult than many models for drug design, because in the last case the point is how to optimize the structure for the maximum of a single activity, while for CT the point is to characterize the many "maxima" corresponding to toxicity, because there are many causes for toxicity, and most are unknown.

4.2. *Similarity*

There has been a wide discussion on similarity concepts,[18] also in the case of chemicals and their properties. The common procedure is that compound A is compared to compound B, according to a given metric. The real measure for similarity assessment is the organism (or cell, or tissue), which undergoes the toxic effect. The organism "sees" the molecule in a way different from the classical chemical one: for instance metabolism is involved, which modifies the structure of the parental compound. We should not simply speak about chemical similarity, but about chemical similarity on the basis of a given property. Thus, the similarity concept should be put inside a given model, and indeed a model is, in away, a system to use patterns or relationships between similar chemical features and properties. But in light of this, the similarity is generally a relationship between a chemical and a series of other chemicals, where the relationship is mediated by some features.

It is important to note that while in most of the cases the compound is compared to the training set (through the features which denote toxicity in the training set) it is also possible the reverse process: to see how much the training set is similar to the compound. The second approach can give more insights on the suitability of the predictive model. While the first approach can be more indicated using chemical descriptors, the second one can be preferable for residues. Using chemical descriptors there is defined range for each of them in the training set, and we can check if the values of the chemical descriptors of the compound to be evaluated are inside or outside the space of the chemical descriptors of the training set. Vice versa, if we have a set of fragments in the training set, it may be that in the compound to be evaluated they are not present. In this case an evaluation for the new chemical is very critical, since we cannot exclude that a different fragment present in the molecule may give toxicity.

5. Building the Model in the Machine Learning Paradigm

We have seen that the model is target oriented. We have to remember this characteristic of the modeling activities.

We consider here a few examples, which show a questionable use of chemical or toxicological data to build up a model.

- Since some programs require full matrix, without empty cells, it may happen that a mathematician is tempted to fill up empty cells, using routines, which can easily do this. However, in some cases for certain molecular descriptors, it is wrong, to define a value. This is the case for instance of a charge on a given atom. For example, if a chemical is without bromine, we cannot put a value for the partial charge on bromine.
- Another example is the case of uncertain experimental data for toxicity values. For instance, we showed in the case of the carcinogenicity database that the chemicals which, according to Gold herself, have a lower reliability, and modeling is poor, while better results have been obtained for the data of higher statistical reliability.[14] This finding is expected, since the quality of these data is poor. It should be always paid attention to the quality of the data, since we have to be very careful if the model gives better performances than experimental data. In the specific case of carcinogenicity, the modeler should pay attention that some classifications, which are currently used by agencies such as the EPA or IARC, are aimed to identify if a chemical is or not carcinogenic to humans. It means that in principle carcinogenicity classes are two: carcinogenic or not. However, some more classes are present in the EPA and IARC classification, because, besides the "black" and "white" classes, there are some "gray" classes, in which the toxicologists put a chemical on the basis of not complete results. But this means that these classes are due to uncertain results, or, in other words, are contaminated by our ignorance. Thus, the current classification is a combination of toxicity and ignorance, due to the limited results available. In this situation, a modeler should be very careful in the use of this kind of data, because it may happen that he can be finally able to model human ignorance. Of course we can always argue that a compound which is carcinogenic for rat, but for which the evidence for man is insufficient, is in any case more toxic than another chemical, which proved to be not carcinogenic. But we have to be aware of this meaning, especially for chemicals in the class for which inadequate or insufficient data exist also for animals.

Advances in technology have enabled us to collect data from laboratory, observations, and experiments. For the scientist to benefit from these data collecting capabilities, it is becoming clear that semi-automated techniques must be applied to find the useful information in the data. In the last ten years, computer science has proposed new methods for dealing with data and taking advantage of their number. Data mining (DM) consists of extracting interesting knowledge from real, large and complex data sets; and is the core step of a broader process, called knowledge discovery from databases (KDD). The KDD process includes pre-processing and post-processing steps; the first to transform the data to facilitate the application of DM algorithms, the latter to validate and refine discovered knowledge. Mining is the discovery of patterns, associations, anomalies, and statistically significant structures in data. It is a multi-disciplinary field, borrowing and enhancing ideas from diverse areas such as statistics, signal and image processing, mathematical optimization, and pattern recognition.

Machine learning emerged precisely as a way of alleviating difficulties raised by knowledge elicitation from humans in building intelligent systems.

Inductive inference is the process of moving from examples to models; the goal is to learn how to classify objects or situations by analysing a set of instances whose classes are known. Classes are mutually exclusive labels, while instances are typically represented as vectors of attribute values. The input to the learning system is a set of such vectors, whose true class is known, and the output is a mapping from attribute values to classes. The learner will then classify both the given instances and the unseen.

While statistics privileges the numeric output of the classifier, AI in interested in concept formation and plausibility of the model.

However, to be of any use the prediction should be statistically significant. Combining the predictions of a set of classifiers has shown to be an effective way to create composite classifiers that are more accurate than any of the component classifiers. Combining the predictions of a set of component classifiers has shown to yield accuracy higher than the most accurate component on a long variety of supervised classification problems. The research on combined classifiers, or ensemble, will give the next advancement also in CT problems.

There are many methods for combining the predictions given by component classifiers, as stacking (using a gating network) or competitive learning, or systematic ways. Modular systems can be developed on the basis of different models tailored for the different chemical classes or mode of action, as in Ref. 20. There are advantages for such an approach. First, its modularity, which allows flexibility and a better modeling of local features. Further, on a practical point of view, it allows to build up individual models in sequence, which can then be improved if better knowledge are available.

However the chemical based classification presents disadvantages; the hypothesis that for a given chemical class a given toxicological mechanism occurs can be incorrect. Indeed, it has been shown that

1) very similar compounds, of the same chemical class, do not exhibit the same toxicological behavior,[15] and
2) vice versa that the same toxicological mechanism can apply to different chemical classes.

Furthermore, we have to pay attention to the three elements for the robust models.

- data of different nature in the data set
- "natural" variability of the toxicity data (for natural variability between organisms)
- presence of noise or poor experimental model (not uniform practical mistakes, etc) .

The model should react in different ways to these three points:

- In the first case the model should not be forced to model all the data. Indeed the data reflect two different phenomena: the toxic value (our target) and the inappropriate experimental performances for some compounds (because exhibit acute toxicity, or because are unsoluble, for instance). It means, that we are dealing with two different outputs: toxicity of interest (our target) and a second phenomenon. While it is possible to model two different outputs, this is more complicated and in any case it should be clearly stated.
- In the second case the modeller should be aware of the limit of predictability
- In this case the modeller could provide better results than experimental values, because the model can be more powerful.

So far we discussed about the possibilities to take advantage of the advanced AI tools to better manage together some studies in an integrated way. However, nowadays it can be possible to design a completely new strategy of modeling. We said that the very first step in knowledge discovery is indeed taxonomies. But computer programs can efficiently generate taxonomies or classes. This process is called clustering, and thus we can introduce clusters (for instance on descriptors, or on fragments) at the basis of our models, and then generate completely new models, because based on new hypothetical knowledge. This is the case of the generation of fragments which then can be evaluated

for their toxicity, as we discussed above, but can be easily extended for any other cluster, used instead of a predefined class. This enhances the discovery capability of our system.

Another new possibility is to use a new concept of experimental model. So far experimental models have been used to mimic real world: rat is the model for man, eventually, daphnia for an ecosystem and so on. Then, so far, mathematical models have been developed to mimic these experimental models. Actually, with computers we can increase the complexity of our model activities, and put the correct target: man or ecosystem. Thus, we can imagine a model with multiple integrated targets to better define aquatic systems and the effect of chemicals on it. In CT the dose is the value defined by a chemical and a toxicity endpoint. We can define a matrix with a collection of these values for each chemical and each endpoint. We can then define a vector considering the same endpoint and different chemicals, as already discussed. But we can also define a vector of different toxicity values for the same chemical. In this case we have studies generally called activity-activity relationship. Actually, we can combine both information, on same toxic properties on other organisms, and on the chemical features, to obtain the toxicity value of interest, as in Buchanan and Feigenbaum.[3]

6. Cases Discussed

We may discuss now in more details some examples of mathematical modelling done on toxicology data.

6.1. *Aquatic toxicity aggregating local chemical models*

The first case is the study of aquatic toxicity, partially reported in Ref. 20, whose task is the prediction of toxicity on fathead minnow (*Pimephales promelas*), according to toxicity data collected by EPA. The measure for acute toxicity is given by LC50 (96h), which means the lethal concentration for 50% of a population of animals within 96 hours. Their data set contains 568 different compounds.

Moreover, we can consider the EU Directive 92/32EEC which classifies the chemicals in 4 classes as shown in Table 2, according to the LC_{50}:

Table 2. EU classification for fish (Directive 92/32EEC annex VI point 5.1).

LC_{50}	**Dangerous for the environment**
< 1 mg/L	Very toxic to aquatic organisms
1 mg/L – 10 mg/L	Toxic to aquatic organisms
10 mg/L – 100 mg/L	Harmful to aquatic organisms
> 100 mg/L	May cause long-term adverse effects in the aquatic environment

We used the computation tools available in WEKA data-mining workbench created by the Department of Computer Science of the University of Waikato, New Zealand.

The data set has 568 chemicals and 156 descriptors for each.

We built a single linear model using all descriptors and the entire data set. With 10-fold cross validation the results are listed in Table 3. Both the low R^2-value, and the comparatively high error measures indicate that the model does not predict toxicity with a satisfactory accuracy.

The model suffers from the excessive number of features and from the noise they introduce. Therefore we reduced the number of descriptors through wrapper-routine,[21]

which uses results of a learning method to select attributes. The new linear regression model with 20 attributes shows that the points are closer to the ideal prediction (Table 4). But after transformation into EU classes of toxicity the accuracy was still poor, with only 59.3% of instances classified correctly.

Table 3. Accuracy of prediction of log (1/LC50) using a single model with all descriptors.

Model	**Descriptors**	**Evaluation Parameter**	**Value**
Linear Regression	156	R (correlation)	0.740
10-fold cross validation	156	R^2(determination coefficient)	0.548
	156	MAE (Mean Absolute Error)	0.643
	156	MSE(Mean Squared Error)	0.956

Table 4. Accuracy of prediction of log (1/LC50) of a single linear regression model with wrapper selection.

Model	**Descriptors**	**Evaluation Parameter**	**Value**
Linear Regression	20	R	0.838
10-fold cross validation	20	R^2	0.701
	20	MAE	0.565
	20	MSE	0.571

To make a better use of the principle of reduced domain of QSAR, we observed that our regression problem can be formulated as the problem of inducing an approximation function from the feature space to the toxicity value, using pairs of data (structure, toxicity). The domain of this function can be divided in K disjoint sub-domains, so to get K sub-problems. Now we have to approximate K functions with a simpler behavior on their domains instead of one function with a very complicated behavior.

In order to define the sub-domains of the function, we observed that the data set contains completely different compounds, which are toxic but structurally diverse. Therefore we split the data set into 13 groups according to EPA chemical classification. Each group contained between 24 and 74 compounds (Table 5).

Table 5. Subsets of chemical classes.

Chemical classes	**Number of compounds**
Hydrocarbons	26
Ethers	24
Alcohols	60
Aldehydes	44
Ketones	39
Acids	68
Nitriles, Sulfur Compounds	33
Amines	74
Benzenes	33
Phenols	49
Heterocyclics	48
Carbamates, Other pesticides	28
Various classes (pasted)	42

All groups but one contain only one or two chemical classes and a number of subclasses. There remained a few classes with a very small number of compounds (fewer than ten), that were pasted.

For each subset the number of descriptors was reduced with the wrapper method. Then linear regression was done on each set to build 13 models, each for the corresponding data set. The 10-fold cross validation results are summarized in Fig 2.

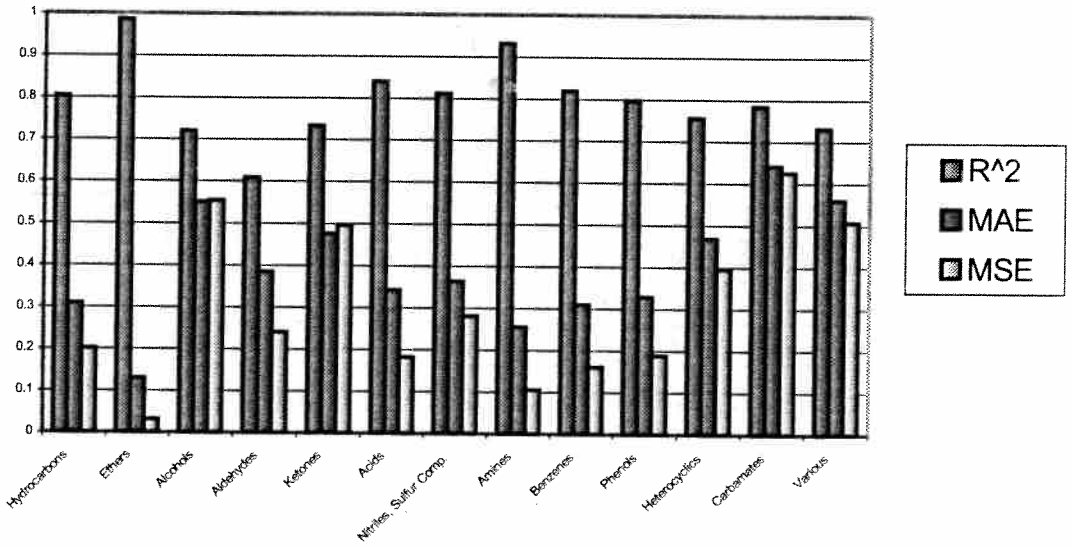

Fig. 2. Results on the 13 models on chemical classes.

The results vary widely. Some models predict very well toxicity as for the ethers and amines. Other models are less satisfactory, as aldehydes. Most models had lower errors than the single model obtained before.

Then we looked at methods to build an integrated model. We discarded ensembles and stacking methods;[22] in fact they combine outputs redundantly. In our case, such linear combination are not useful because it would lose the local influence of our experts, built and adapted on strictly separate areas of the data space. To keep this advantage we decided to use a competitive strategy,[23] thus selecting the appropriate expert for each instance. This choice benefits from the fact that we used the chemical classification to separate the data space. The classification strongly depends on the structure chemicals, i.e. the existence of specific atoms or functional groups.

To develop the combination classifier, i.e. the classifier able to assign a compound to its chemical class, we applied a meta classifier scheme that is able to handle multi-class data sets with two-class classifiers. It was applied to the J48-algorithm, which implements the C4.5 algorithm,[24] using the error correction code to improve the accuracy. The validated model percentage of correctly classified compounds is 85.4.

The output of this classifier was used to select the appropriate expert model for each compound in the data set. These models were set to predict all instances of the entire data set. After combining the sub-models the results improved. R^2 rose about one decimal point and the error values decreased by nearly 30 percent compared with the best result of the single linear model used before. The classification in toxicity classes built after this regression increased as well: 12 percent more instances were correctly classified, which amounts to more than 70 chemicals (see Table 6, Table 7 and Fig. 3).

Table 6. Accuracy of prediction of log (1/LC50) of the combined model.

Model	**Descriptors**	**Evaluation Parameter**	**Value**
Linear Regression	20	R	0.896
10-fold cross validation		R^2	0.802
		MAE	0.417
		MSE	0.390

Table 7. Accuracy of classification of the combined model.

	Number of compounds	**Percentage**
Instances Classified Correctly	409	72.0 %
Instances Classified Incorrectly	159	28.0 %
Total	568	

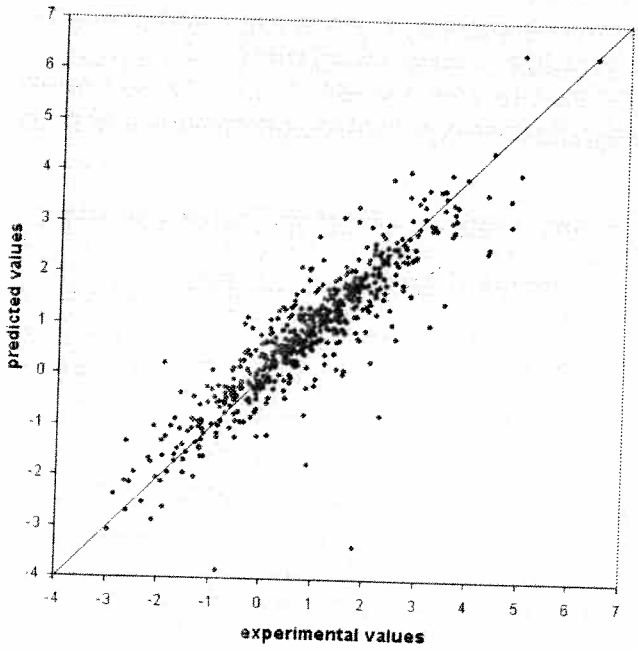

Fig. 3. Prediction dispersion of a combined model of 13 subsets of chemical classes.

To gain a better confidence in the model, we evaluated the performance on external test set. We split the data set in a ratio 80: 20 into a training set (456 cases) and a test set (112 cases), according to the distribution in chemical and toxicological classes. The chemical classifier showed a prediction accuracy of about 88.4% correctly classified instances on the test set and nearly 99.1% on the training set. For the combined regression model we obtained $R^2 = 0.745$, compared to 0.630 for monolithic model with selected attributes.

6.2. *Building a model after clustering data and building local models*

To see whether the supervised subdivision in chemical classes introduces errors, we also used clustering instead of classification.

- we built and trained a self-organized neural network to split the initial training data into a given number k of clusters and we obtained the training sets for the domains.
- for every cluster we built a QSAR model training a feed forward neural network with three layers (input, hidden, output)

From the initial dataset, we sorted it in ascending order according to the output and every 4 and 8 from 10 consecutive compounds were extracted to obtain the test set. So, we had 80% of the chemicals in the training set and 20% for test. Then the values for the training set were scaled in [−1,1]; after the test set was scaled in the same way, using the information from training sets (min and max values of the training sets descriptors).

We built and trained a competitive network (self-organized neural network) to group the training sets into different clusters in the input space. After experiments with different numbers of clusters, it became clear that this strongly influenced the result. Generally speaking, increasing the number of clusters increases the performances for a while; but from a certain point some clusters remain "empty" which indicates that the final number of clusters was reached. In our experiments this point was reached between 12 and 14 clusters, a number very similar to the chemical classes defined by EPA. To have enough molecules in each cluster we made the following experiments with 9 clusters.

The dimension of the vector space in the training data was reduced choosing one from all the descriptors correlated to each other (with $R > 0.9$).

For every cluster, the training data was used to train the supervised neural networks. The final combined model, illustrated in Table 8, has the indicated accuracy.

Table 8. Accuracy of prediction of log (1/LC50) on the external test set.

Evaluation Parameter	Value
R	0.92
MAE	0.392
MSE	0.252

We observe that the results are slightly better than the results of the supervised method presented above. On the other side, we can observe that the models have less interpretability since they are built on clusters that are not necessarily of similar chemicals. A further investigation is needed to see whether the better performances are given from a better choice for chemicals that belong to more classes, or from the bigger number of elements in the clusters. Other similar analysis are reported in a parallel work on neuro fuzzy methods.[25]

6.3. *SAR models (Structure Activity Relationships)*

In the following subsection we consider a different method to build up models useful for ecotoxicity prediction using as a target the regulation of chemicals (based on Directive 92/32/EEC). Lethal concentration for 50% of the animals (LC_{50}) on different species are the endpoints, and extending SAR we approach a multiclass classification problem.

The theory of inducing classification trees has been implemented in C4.5,[24] CART,[26] and their clones.

A classification tree is an empirical rule for predicting the class of an object from values of predictor variables. Common features of classification tree methods are

- Merging: relative to the target variable, non-significant predictor categories are grouped with the significant categories.
- Splitting: a variable to split population is chosen by comparison to all others. The method recursively splits nodes until a stopping rule is triggered[27].
- Stopping: rules to determine how far to extend the splitting of nodes.
- Pruning: branches that add little to the predictive value of the tree are removed.

The classification analysis has been performed with SCAN (Minitab Inc., USA) After a tree has been built, two verification methods have been used: partitioning and cross-validation (leave-one-out). The misclassification matrix, calculated without and with cross-validation, has rows corresponding to the true classes, and columns corresponding to the assigned classes. In a perfect classification, all the off-diagonal elements of the misclassification matrix are zero.

For the set of the pesticides, we defined three toxicity classes for trout, with toxicity ranges from 1 to 0.66, from 0.66 to 0.33 and from 0.33 to 0, in the normalised logarithmic scale. In this way the population of each class is similar.

We obtained an error in validation of 26.3% using the leave-one-out method for validation.

Table 11 shows the real class and the class predicted in validation, and gives also the indication that the classification error is always the real class plus or minus 1, never plus or minus 2.

In a similar way we analyzed only the OrganoPhosphorous subset of compounds, with four toxicity classes:

- with toxicity values (antilog of LC_{50} for trout, scaled between –1 and 1) between –1 and –0.5;
- with toxicity values between –0.5 and 0;
- with toxicity values between 0 and 0.5
- with toxicity values between 0.5 and 1.

Table 11. The predicted classes for the pesticides.

		Predicted class		
		1	**2**	**3**
Real	**1**	16	13	0
	2	11	73	9
	3	0	10	32

The four classes are quite balanced. In leave-one-out the Error Rate was 0.12.

The class 3 (the most represented in the training set) has been correctly assigned. The classification tree for trout, using the OPs, is illustrated in Figure 7. We can see that for a class there may be more than one leaf.

The chemical descriptors used in the tree may be useful to have information on the molecular features involved in the toxic mechanism. For instance, in the illustrated tree some descriptors are topological, as "Average2", "Kier sh1", "Randic 2"and give information on atomic connectivity in the molecule. Other descriptors are constitutional, such as "Number O", the number of oxygen atoms. Others are electrostatic, such as "HA depen" and "PNSA-1P", and reflect characteristics of the charge distribution of the molecule. Finally, some descriptors are geometric, referring to the moment of inertia "Moment o", or to the molecular surface area "Molecula".

However, it is not easy to obtain simple and stable rules. Similar models can be obtained with CART starting from different selected descriptors, and even if the trees perform equally well, the descriptors and the critical values for the nodes may change.

CART is sensitive to the elements and their features. However CART has the advantage of transparency, and it gives indications on the used molecular descriptors which can be valuable in the study of the mechanism responsible of the toxic effect. Even when we split compounds in some chemical classes, each subset includes compounds which very probably produce toxic effects on the basis of different mechanisms.

However these classifiers are suitable to be combined within multiple classifiers systems (MCS) as discussed before.

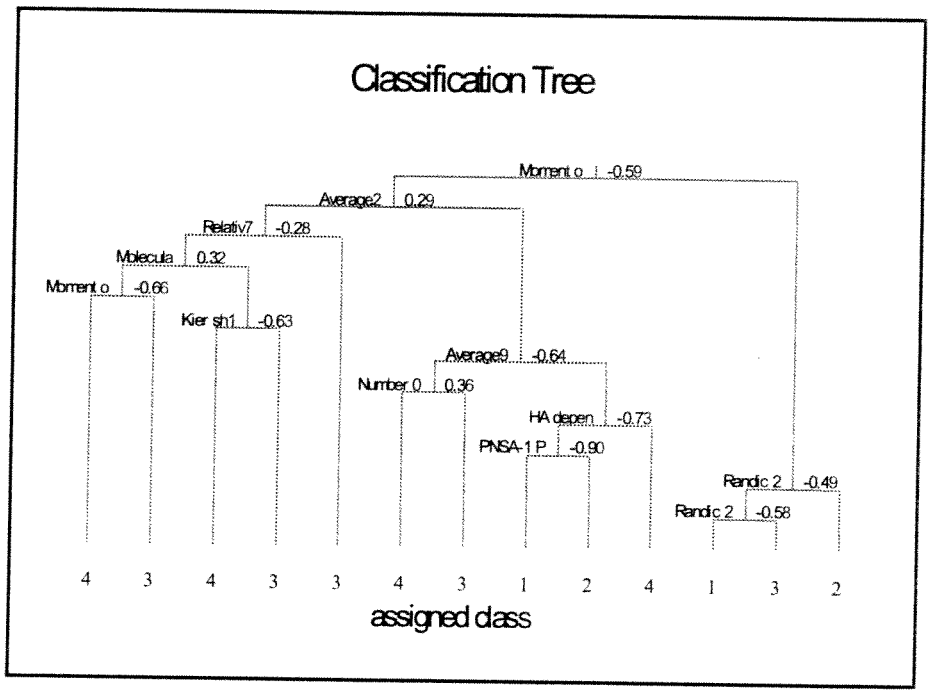

Figure 7. The classification tree for trout toxicity obtained with CART.

7. Conclusions

Models are part of research. In order to take better advantage from them we have to know their limits and capabilities, and understand the rules to which they have to obey. As a matter of fact, research is unformulated, and models represent a help in our discovery process .

We described above the issues related to virtual modeling, involving an extensive use of IT tools. Using the same starting point, of a series of data related to the toxicity behavior of chemical compounds, we can compare the human and the computer approach. The human approach is

1. more qualitative, because often it is related to behavior such as toxic or not, presence of given residues, evidence from experiments done on other species, or done on related compounds. This kind of knowledge is very useful, but it is less suitable to be included in simple quantitative relationships.

2. For the reasons above mentioned, human knowledge is in many cases heterogeneous. This can be a disadvantage when we want to give a clear structure to our knowledge, but it is an advantage as well, because it can obtain knowledge from the numerous instances offered.
3. The human expert takes into consideration instances in a sequential order.
4. (S)he can cope with relationships involving only a few parameters or features (as molecular descriptors, for example).
5. Similarly, (s)he can deal simultaneously with a quite limited number of instances or objects (as molecules).

Vice versa computer models are:

1. More quantitative, and actually in many cases they deal with quantities, even though there are software capable to use symbolic knowledge.
2. In the past, most of the models have been dedicated to a specific focused aspect of the problem, and thus programs have been addressed to well defined toxicity and chemicals. Actually, more modern computer approaches are nowadays capable to keep into consideration knowledge of different sources, and in this they can mimic human capabilities.
3. Computer programs can process knowledge in parallel computations, and can keep into account in an efficient way relationships between variables.
4. Computer programs can work on a number of parameters which is well above any human possibility.
5. Similarly, they can deal with a very high number of instances.

The exploring capabilities offered by data mining, the powerful computational performances of parallel computing, the enormous possibilities of complex mathematical algorithms can be arranged in a way to offer a virtual laboratory, allowing to make toxicological experiments *in silico* in the near future. Man will have to arrange new strategies to exploit these possibilities, but virtuality will not spoil man of his role and mission. Virtuality will have a meaning as far as it will represent a new perspective to get insights in nature and knowledge.

Acknowledgements

We acknowledge the EC projects Demetra and ION, and the Cost 282 Action of the European Science Foundation.

References

1. L. Hunter (ed.), *Artificial Intelligence and MolecularBiology* (Cambridge, MA: MIT Press, 1993).
2. G. C. Gini and A. R. Katritzky (eds.), *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools* (AAAI Press, Menlo Park, CA, USA, 1999).
3. B. G: Buchanan, E. A. Feigenbaum, DENDRAL and Meta-DENDRAL: Their Applications Dimension, *Artificial Intelligence* **11**, (1978) pp.5-24.
4. R. K. Lindsay, B. G. Buchanan, E. A. Feigenbaum, J. Lederberg, J. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation, *Artificial Intelligence* **61** (1993) pp. 209-261.
5. E. Benfenati, N. Piclin, A. Roncaglioni' M. R. Varì. Factors Influencing Predictive Models For Toxicology, *SAR and QSAR in environmental research* **12** (2001), pp. 593-603
6. C. Hansch, D. Hoekman, A. Leo, L. Zhang, P. Li, The expanding role of quantitative structure-activity relationships (QSAR) in toxicology, *Toxicology Letters* **79** (1995), pp. 45-53.

7. Chemical Hazard Data Availability Study from EPA (Environmental Protection Agency), http://www.epa.gov/opptintr/chemtest/hazchem.htm.
8. Press release, http://europa.eu.int/comm/environment/chemicals/whitepaper.htm
9. P. Langley, H. A. Simon, G. L. Bradshaw, J. M. Zytkow, *Scientific Discovery: Computational Explorations of the Creative Process* (Cambridge, Massachusetts: MIT Press, 1987).
10. A. Golbraikh, A Tropsha, Beware of q^2! *Journal of Molecular Graphics and Modelling* **20** (2002) pp. 269-276.
11. D. M. Hawkins, The problem of overfitting. J of Chemical Information and Computer Sciences **44** (2004), pp. 1-12.
12. G. M. Rand, *Fundamentals of aquatic Toxicology; effects, environmental fate and risk assessment*, (CRC Press, 1995).
13. P. Suppes, Models of data, in: *Logic, Methodology and the Philosophy of Science*: Proceedings of the 1960 International Congress, E. Nagel, P. Suppes and A. Tarski (eds) (Stanford University Press, Stanford, CA, 1962) pp. 252-261.
14. G. C. Gini, E. Benfenati, M. Lorenzini, M. Bruschi, P. Grasso, Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network, *J of Chemical Information and Computer Sciences* **39** (1999) pp. 1076-1080.
15. C. L. Russom, S. P. Bradbury, S. J. Broderius, D. E. Hammermeister, R. A. Drummond, Predicting modes of action from chemical structure: Acute toxicity in the Fathead Minnow (Pimephales Promelas), *Environmental Toxicology and Chemistry* **16** (1997) pp. 948- 957
16. A. R. Katritzky, V. S. Lobanov, M. Karelson, *CODESSA: Comprehensive Descriptors for Structural and Statistical Analysis, version 2.2.1. Reference Manual.* University of Florida (Gainesville, Florida, U.S.A, 1994).
17. R. D. Cramer, D.E. Patterson, J.D. Bunce, Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 18 (1988) pp. 5959-67.
18. R. Singh, Reasoning About Molecular Similarity and Properties, Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)
19. M. Berthold, D. J. Hand, *Intelligent Data Analysis – An introduction* (Springer, Berlin, 1999).
20. C. König, G. C. Gini, M. Craciun, E. Benfenati, Multi-class classifier from a combination experts: towards distributed computation real-problem classifiers. *International Journal of Pattern Recognition and Artificial Intelligence* **18**, 5 .(2004) pp. 801-817.
21. R. Kohavi and G. H. John, Wrappers for Feature Subset Selection, *Artificial Intelligence*, **97** (1-2) (1996) pp. 273-324.
22. D. H. Wolpert, Stacked Generalization, *Neural Networks*, **5**,. (1992) pp. 241-259.
23. A. J. Sharkey, *Combining Artificial Neural Nets - Ensemble and Modular Multi-Net Systems* Springer, London, 1999.
24. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, San Mateo, CA, 1993)
25. C-D. Neagu and G. C. Gini, Neuro-Fuzzy knowledge integration applied to toxicity prediction, in *Innovations in Knowledge Engineering*, R. Jain, A. Abraham, C. Faucher, B. Jan van der Zwaag (eds), Advanced Knowledge International Pty Ltd, Australia, 2003.
26. L. Breiman et al., "Classification and Regression Trees (CART)", Wadsworth & Brooks, 1984.
27. J. Mingers, An Empirical Comparison of Pruning Methods for Decision Tree Induction, *Machine Learning*, **4**, (1989) pp. 227-243.