Molecular Descriptors
the free online resource

Molecular descriptors and chemometrics: a powerful combined tool for pharmaceutical, toxicological and environmental problems.

Roberto Todeschini
Milano Chemometrics and QSAR Research Group - Dept. of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1 – 20126 Milano (Italy)

The concept of molecular structure is one the most important concepts in the development of the scientific knowledge of the XX century. As a matter of fact the reasoning based on the molecular structure has been the main engine for the great development of physical chemistry, molecular physics, organic chemistry, quantum chemistry, chemical synthesis, polymer chemistry, medicinal chemistry, etc.
By definition, a system is complex when its behaviour as a whole is not derivable from the properties of its parts: a molecule, together with its imbedded concept of molecular structure, exactly fulfils these conditions. In fact, molecule properties do not depend only on the properties of the component atoms but also on their mutual connections: it is in principle a holistic system, i.e. its emergent properties cannot be derived as the sum of the properties of its parts, but they are inherent to the whole molecule organisation and stability.
As a consequence of its complexity, molecular structure cannot be represented by a unique formal model; several molecular representations can represent the same molecule, depending on the level of the underlying theoretical approach and these representations are often not derivable from each other.

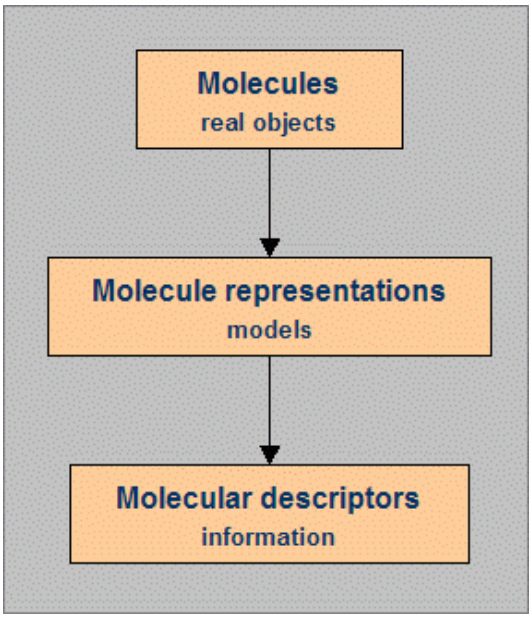

Figure 1

Different molecule representations were proposed, such as the 3-dimensional Euclidean representation, 2-dimensional representations based on the graph theory, or vectorial representations (*fingerprints*) where the frequencies of several molecular fragments are stored. Each representation constitutes a different conceptual model of the molecule and by each model different sources of chemical information become available.

The molecule – thought as a real object – implicitly contains all the chemical information, but only a part of this information can be extracted by experimental measurements. Molecular descriptors are numbers able to extract small pieces of chemical information from the different molecule representations (Figure 1).

The role of the molecular descriptors

In the last decades, several scientific researches have been focussed on studying how to catch and convert – by a theoretical pathway - the information encoded in the molecular structure into one or more numbers – called *molecular descriptors* – used to establish quantitative relationships between structures and properties, biological activities and other experimental properties.

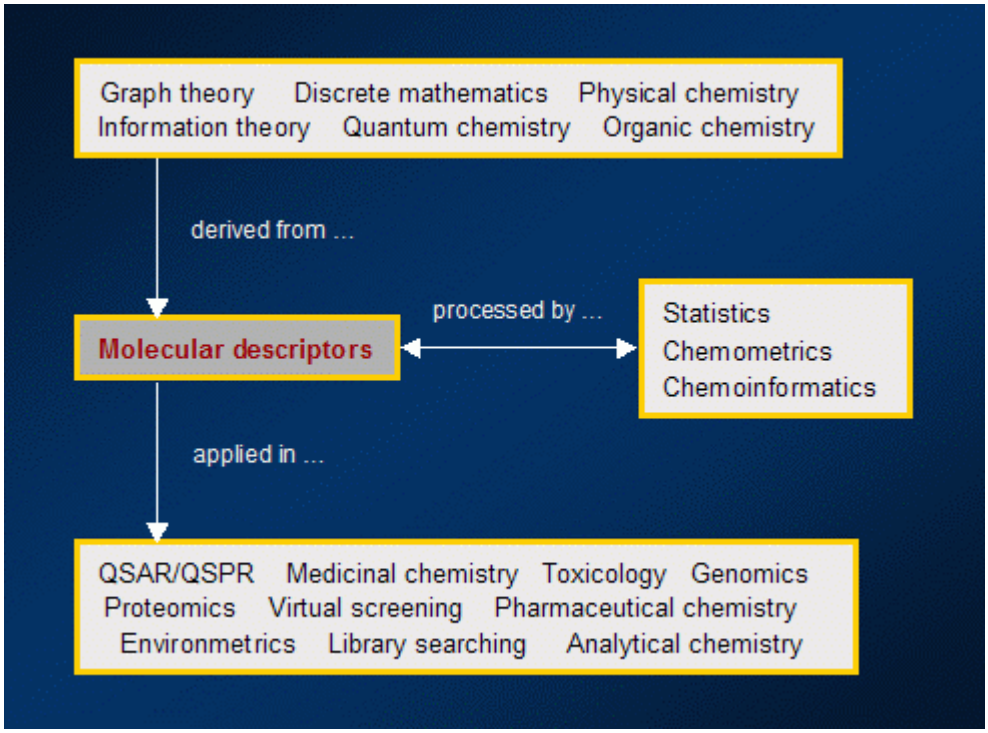

Figure 2

Therefore, molecular descriptors are now playing a key role in scientific research (Figure 2). In fact they are derived from several different theories, such as quantum chemistry, information theory, organic chemistry, graph theory, etc. and are applied in modelling several different properties in fields such as toxicology, analytical chemistry, physical chemistry, medicinal and pharmaceutical chemistry, environmental and toxicological studies and regulatory tools.

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 2000 of descriptors [1] are actually defined and

computable by using dedicated software tools. Each molecular descriptor takes into account 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. By now molecular descriptors are became among the most important variables used in molecular modelling, and as a consequence of that they have a strong relationship with statistics, chemometrics and chemoinformatics.

Statistics, chemometrics and chemoinformatics are the fields where methods for data elucidation, data mining and modelling are developed. In particular, chemometrics since about 30 years has developed several classification and regression methods able to provide – although not always - reliable models, both in reproducing the known experimental data and in predicting the unknown data. In fact, the modelling process has usually not only explanatory purposes, but – in particular in these last years – predictive purposes, i.e. it aims at developing models with reliable predictive qualities.

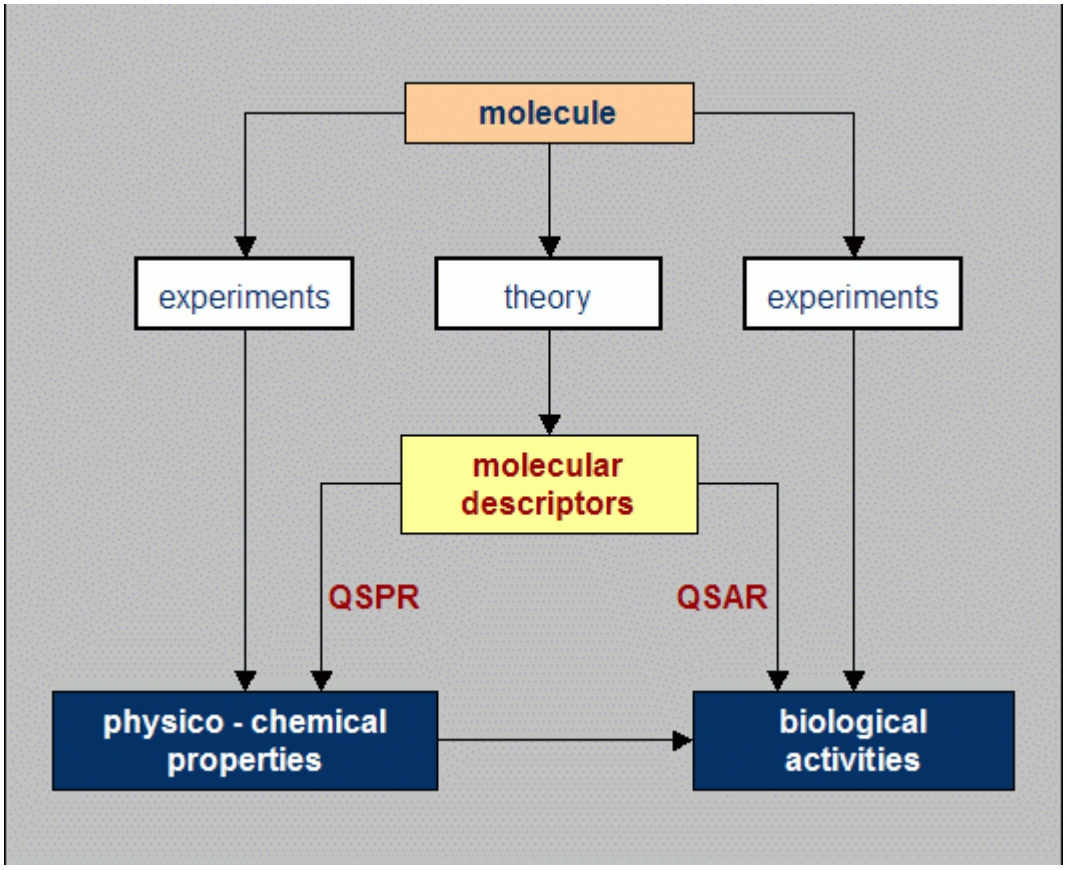

Figure 3

It has to be noted that the use of the molecular descriptors provided a big change in the scientific paradigm. In fact, while until 30 years ago molecular modelling mainly consisted in searching for mathematical relationships between experimentally measured quantities, now it is mainly performed modelling a measured property by the use of molecular descriptors able to catch structural chemical information (Figure 3).

To explain the complex relationships between molecules and observed quantities, two main streams were developed, the first related to the search for relationships between molecular structures and physico-

chemical properties (QSPR, *Quantitative Structure-Property Relationships*) and the second between molecular structures and biological activities (QSAR, *Quantitative Structure-Activity Relationships*).
The successes reached by using these approaches have encouraged the scientific community to apply them in other fields and relationships between molecular descriptors and environmental, toxicological, and technological properties are now widely investigated (Table 1).

Physico-chemical properties	Biological / toxicological activities	Environmental properties	Technological properties
boiling point	binding affinities	biodegradation	rubber vulcanization
melting point	pharmacological activties	bioconcentration	rheological behaviours
flash point	mutagenicity	half-life time	conductivity
solubility	carcinogenicity	mobility	stress
vapor pressure	lethal dose	athmospheric persistance	
molar volume	inhibition concentration	COD / BOD	
molar refractivity	druglikeness		

Table 1

Can mathematical models replace experiments?
Several people – and I am among them – foresee that in the future several quantities will be obtained by using predictive mathematical models, avoiding heavy and expensive experimental measurements. This extreme idea is supported by the successes of well established strategies for building regression and classification models, and by the capabilities of the developed models to suggest new active molecules (*drug design*) and new molecules with the required technological properties, to build priority lists of molecules for toxicological risk, to help in understanding molecule modes of action, to evaluate the environmental risks.
In more details, chemometric models are empirical mathematical relationships (*functions f*) obtained between a dependent experimental property (a measure *y* or a membership to a predefined class *c*) and some independent variables which are *related to* and *relevant for* the studied property:

$$y = f\left(x_1, x_2, \ldots x_p\right) \quad \text{or} \quad c = f\left(x_1, x_2, \ldots x_p\right)$$

In this context, the independent variables x are molecular descriptors and the models are built to reproduce to the greatest extent the known experimental responses by searching for relationships with these theoretical variables (Figure 4, step 1).
The validation tools developed in these last 20 years allow to evaluate not only the degree of agreement between the calculated and experimental responses, but also to estimate the future model capability to predict unknown responses.
If the predictive quality of the model is considered satisfactory, the model can be used for real future predictions of the modelled property (Figure 4 – step 2), resulting in a significant cost and animal testing reduction. In fact, the cost of the using the model as a predictive tool (second step) is almost zero (except the costs of the operator time and the current for the PC)!

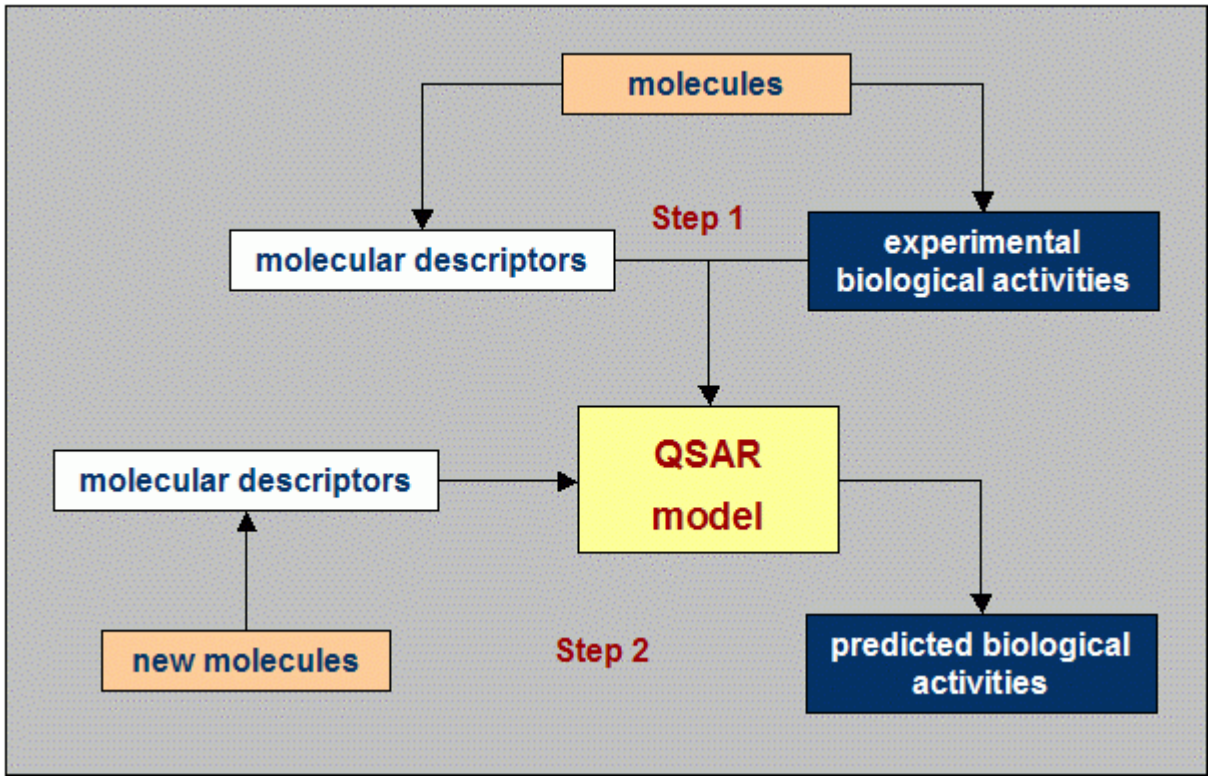

Figure 4

Conclusions

We can say that chemical research based on chemometric modelling and molecular descriptors is by now well established and further relevant results are expected. These expectations are in fact well founded on the two basic principles that 1) biological activities, as well as physico-chemical and chemical properties of organic compounds, are related to the molecular structure and that 2) similar compounds behave in similar way.

The reality is unique and, certainly, models are an approximation of the reality; however, different models represent different perspectives of the same reality and some of them might be useful. The fact that reality is represented by more than one model is often considered as a weakness of the theory due to our still limited knowledge: the scientific goal should be to reach a unique accepted model. This philosophical position is – in my opinion – mistaken: being a model only one possible interpretation of the reality, the availability of several models simply reflects the complexity of the studied systems where each model is able to catch a part of the whole information, i.e. it is a point of view: in summary, it is better to watch at a beautiful landscape by different windows.

References

1. R. Todeschini and V. Consonni: *Handbook of Molecular Descriptors*, WILEY-VCH, 2000.