How the Parts Organize in the Whole? A Top-Down View of Molecular Descriptors and Properties for QSAR and Drug Design

ERNESTO ESTRADA*

Complex Systems Research Group, X-Ray Unit, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, 15782 Santiago de Compostela, Spain.

Running header: Top-Down Molecular Descriptors

* E-mail: estrada66@yahoo.com; Fax: 34-981-547 077

ABSTRACT

Sometimes the complexity of a system, or the properties derived from it, do depend neither on the individual characteristics of the components of the system nor on the nature of the physical forces that hold them together. In such cases the properties derived from the "organization" of the system given by the connectivity of its elements can be determinant for explaining the structure of such systems. Here we explore the necessity of accounting for these structural characteristics in the molecular descriptors. We show that graph theory is the most appropriate mathematical theory to account for such molecular features. We review a method (TOPS-MODE) that is able to transform simple molecular descriptors, such as logP, polar surface area, molar refraction, charges, etc., into series of descriptors that account for the distribution of these characteristics (hydrophobicity, polarity, steric effects, etc) across the molecule. We explain the mathematical and physical principles of the TOPS-MODE method and develop three examples covering the description and interpretation of skin sensitisation of chemicals, chromosome aberration produced by organic molecules and drug binding to human serum albumin.

Keywords: molecular descriptors, drug design, QSAR, TOPS-MODE, QSPR

1. INTRODUCTION

A simple observation of the world gives us the clear idea that everything is made up of parts. Then, it is reasonably to try to figure out what an object does by figuring out what the parts do. In theoretical chemistry, bottom-up approaches are generally used to calculate molecular properties [1]. Usually, the contributions of atoms, bonds and molecular regions to a property are estimated by means of the quantum mechanical approaches to chemistry. The approach is essentially a bottom-up one [2] where the individual base elements of the system, e.g., atomic orbitals, are first specified in great detail. Then, these elements are linked together forming larger subsystems, e.g., molecular orbitals, which then in turn are linked, until the complete top-level system is formed. Other molecular descriptors used to study quantitative structure-property (QSPR) and structure-activity (QSAR) relationships also follow a bottom-up approach. For instance, substituent constants like the ones used in the Hansch or the Free-Wilson approaches to QSAR start by a detailed description of the parts to build the whole molecular property [3]. The risk in using a bottom-up approach in science is that the *"continual breaking down of the parts into their components parts progresses until we forget what it was we were trying to do in the first place!*" [4].

On the other hand, there are properties that do not depend on the nature of the molecular components but on the way these components are organized in the molecule. As a matter of example we can consider the number of isomers that exist with a given chemical formula. This question, which is relevant to combinatorial chemistry [5], cannot be responded by using a detailed description of the atoms or bonds in the molecules. The reason is that it depends on the way the atoms and bonds are combinatorially disposed in the molecule and not on their physical or chemical nature. The same philosophy can be

applied to any property/activity. In general (some exceptions are mentioned below), we are able to predict the biological and toxicological activity of a molecule, but we can say very little about the way in which the molecular parts contribute to this global activity or property. The reason is simply because we commonly use bottom-up approaches to predict properties/activities. However, it is possible to use an approach which consists in formulating a general overview of the system, which specifies, but not details, any first-level subsystems. Then, each subsystem is refined to increase their details until the entire specification is reduced to base elements. This strategy constitutes the *top-down approach* to science [2, 6]. The question about the existence of any general mathematical approach to account for a top-down view of the molecular structure in which we can analyze how the atoms/bonds organize in the molecule is important and necessary. This mathematical theory is expected to complement, more than rivaling, with quantum or extrathermodynamic (Hansch, Free-Wilson, etc.) views of molecular structure. Physicists have started to understand the importance of these organizational principles in the functioning of complex systems beyond the study of the nature of the elements that compose them. Stephen Hawkins has said that "*the next century will be the century of complexity*" [7]. This of course begs the question: How Chemistry is positioned to apply these ideas to the study of molecular properties?

2. A SHORT LESSON FROM COMPLEXITY

The aim of a complexity theory is the discovery of the laws of form that govern any collection of interacting parts, such as atoms and molecules, organisms like bacteria or mammals, individuals in a society, traders in a stock market, and even nations, ***regardless***

of what they are made of [8]. Consequently, some of the deepest truths about such systems can be truths about the organization of their components, rather than about what kinds of things make up such components and how they behave individually. This level of organization is represented through complex networks in which components are dots connected by lines that represent the interconnection between them [6, 8].

In recent years there has been a renaissance of the study of networks in physics and mathematics which has produced a number of new findings, documenting the power of networks in everything from business economy to drug discovery [6, 8-10]. In this context Barabási has cleverly stated that "*networks have become the X-ray machines of our connectedness, diagnosing the cell or the Web with the same ease*" [10]. This situation is in contrast with that existing in Chemistry, the scientific discipline which first welcomed the development and application of graph and network theory, where bottom-up approaches to molecular structure are still preferred. These approaches are well justified if we are interested in studying the nature of the chemical bonds or in molecular properties which are derived from it. However, if we are interested in the study of chemical properties derived from the connectivity of atoms in the molecule, we necessarily have to study their graph theoretical (network) features using a top-down approach.

3. IS A TOP-DOWN VIEW OF MOLECULAR DESCRIPTORS NECESSARY?

Despite there are methods like CoMFA [11] and Catalyst (available at www.accelerys.com) that permit to obtain maps of the molecular regions contributing to a given property or activity, most of the molecular descriptors existing today measure global

structural properties more than the distribution of such properties across the molecular structure [12]. For instance, partition coefficients or any of the descriptors quantifying hydrophobic properties of molecules do not gives any information about how hydrophobicity is distributed across the molecule. Net polar surface area does not indicate whether polarity is concentrated or spread across the molecule, and the situation is repeated for most of the molecular descriptors currently in use. This situation obligates in many cases the use of molecular descriptors in an indiscriminate way to obtain statistically significant QSPR/QSAR models. In such cases some descriptors are used to compensate the lack of information shown by the others, giving rise to very convoluted models in which information useful to chemists is encrypted in a way that makes the model useless.

It could be desirable to have a sort of map for the distribution of such molecular descriptors across the molecule in which we can "visualize" the contributions of different molecular regions to the global descriptor or property we are studying. The reason for the existence of molecular descriptors is their usability in describing other experimental properties, such as physicochemical or biological ones. Consequently, the main question here is whether these maps illustrating a descriptor's distribution across a molecule are necessary for describing other properties through quantitative relations. It is obvious that these "distributions across the molecule" depend on the connectivity pattern of such properties in the molecule. It is known that through the study of topological properties of networks we can identify which groups in a "social" network are more at risk of spreading an infection or which groups of Internet nodes are most susceptible to an attack [6, 8-10]. In a similar way we can identify which molecular regions have more or less "concentration" of any molecular property/descriptor. This approach is definitively a top-down approach to

molecular descriptors in which we start by defining some global properties of the molecules and then going down to "see" how they organize at atomic level.

Consider for instance, the aquatic toxicity of chlorobenzenes, which is believed to depend almost exclusively on their hydrophobicity [13]. It is known that chlorobenzene is less toxic to *Daphnia magna* than dichlorobenzenes, these are less toxic than trichlorobenzenes and so forth. Pentachlorobenzene is 50 times more toxic to *Daphnia magna* than chlorobenzene. A QSAR model obtained by Marchini *et al*. [13] show this general trend for seven arylbenzenes: $\log(1/\mathrm{EC50}) = 0.71 \cdot (\log \mathrm{P_{OW}}) - 3.53$. However, what happens if we analyze this trend in more detail by using a looking glass? We can see that, for instance, trichlorobenzenes (TCB) do not follow this expected general trend. 1,2,4-TCB and 1,2,3-TCB have similar logP values determined experimentally (4.02 and 4.05, respectively). 1,3,5-TCB has a logP value slightly higher than its isomers (4.19). However, 1,2,3-TCB is almost three times more toxic than 1,2,4- and almost seven times more toxic than 1,3,5-TCB. In fact, 1,2,3-TCB is as toxic as 1,2,3,4-tetrachlorobenzene, in spite of the fact that the last is significantly more hydrophobic than the first.

A close look at this trend of toxicity, 1,2,3-TCB > 1,2,4-TCB > 1,3,5-TCB, give us insights about what is happening. In 1,2,3-TCB the hydrophobicity is "concentrated" in a smaller molecular region than in 1,2,4-TCB and 1,3,5-TCB, where the hydrophobic groups, i.e., chlorines, are distributed across the molecule. Then a top-down approach that permits to account for the distribution of hydrophobicity across the molecule is necessary to explain these observed facts.

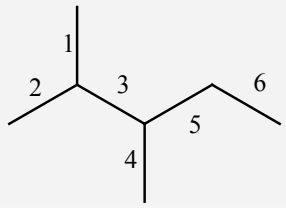
4. A TOP-DOWN APPROACH TO HYDROPHOBICITY

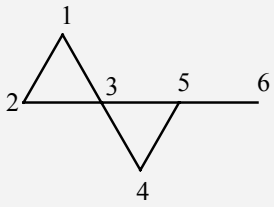
The logarithm of the n-octanol/water partition coefficient, logP, is considered a fundamental molecular descriptor in QSAR [14]. It accounts for the capacity of a molecule of distributing between an aqueous and lipid phase. In fact, until 2001, 965 out of 2129 QSARs for purified enzymes or more or less purified receptors, and 300 out of 709 QSARs for receptors contain hydrophobicity descriptors [15]. It is also very significant that 2937 out of 3677 QSARs developed for more complex systems, from organelles to whole organisms, also contain hydrophobicity terms [15].

In order to account for hydrophobicity across a molecule lets consider the following approach. First, let represents a molecule by means of a bond adjacency matrix, **B** [16] (see Box 1). Now, let considers the particular case of a "hydrophobicity bond matrix", $\mathbf{B}(H)$, where the main diagonal entries B_{ii} are the contribution of this bond to the partition coefficient n-octanol/water of the molecule. Then, obviously we have that the partition coefficient is the simple sum of the diagonal entries of this matrix, $\log P = \sum_{i=1}^{N} B_{ii}$. The sum of the diagonal entries of a matrix is known in mathematics as the spectral moment because it is also equal to the sum of the eigenvalues of such matrix. Thus we have that $\log P = \mu_1(H) = \sum_{j=1}^{N} \lambda_j(H)$, where $\lambda_j(H)$ are the eigenvalues of $\mathbf{B}(H)$ (see Box 1). As we already know this first moment of the hydrophobicity matrix does not reflect the distribution of the hydrophobicity across the molecule. But, what about the higher order moments, $\mu_{k>1}(H)$? The hydrophobicity moment of order k are defined as in Box 1. In Figure 1 we illustrate how the higher order moments account for the distribution of the hydrophobicity in the three TCB isomers. In order to account for the total effect of higher moments we use the following formula which gives the largest weight to the lower

Box 1 | **Spectral moments of the bond matrix**

The bond adjacency matrix is a square symmetric matrix whose non-diagonal entries are zeroes or ones as the corresponding bonds are adjacent or not, respectively. Two bonds are adjacent if they share a common atom. The bond matrix corresponds to the adjacency matrix of the line graph of the graph. A line graph is that built by representing any bond of the graph as a vertex in the line graph. Two vertices are adjacent in the line graph if the corresponding bonds are adjacent in the graph.





$$\mathbf{B} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Graph with labeled bonds Line graph Bond matrix

The kth spectral moment of the bond matrix, μ_k, corresponds to the trace of the kth power of the matrix. The trace is the sum of the diagonal elements of the matrix. For instance, for calculating the second moment we square the matrix and then sum the diagonal entries of this matrix giving the second spectral moment, which in this case is equal to 14. Each of the diagonal entries of this matrix are the bond contributions to the second spectral moment, for instance, bonds 1, 2 and 4 have contributions to the second moment of 2, while bond 3 has a contribution of 4. These contributions represent the number of pairs of adjacent bonds in which the corresponding bond participates. The relationship between spectral moments of the bond matrix and the spectrum of a graph with m bonds, i.e., the set of its eigenvalues, λ_i, is given by the following expression:

$$\mathbf{B}^2 = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 0 \\ 1 & 2 & 1 & 1 & 1 & 0 \\ 1 & 1 & 4 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 1 & 3 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}$$

$$\mu_k = Tr\left(\mathbf{B}^k\right) = \sum_{i=1}^{m} (\lambda_i)^k$$

Eigenvalues are a special set of scalars associated with the matrix **B**. If there is a vector

$$\mathbf{v} \in \Re^n \neq 0$$

such that

$$\mathbf{Bv} = \lambda\mathbf{v}$$

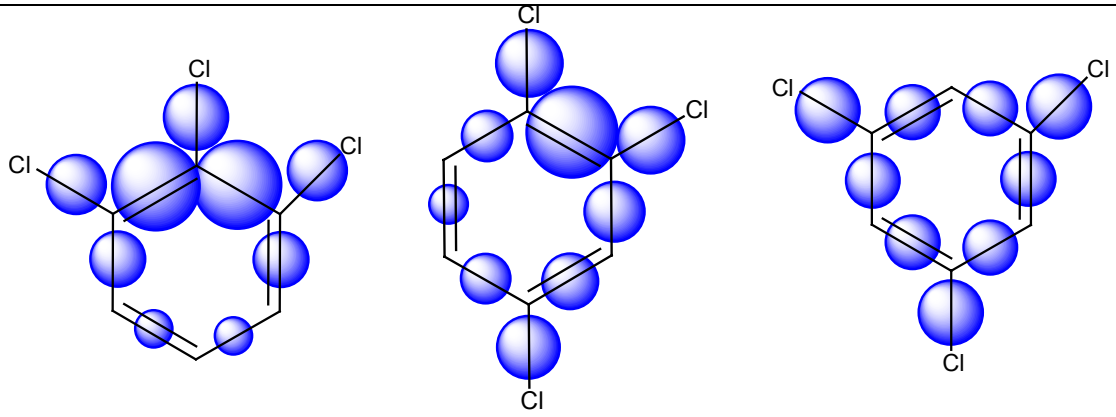
for some scalar λ, then λ is called an eigenvalue of **B** with corresponding (right) eigenvector **v**.

moments and gives lower weights to the higher moments [17], and we introduce a "hydrophobicity descriptor" or *HD* for brief

$$HD = \sum_{k=1}^{\infty} \mu_k / k! = \frac{1}{N} \sum_{j=1}^{N} e^{\lambda_j(H)} \quad (1)$$

It is easy to see that the molecular *HD* can be expressed in terms of bond contributions, $HD = \sum_{i=1}^{N} HD(i)$, where the bond contributions are given by the following expression (see Box 2) $HD(i) = \frac{1}{N} \sum_{j=1}^{N} [\gamma_j(i)]^2 e^{\lambda_j(H)}$, where $\gamma_j(i)$ is the i th component of the eigenvector associated with the j th eigenvalue of the bond matrix with hydrophobicity parameters in the main diagonal (see Box 1).

The values of $HD(i)$ are represented graphically in the Figure 1 for the three isomers under analysis. It can be seen that 1,2,3-TCB has a great "concentration" of hydrophobicity around chlorine atoms. In particular, the bonds in the benzene ring which are between two C-Cl bonds have the highest hydrophobic contribution. This molecule resembles a dipole, e.g., a "hydrophobic dipole", having a pole of high hydrophobicity and another less hydrophobic. In 1,2,4-TCB the hydrophobicity is distributed in a more homogeneous way across the molecule with the region around the two chlorines in ortho position still having a significantly higher contribution than the rest of the molecule. However, in 1,3,5-TCB this distribution is symmetrical across the molecule with each bond having approximately the same hydrophobicity. The values of *HD* also reflect this trend as can be seen in the Figure 1. These values follow the ecotoxicological profile observed experimentally [13] for these compounds which are also given in this figure. This simple example begs the question about whether we can extend this idea to any other property/activity.



Log(1/EC_{50}	-0.33	-0.76	-1.14
LogP	4.05	4.02	4.19
μ_1	3.89	3.89	3.89
μ_2	26.19	26.18	26.17
μ_3	46.75	47.16	47.58
μ_4	170.40	168.08	165.80
HD	2.630	2.621	2.612

Fig (1). Isomers of trichlorobenzene and its n-octanol/water partition coefficients experimentally determined, as well as the spectral moments of the hydrophobicity matrix and toxicity values for polar narcosis to *Daphnia magna* at 48 h (Log(1/EC_{50}).

5. FROM GLOBAL TO LOCAL CONTRIBUTIONS

The first thing we need to understand is whether we need global molecular descriptors in QSAR/QSPR. Why not directly to use local descriptors defined for atoms or bonds? If we were interested in studying congeneric sets of organic compounds there is no difficulty in relating the property P to atomic or bond parameters of the compounds under study.

The problem arises when we attempt to study heterogeneous datasets of organic molecules. In this case there is not necessarily an atomic/bond pattern which is repeated in all the molecules under study. As a matter of example lets consider a dataset which contains an alkane, an α,β-unsaturated aldehyde, and an aromatic amine. Then we have not a common atom or bond for which we can calculate the atomic/bond parameter which will be related to the property P. As a consequence we have to use molecular descriptors like the electronic chemical potential, the molecular electronegativity, the chemical hardness, or other global molecular indices [12].

This question immediately poses another, which is whether we can obtain structural information at a local scale from the models developed using global molecular descriptors. The only information that we need to transform the global model into the atomic/bond contributions is the mathematical relationship between the global molecular descriptor and the local contributions.

A possible strategy to account for this aspect of molecular complexity is to use descriptors based on a graph theoretical representation of molecules. The great advantage of using graph theory based molecular descriptors [18] is that we can always obtain a mathematical relationship between the global index and the structural molecular fragments. This connection is guarantied by the Baskin et al. [19] theorems, which prove that *any topological index can be uniquely represented as i) a linear combination of occurrence numbers of some structural fragments, both connected and disconnected, or ii) a polynomial on occurrence numbers of connected substructures of the corresponding*

molecular graph. Then, we can build efficient ways of transforming global descriptors/properties into local distribution maps.

6. A TOP-DOWN APPROACH TO QSAR/QSPR

In the last few years the TOPS-MODE approach to QSAR/QSPR has been developed to account for the contributions of molecular parts to the global molecular properties [20]. TOPS-MODE (Topological Sub-Structural Molecular Descriptors/Design) is based on the spectral moments $\mu_k(w)$ of bond matrices [20], where w represents the weights used in the diagonal of the matrices, to account for hydrophobicity, polar surface area, polarizability, molar refractivity, van der Waals radii, and electronic charges.

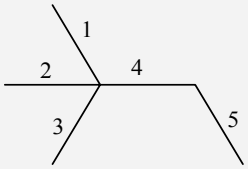
This approach has been applied to the study of chromatographic [21], diamagnetic, magneto-optic properties [22] and the permeability coefficients through low-density polyethylene [23] of organic compounds as well as the soil sorption coefficients for pesticides [24]. Many studies have reported the application of TOPS-MODE in QSAR and drug design, which include the design of new sedative/hypnotic [25], anticonvulsant [26], anticancer [27], antiinflamatory [28], herbicides [29], antibacterial [30], and central nervous system activity [31]. Other studies reported the design of anti-HIV nucleosides [32, 33], antioxidants analogues of compounds in Brazilian propolis [34], adenosine receptors inhibitors [35], and antifungal compounds [36].

TOPS-MODE has been applied to predict ADMET (Administration, Distribution, Metabolism, Excretion and Toxicity) parameters of drugs and drug-like compounds. They include the physicochemical, absorption and pharmacokinetics properties of 6-fluoroquinolone derivatives [37], the prediction of blood-brain barrier permeation [38],

human intestinal absorption [39], binding to P-glycoprotein substrates [40], and binding of drugs to human serum albumin [41]. Other works have been devoted to the understanding of skin sensitization mechanisms [42, 43], the study of mutagenic activity in dental monomers [44], the prediction of rodent carcinogenicity [45], the prediction of nitrocompounds carcinogeneicity [46] and the study of chromosome aberrations produced by drugs and drug-like compounds [47]. The prediction and understanding of the behaviour of organic chemicals in the environment or human health after the exposition of diverse doses of such compounds has also been studied by using TOPS-MODE [48, 49].

An important question related to this method is related to its top-down nature. The three theorems of Baskin et al. [19] previously mentioned were proved for labelled graphs, that is for graphs in which vertices and/or edges are weighted by some real numbers. Consequently, they assure that we can always obtain a linear combination of the TOPS-MODE descriptors in terms of structural fragments of the molecules under study. The general strategy for obtaining local contributions from global spectral moments is illustrated in the Box 2. In closing, we can say that *if any structure-property data is sufficiently large to allow building statistically significant models with TOPS-MODE descriptors, then we can express this property as an additive function of bond contributions*.

Box 2 | **Calculation of bond contributions**



Using a QSPR model for the molar refraction of alkanes [20] we calculate bond contributions for the molecule of 2,2-dimethylbutane with the bond numbering given in the figure. The total spectral moments can be expressed as sum of bond spectral moments of the form:

$$\mu_k = \sum_i \mu_k(i)$$

In terms of the eigenvalues λ_j of the **B** matrix and the corresponding eigenvectors $v_j(i)$ the local spectral moments can be expressed as

$$\mu_k(i) = \sum_j [v_j(i)]^2 \lambda_j^k$$

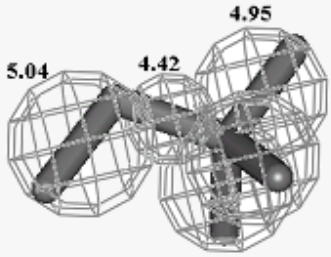
The bond spectral moments for the bonds of this molecule are as follows:

Bond/μ_k	k=0	k=2	k=3	k=4
1	1	3	6	22
2	1	3	6	22
3	1	3	6	22
4	1	4	6	24
5	1	1	0	4

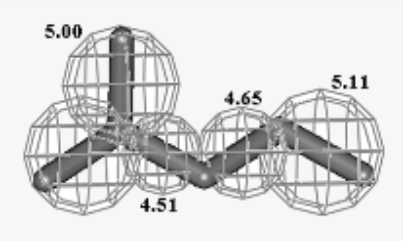
Now we will substitute these expressions into the QSPR model obtaining bond molar refractions as exemplified for bond 1:

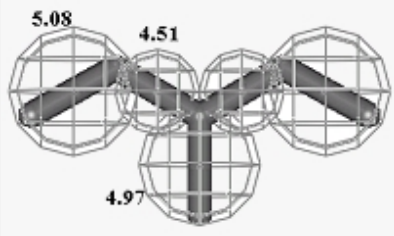
$$\mathrm{MR}(1) = 5.506 \cdot 1 - 0.329 \cdot 3 + 0.193 \cdot 6 - 0.033 \cdot 22 = 4.95 \mathrm{cm}^3$$

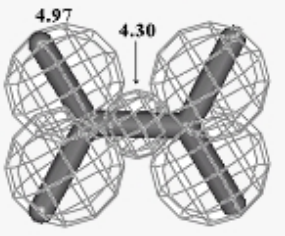
In a similar way the bond contributions of the other bonds are obtained and graphically visualized. It is clear that the sum of these bond contributions plus the intercept of the QSPR model (13) gives the value of the molar refraction of the molecule: 30.026 cm^3.



Bond contributions for the other three hexane isomers are given below:







7. APPLICATIONS TO THE QSAR/DRUG DESIGN

7.1. Skin sensitization of organic compounds.

The potential of a chemical to develop skin sensitization in humans is of tremendous importance for the topical application of such substance. Skin sensitization is an important aspect of the allergic contact dermatitis, which is produced as a complex process involving the stimulation of the immune system producing an inflammatory response in the skin. Using TOPS-MODE we have developed a quantitative model which predicts the potential of an organic compound to develop skin sensitization. The training set used to develop this model was formed by 93 organic molecules of different classes, which include alkyl halides, aldehydes, amides, esters, ketones, nitriles, nitrocompounds, aromatic amines, phenols, sulfides, among others. This structural heterogeneity obligates to use global molecular descriptors like the TOPS-MODE ones. The model developed classifies these compounds according to their potencies as strong/moderate, weak and extremely weak/non-sensitizers [42]. Using this global structural information we have obtained the bond contributions for all chemical bonds in the molecules studied. In the Figure 2 we illustrate some of these contributions for the bonds identified as responsible for the skin sensitization of two aromatic amines and two aldehydes [42, 43].

The information about the groups having a positive contribution to the skin sensitization has been used to propose structural alerts based on the presence of certain toxicophores in the molecules to be analyzed [43]. Søsted et al. [51] used this model for ranking 229 hair dye substances according to their predicted skin sensitization potency.

None of these substances was previously included in our models. Recently some of these predictions of skin sensitizers have been confirmed experimentally [52, 53].

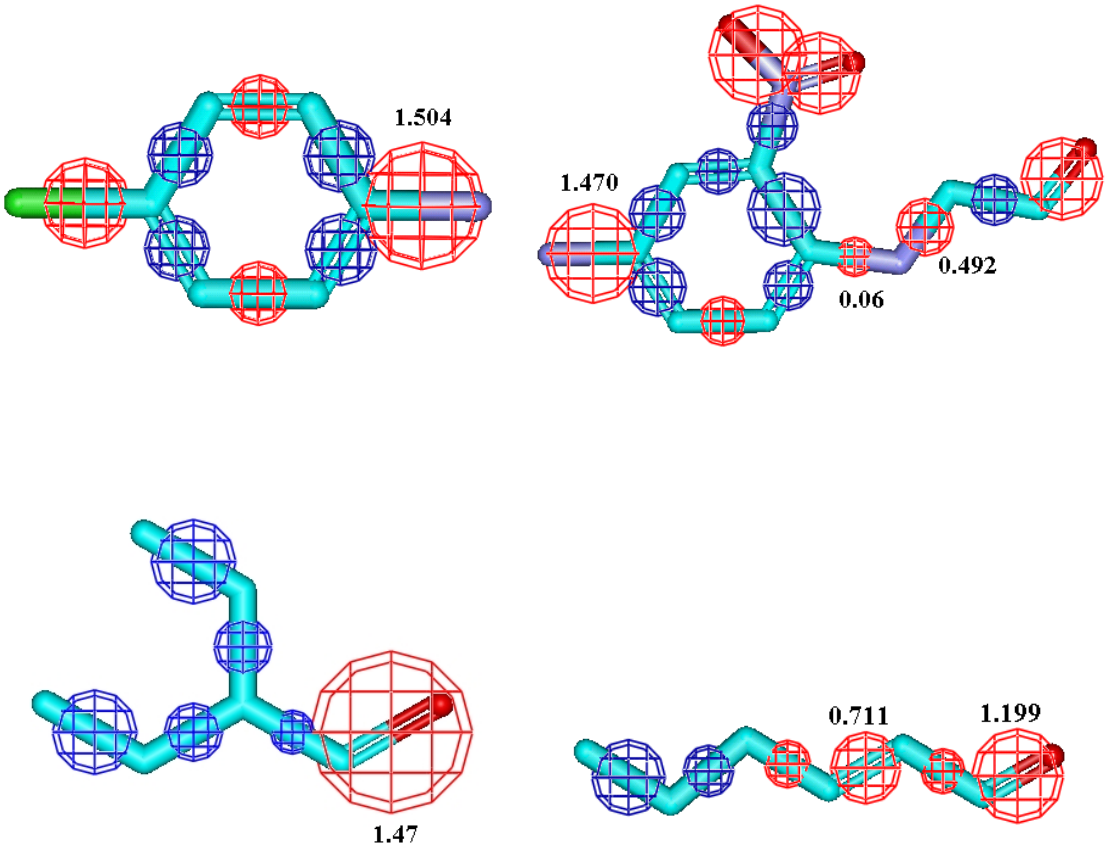


Fig (2). Visualization of the bond contributions to skin sensitization of a primary and a secondary aromatic amine as well as for a saturated and an α,β-unsaturated aldehyde according to the predictions made using TOPS-MODE. Gray spheres correspond to positive contributions to skin sensitization, some of them display the contributions. Black spheres correspond to negative contributions.

7.2. Chromosome aberration of organic compounds.

We have recently used TOPS-MODE to generate structural alerts that predict the clastogenic potential of an organic molecule [47]. Clastogens, which are chromosome breaking chemicals, can induce chromosome aberrations by different mechanisms. These

mechanisms include DNA alkylation, inhibition of deoxyribonucleotide synthesis, denaturation or degradation of DNA, production of labile DNA by chemical reaction and/or incorporation of abnormal precursors as well as removal of DNA bound metals [47]. Our strategy to generate structural alerts consists in identifying those molecular "positive" regions which are repeated in several molecules. Positive regions refer to those having positive contributions to the property/activity under consideration. These structural alerts can be easily implemented in expert systems for the prediction of toxicity or biological activity. In this study we have used a data set of 383 organic compounds which were classified as clastogenic/non-clastogenic [47]. Using this information we have generated 22 structural alert rules, which include those for N-nitrosoureas, N-nitrosourethanes, nitro compounds, alkyl esters of phosphoric acids, alkyl methanesulfonates, epoxides, amines, phenols, urethanes, α, β- unsaturated carboxylic acids, amides, esters and ketones, among others. In Figure 3 we illustrate some N-nitrosoureas which are included in the dataset. In addition we also illustrate the strategy followed for the generation of the structural alert found for these compounds [47].

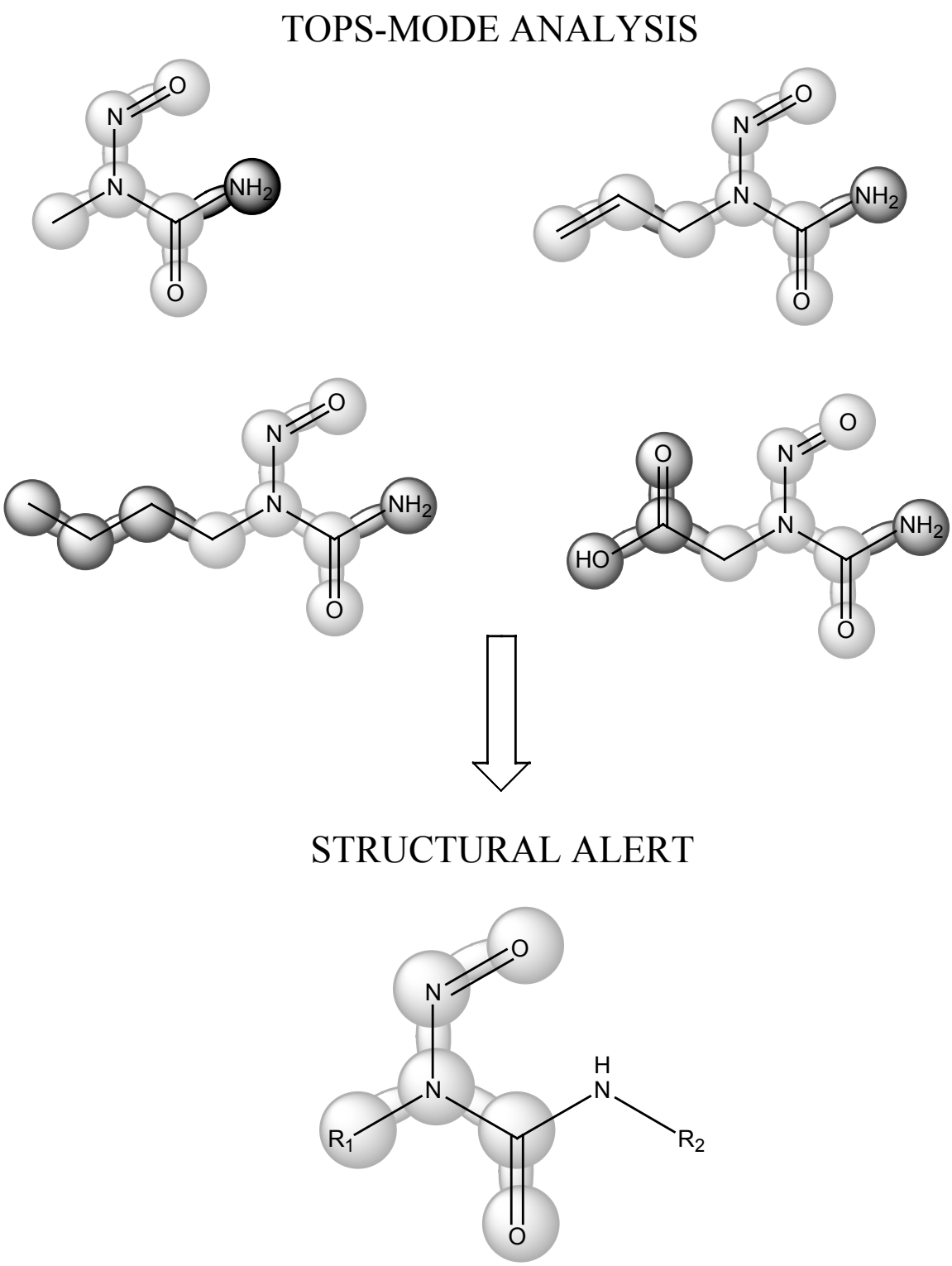


Fig (3). Illustration of the structural alert generation process. In clear/dark we have marked those regions which are predicted to have/not-have a positive contribution to the

chromosome aberration. A structural alert is obtained by the maximal common fragment which is present in the majority of the structures studied, which here corresponds to the N-nitrosamide moiety.

7.3. Drug binding to human serum albumin (HSA).

HSA is the most abundant protein in plasma, which is known to form mainly non-covalent complexes with exogenous ligands. Most of the drugs that bind to HSA form complexes in which the drug is located at one of the two main binding sites of HSA. The drug site I has a predominantly apolar interior with two polar clusters. The drug site II consists of a largely hydrophobic cavity with distinct polar features. In this study a robust QSAR model was obtained by using the TOPS-MODE approach for 78 drugs in the training set and 10 others used for prediction [41]. Following our top-down approach to QSAR/QSPR we have calculated the bond contributions to the drug-binding to HSA for the 88 molecules studied. These bond contributions were transformed into the contributions of fragments or functional groups. The sum of contributions for all bonds forming the fragment is considered to be the fragment's global contribution. In this way, we have calculated the contribution of 65 different groups to the drug-HSA binding. The contribution for the same group in different molecules is averaged and reported as the group contribution for this specific fragment independent of the molecule in which it is located.

A perfect agreement exists between the group/fragment contributions found by TOPS-MODE and the specific interactions of drugs with HSA [41]. These results indicate a preponderant contribution of hydrophobic regions of drugs to the specific binding to drug binding sites 1 and 2 in HSA and specific roles of polar groups which anchor drugs to HSA

binding sites. For instance, warfarin is a drug that binds to site I. In Figure 4 we show the contributions to the HSA-binding for the main groups of this drug [41]. TOPS-MODE identifies the main contributions of the hydrophobic moieties, which are located at major hydrophobic pockets of the protein as well as the electrostatic interactions between the oxygen of the hydroxyl and carbonyl groups. These groups form stabilizing hydrogen bonds as well as destabilizing interactions with the residues of binding site I. In closing, the top-down approach based on TOPS-MODE fits very well with the experimental molecular models for the drug-HSA interactions, which illustrates its utility beyond the classical QSAR/QSPR applications.

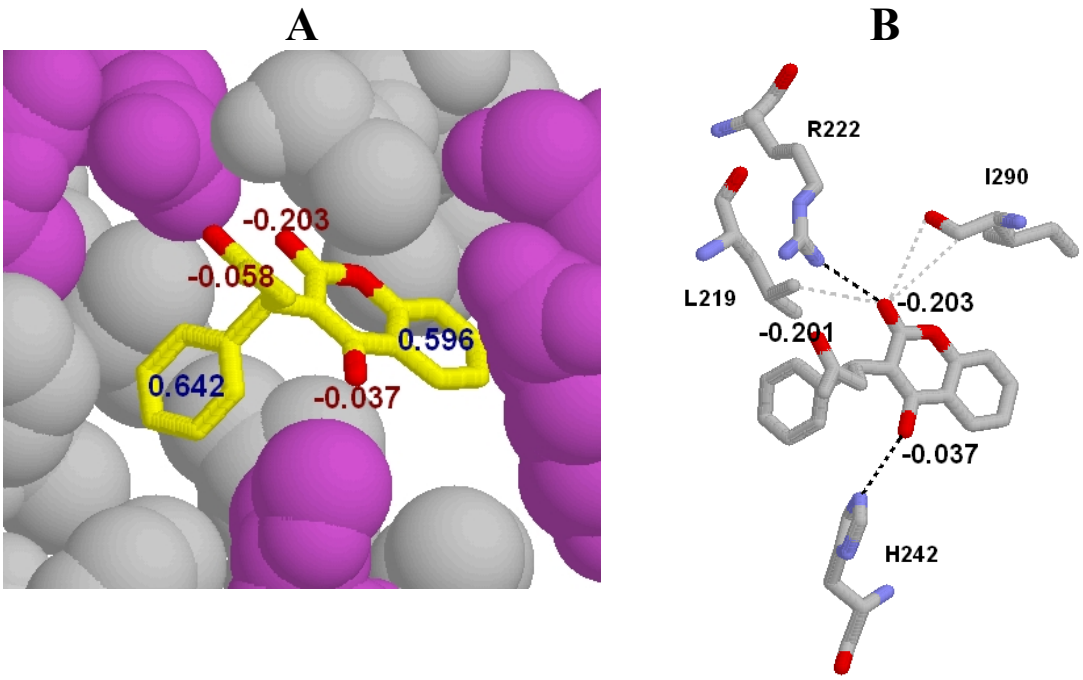


Fig (4). A) Illustration of the contributions of fragments to the interaction of warfarin at the drug binding site I of HAS . B) Interactions of carbonyl and hydroxyl oxygen atoms of warfarin with residues of HSA. In black dotted lines are the hydrogen bonding formed with R222 and H242 which stabilize the HSA-warfarin complex. In gray dotted lines are the

hydrophobic-hydrophilic and oxygen-oxygen repulsion, which destabilize the warfarin-HSA complex.

8. CONCLUDING REMARKS

The complexity of the molecular structure depends on the scale used for its description. In general, as we get to very large scales the complexity of the description is significantly low. However, as we get further and further by reducing the scale, the complexity increases non-linearly in a dramatic way. This is exactly what happens when we use different molecular descriptors for studying the molecular structure. At very large scales we can represent a molecule by a single dot, if we are interested only in their statistical mechanic properties. As we reduce the scale we can represent the atoms and bonds as nodes and links of a simple graph. Then, we can analyze any of the properties which arise as a consequence of the connectivity pattern of a molecule. At this level we are investigating the "topological world" of the molecular structure. The complexity of this description is large enough compared to the previous representation, but it is tiny in comparison with that obtained by reducing the scale up to the "quantum world". At this scale we study the internal nature of the atoms and bonds, which increase considerably the complexity of the system. But, think about the complexity of reducing even more the scale and "see" the simultaneous movement of the electrons, vibration of atoms and so forth. In Figure 5 we represent the complexity as a function of scale for the molecular structure.

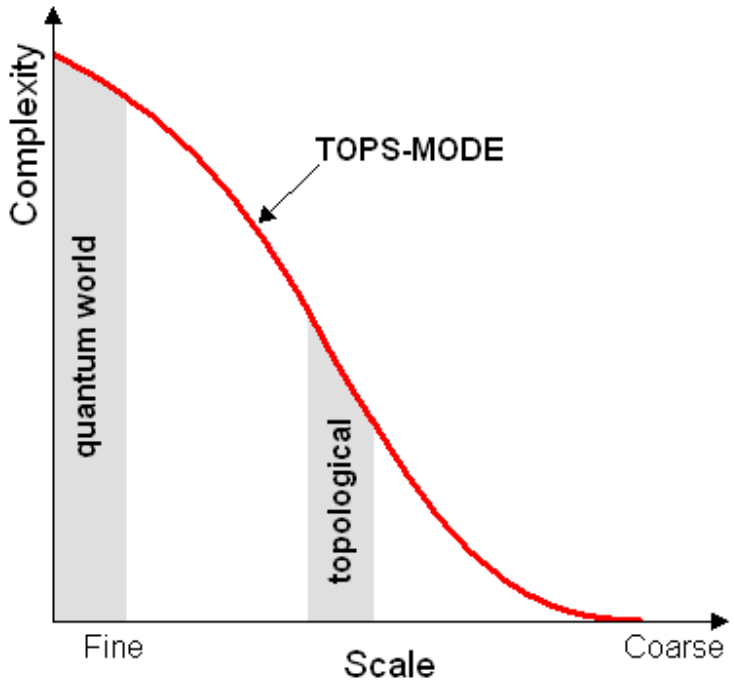


Fig (5). The place of topological and quantum worlds in the complexity-scale plot. The TOPS-MODE is expected to occupies an intermediate position.

On the topological side of the scale we make emphasis on the way in which the parts of the system are organized. On the other extreme we make emphasis on the nature of such parts, e.g., atoms, bonds, electrons, etc. However, it has been stated that "*a system can be fully understood in terms of its parts and the interactions between them*" [54] by recognizing that such interactions "*often lead to global patterns of organization that cannot be traced to the particular parts*" [8]. The TOPS-MODE approach is a sort of intermediate state, or using a physicists language it is a sort of "meso-scale" description of the molecular structure. In this approach we use information at the local scale, such as bond properties, which indeed can also be extracted from the quantum world. Then, this information is combined in a topological way to extract the information arising from their

interrelationships. As a consequence, we have placed the TOPS-MODE somewhere in between the topological and the quantum worlds. Of course, it is possible that other intermediate approaches like this exist and there are enough places in the plot to locate them. TOPS-MODE has been recognized as a useful tool to investigate practical problems related to the molecular structure. For instance, it has been recognized that this approach "*provides a mechanistic interpretation at a bond level and enables the generation of new hypotheses such as structural alerts*" [55]. Thus the use of top-down approaches to the study of molecular structure is not only useful but also a necessary approach in chemistry.

ACKNOWLEDGEMENTS

The author thanks partial financial support by the program "Ramón y Cajal", Spain. The author wants to thank all his coworkers and user of the TOPS-MODE approach, whose names appear in the list of references of this work.

REFERENCES

[1] Primas, H. Chemistry, Quantum Mechanics and Reductionism. Springer: Berlin, **1983**.

[2] Ellis, F.F.R. *Found. Phys*. **2006**, 36, 227.

[3] Kubinyi, H. QSAR: Hansch Analysis and Related Approaches. John Wiley & Sons: New York, **1993**.

[4] Bar-Yam, Y. Making Thinks Work. Knowledge Press: Cambridge, MA, **2005**.

[5] El Basil, S. Combinatorial Organic Chemistry: An Educational Approach. Nova Science: New York, **2000**.

[6] Bray, D. *Science* **2003**, 301, 1864.

[7] Hawking, S. *San José Mercury News*, January 23th, 2000.

[8] Buchanan, M. Nexus. W. W. Norton & Co.: New York, 2002.

[9] Strogatz, S. *Nature* **2001**, 410: 268.

[10] Barabási, A.-L. Linked. Plume, NY, **2003**.

[11] Kubinyi, H. In Encyclopedia of Computational Chemistry; J. Gasteiger, Ed.; Wiley, London, **1998**.

[12] Todeschini, R.; Consonni, V. Handbook of Molecular Descriptors. Wiley-VCH: Weinheim, **2000**.

[13] Marchini, S.; Passerini, L.; Hoglund, M. D.; Pino, A.; Mendza, M. *Environ. Toxicol. Chem*. **1999**, 18, 2759.

[14] Leo, A. J.; Hansch, C. *Persp. Drug Discov. Des*. **1999**, 17, 1.

[15] Hansch, C.; Kurup, A.; Garg, R.; Gao, H. *Chem. Rev*. **2001**, 101, 619.

[16] Estrada, E. *J. Chem. Inf. Comput. Sci*. **1995**, 35, 31.

[17] Estrada, E.; Rodríguez-Velázquez, J. A. *Phys. Rev. E* **2005**, 71, 051601.

[18] Devillers, J.; Balaban, A. T. (Eds.). Topological Indices and Related Descriptors in QSAR and QSPR. Gordon & Breach: Amsterdam, **1999**.

[19] Baskin, I. I.; Skvortsova, M. I.; Stankevich, I. V.; Zefirov, N. S. *J. Chem. Inf. Comput. Sci*. **1995**, 35, 527.

[20] Estrada, E.; Uriarte, E. *Curr. Med. Chem*. **2001**, 8, 1699.

[21] Estrada, E.; Gutierrez, Y. *J. Chromatogr. A* **1999**, 858, 187.

[22] Estrada, E.; Gutierrez, Y.; González, H. *J. Chem. Inf. Comput. Sci*. **2000**, 40, 1386.

[23] Pérez-González, M.; Morales-Helguera, A. *J. Comp.-Aided Mol. Des.* **2003**, 17, 665.

[24] Pérez-González, M.; Morales-Helguera, A.; Collado, I. G. *Mol. Div.* **2006**, 10, 109.

[25] Estrada, E.; Peña, A.; García-Domenech, R. *J. Comp.-Aided Mol. Des.* **1998**, 12, 583.

[26] Estrada, E.; Peña, A. *Bioorg. Med. Chem*. **2000**, 8, 2755.

[27] Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem*. **2000**, 43, 1975.

[28] Pérez-González, M.; Carlos Dias, L.; Morales-Helguera, A.; Morales-Rodríguez, Y.; Gonzaga de Oliveira, L.; Torres-Gomez, L.; González-Díaz, H. *Bioorg. Med. Chem*. **2004**, 12, 4467.

[29] Pérez-González, M.; González-Díaz, H.; Molina-Ruiz, R.; Cabrera-Pérez, M. A.; Ramos-de-Armas, R. *J. Chem. Inf. Comput. Sci*. **2003**, 43, 1192.

[30] Molina, E.; González-Díaz, H.; Pérez-González, M.; Rodríguez, E.; Uriarte, E. *J. Chem. Inf. Comput. Sci*. **2004,** 44, 515.

[31] Cabrera-Pérez, M. A.; Bermejo-Sanz, M. *Bioorg. Med. Chem*. **2004**, 12, 5833.

[32] Estrada, E.; Vilar, S.; Uriarte, E.; Gutierrez, Y. *J. Chem. Inf. Comput. Sci*. **2002**, 42, 1194.

[33] Vilar, S.; Estrada, E., Uriarte, E. Santana, L., Gutierrez, Y. *J. Chem. Inf. Comput. Sci*. **2005**, 45, 502.

[34] Estrada, E., Quincoces, J., Patlewicz, G. *Mol. Div.* **2004**, 8, 21.

[35] Pérez-González, M.; Teran-Moldes, M. C. *Bioorg. Med. Chem*. *Lett.* **2004**, 14, 3077.

[36] Saíz-Urra, L.; Pérez-González, M.; Collado, I. G.; Hernández-Galán, R. *J. Mol. Graph. Modell.* **2007**, 25, 680.

[37] Cabrera-Pérez, M. A.; Ruiz-García, A.; Fernández-Teruel, C.; González-Álvarez, I.; Bermejo-Sanz, M. *Eur. J. Pharm. Biopharm*. **2003**, 56, 197.

[38] Cabrera-Pérez, M. A.; Bermejo, M.; Pérez, M.; Ramos, R. *J. Pharm. Sci*. **2004**, 93, 1701.

[39] Cabrera-Pérez, M. A.; Bermejo-Sanz, M.; Ramos-Torres, L.; Grau-Ávalos, R.; Pérez-González, M.; González-Díaz, H. *Eur. J. Med. Chem*. **2004**, 39, 905.

[40] Cabrera-Pérez, M. A.; González, I.; Fernández, C.; Navarro, C.; Bermejo, M. *J. Pharm. Sci*. **2006**, 95, 589.

[41] Estrada, E., Uriarte, E., Molina, E., Simón-Manso, Y., Milne, G. W. *J. Chem. Inf. Model.* **2006**, 46, 2709.

[42] Estrada, E., Patlewicz, G., Chamberlain, M., Basketter, D., Larbey, S. *Chem. Res. Toxicol.* **2003**, 16, 1226.

[43] Estrada, E., Patlewicz, G., Gutierrez, Y. *J. Chem. Inf. Comput. Sci*. **2004**, 44, 688.

[44] Pérez-González, M.; Teran-Moldes, M. C.; Fall, Y.; Carlos Dias, L.; Morales-Helguera, *Polymer* **2005**, 46, 2783.

[45] Morales-Helguera, A.; Cabrera-Pérez, M. A.; Pérez-González, M.; Molina-Ruiz, R.; González-Díaz, H. *Bioorg. Med. Chem*. **2005**, 13, 2477.

[46] Morales-Helguera, A.; Cabrera-Pérez, M. A.; Combes, R. D.; Pérez-González, M. *Toxicology* **2006**, 220, 51.

[47] Estrada, E.; Molina, E. *J. Mol. Graph. Modell*. **2006**, 25, 275.

[48] Estrada, E.; Molina, E.; Uriarte, E. *SAR QSAR Environ. Res*. **2001**, 12, 445.

[49] Estrada, E., Uriarte, E.; Gutierrez, Y.; Gonzalez, H. *SAR QSAR Environ. Res*. **2003**, 14, 145.

[50] Estrada, E. *J. Chem. Inf. Comput. Sci*. **1996**, 36, 844.

[51] Søsted, H.; Basketter, D. A.; Estrada, E.; Johansen, J. D.; Patlewicz, G. Y. *Contact Dermatitis* **2004**, 51, 241.

[52] Søsted, H.; Nenéd, T. *Contact Dermatitis* **2005**, 52, 317.

[53] Katugampola, R. P.; Statham, B. N. *Contact Dermatitis* **2005**, 53, 234.

[54] Here the term "interaction" is used in the sense of "interconnection" and is not referred to the nature of the physical forces keeping together the elements of the system.

[55] Environment Directorate OEMC. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models. OECD Environmental Health and Safety Publications. Series on Testing and Assessment No. 69; **2007**.