Universitat de Girona

MOLECULAR QUANTUM SIMILARITY IN QSAR: APPLICATIONS IN COMPUTER-AIDED MOLECULAR DESIGN

Ana GALLEGOS SALINER

ISBN: 84-689-0623-9
Dipòsit legal: GI-1168-2004
http://hdl.handle.net/10803/7937

Doctoral Thesis for the obtaining of the degree of
Doctor in Theoretical and Computational Chemistry

MOLECULAR QUANTUM SIMILARITY IN QSAR: APPLICATIONS IN COMPUTER-AIDED MOLECULAR DESIGN

Ana Gallegos Saliner

Girona, June 2004

Institut de Química Computacional
Departament de Química
Universitat de Girona

El sotasignat, Professor **Ramon Carbó-Dorca i Carré**, Catedràtic d'Universitat del Departament de Química de la Universitat de Girona

CERTIFICA:

Que Ana Gallegos i Saliner, llicenciada en Química per la Universitat de Girona, ha realitzat sota la meva direcció, a l'Institut de Química Computacional i al Departament de Química de la Universitat de Girona, el treball d'investigació titulat:

"Molecular Quantum Similarity in QSAR: Applications in Computer-Aided Molecular Design"

que es recull en aquesta memòria i es presenta en pública defensa per a optar al grau de Doctora en Química Teòrica i Computacional.

I, perquè consti als efectes legals escaients, signo aquest certificat a Girona, el 20 d'Abril de 2004.

Ramon Carbó-Dorca
Director de l'Institut de Química Computacional
Professor Catedràtic de la Universitat de Girona

Girona, 20 d'Abril de 2004

A la meva família,

A les meves amigues de Figueres,

Als de Castelló,

Als companys i amics IQC,

Per tots vosaltres!

Ara mateix enfilo aquesta agulla
amb el fil d'un propòsit que no dic
i em poso a apedaçar. Cap dels prodigis
que anunciaven taumaturgs insignes
no s'ha complert i els anys passen de pressa.
De res a poc, i sempre amb vent de cara,
quin llarg camí d'angoixa i de silencis.
I som on som; més val saber-ho i dir-ho
i assentar els peus en terra i proclamar-nos
hereus d'un temps de dubtes i renúncies
en què els sorolls ofeguen les paraules
i amb molts miralls mig estrafem la vida.
De res no ens val l'enyor o la complanta,
ni el toc de displicent malenconia
que ens posem per jersei o per corbata
quan sortim al carrer. Tenim a penes
el que tenim i prou: l'espai d'història
concreta que ens pertoca i un minúscul
territori per viure-la. Posem-nos
dempeus altra vegada i que se senti
la veu de tots, solemnement i clara.
Cridem qui som i que tothom ho escolti.
I, en acabat, que cadascú es vesteixi
com bonament li plagui, i via fora,
que tot està per fer i tot és possible.

Ara mateix

Miquel Martí i pol

MOLECULAR QUANTUM SIMILARITY IN QSAR:

APPLICATIONS IN

COMPUTER-AIDED MOLECULAR DESIGN

Ana Gallegos Saliner

Contents

Preface

1 PRESENTATION

The present thesis, entitled *Molecular Quantum Similarity in QSAR: Applications in Computer-Aided Molecular Design,* is the summarised memory of the work developed during the period of four and a half years in fulfilment of the requirement for the defence and obtaining of the Ph.D.[†] degree in Theoretical and Computational Chemistry. Held in the subgroup of **Molecular Quantum Engineering** (**EMQ**)[‡], in the **Institute of Computational Chemistry** (**IQC**)[‡] of **University of Girona** (**UdG**)[‡], the advisory directives have been provided by Professor **Ramon Carbó-Dorca**. In addition, the tight collaboration with the other components of the group, Emili Besalú, Lluís Amat, David Robert, and, specially, Xavier Gironés, has not only been enlightening but also productive.

The long journey of doctorate began in 2000, with the instruction learned in the first edition of the **Interuniversity Course in Theoretical and Computational Chemistry,** held in Castelló. This training provided the foundation for the basic knowledge in the theoretical chemistry field. Already in the lab, the following stage was devoted to the learning of the basic gear of **Quantum Similarity Theory** and **QSAR techniques**, as well as to the initiation in computing strategies. This primary learning phase concluded with the elaboration of the master research project to obtain the **Advanced Studies Diploma** (**DEA**) [‡], defended in June 2001.

After this period, several research projects have been proposed and have prospered, resulting in different publications. Among the ensemble of cooperation opened with other scientific research groups that conforms this thesis, some of them are still in progress.

[†] Doctor of Philosophy

[‡] Acronyms derived from Catalan language

2 STRUCTURE OF THE MEMORY

The structure of the memory, composed by five chapters, has been conceptually divided into two differentiated parts. On one hand, the **theoretical background** section, which is composed by **Chapter 2** and **Chapter 3**, begins with a historical review compiling the previous antecedents in the literature. Then, the specific techniques useful to follow the discursion of the presented work are discussed in detail. These methodological chapters follow a conducting thread throughout theoretical methods, computational tools and statistical techniques leading to the comprehension of the ultimate purpose: the understanding of the nature manifested in the mechanistic and chemical behaviour of any particular system. On the other hand, the **Chapter 4**, devoted to **applications and results**, deepens into specific practical applications, showing the framework of the theory. Exceptionally, after the concluding remarks, an annex with the printouts of various contributions is appended.

Next, the **content of the chapters** is briefly described.

■ **Chapter 1. Introduction.**

This preliminary chapter attempts to overlook the driving insights of this work, as well as to present a brief sketch that will guide the reader throughout the development of the dissertation.

■ **Chapter 2. Quantum Similarity Theory.**

This is an introductory chapter were the main concepts and definitions related to quantum similarity theory are presented. This chapter begins with a historical revision, the mathematical formulation, and the type of quantum similarity measures, followed by a brief enumeration of the multiple applications derived from the seminal idea. Without deepening into computational facets, the section supplies a clear perception of the methodologies employed in the calculation of Molecular Quantum Similarity Measures (MQSM).

In the second section of the chapter, the basic theoretical background of the topological approach, based on graph theory, is also sketched. From this basis, the connection between quantum similarity theory and the classical topological approach is derived, and the resulting Topological Quantum Similarity Indices (TQSI) are properly defined.

■ **Chapter 3. QSAR Analysis.**

This episode presents the new Computer-Aided Molecular Design (CAMD) advances and highlights the recent development of Quantitative Structure-Activity Relationships (QSAR) techniques. First, it introduces the historical birth and development of QSAR techniques. The QSAR field has significantly evolved since its qualitative origins, to the actual three-dimensional and higher dimensional models, going through the linear free energy relationships, the Hansch analysis, and the QSAR based on topological descriptors. In addition, this section also describes the generation of descriptors, the statistical treatment of Similarity Matrices (SM), and the validation of results. The most common chemometric techniques are also emphasised; among them, the multivariate dimensionality reduction and statistical validation techniques. The statistical tools are not exclusive from QSAR but they can be applied to any data set. In any case, the objective is to build a mathematical model relating the molecular descriptors with the experimental data, namely the biological activity.

■ **Chapter 4. Applications of Quantum Similarity Measures to QSAR.**

This chapter illustrates some technical aspects of the process. Here, a exposition and analysis of the obtained results by employing similarity measures as a source of molecular descriptors, as well as the influence of several factors in the QSAR model, are discussed. Also the chronological synthesis of the evolution suffered in the different procedures involved in the QSM calculation is presented. Here, practical application examples are displayed, some of them complemented by the contributions annexed at the end of the memory. They cover several fields of interest, i.e. pharmacology, toxicity, and property studies.

■ **Chapter 5. Conclusions.**

Finally, the last chapter includes the final considerations and summarises the concluding remarks, together with future perspectives. Besides, a succinct enumeration of the published contributions in journals, and poster and oral communications, is listed.

■ **Annex. Contributions.**

The printouts of three published papers have been annexed as an appendix.

3 ACKNOWLEDGEMENTS

Finally, I would like to acknowledge my advisor, Professor Ramon Carbó-Dorca, for the given opportunity to develop this thesis in his group, and specially Dr. Xavier Gironés, who has been my unofficial second supervisor from the very beginning. I also thank the rest of the former members of EMQ, who introduced me in the quantum chemistry world, and all the members of the IQC, who stand by me day after day and support me in the good and bad times, creating a warm working atmosphere.

Besides, I would also like to acknowledge the opportunity to collaborate with the Professors who kindly received me in their labs. The temporary research stays with Prof. Robert Ponec (in Prague), Prof. Mark Cronin (in Liverpool), and Andreas Savin (in Paris) have been not only enlightening but also productive.

Finally, I should also acknowledge the financial support, a non-negligible requirement to devote almost a lustrum to scientific research. This Ph. D. has been mainly supported by a pre-doctoral fellowship from the *Ministerio de Ciencia y Tecnología*.

Ana Gallegos Saliner
Girona, June 2004

Introduction

There is no reason why anyone
would want to have a computer in their home

K. Olson
President of Digital Equipment Corp., 1977

1 GENERAL INTRODUCTION

Computer-Aided Molecular Design (CAMD), together with computational chemistry, is a relatively new discipline of chemistry with outstanding projection. The possibility to virtually design new useful compounds with well-defined properties reducing the high costs of experimental synthesis has recently promoted the investment in theoretical research.

Nowadays, CAMD methods are of special relevance in the rationalization of the discovery of compounds with specific pharmaceutical properties and drug research and development. Hence, the effective design of chemical structures with the desirable therapeutic properties is directed towards **Computer Aided-Drug Design (CADD) [**1-5], a well established area of CAMD [6], among others. These techniques comprise new methodologies, such as molecular modelling, computer simulation, and the recently rationalized and systematized discipline of **Quantitative Structure-Activity Relationships (QSAR).**

The main applications of CAMD are the elucidation of the basic requirements for a compound to elicit a determined activity, the simulation of the binding between a ligand and the receptor, the proposal of new mechanisms to comprehend biological processes, the prediction of chemical reactivity, the discovery of new active principles or prototypes, the screening for active lead compounds, and the prediction of activities for non-synthesised analogues. These applications convert CAMD in a highly suitable tool to be used in molecular design, and, in particular, in drug design.

The process to synthesise new drugs usually implies first of all the discovery of the potential active principle of interest. Once identified the candidate structure, there is a research of analogue compounds with the optimal desired properties, which improve the biological activity and the pharmacokinetic characteristics, and simultaneously diminish the secondary effects and the toxicity. The biological phases include animal and human testing for the activity, specificity, bioavailability, lack of toxicity, and, also, medical need, manufacturing requirements, and market potential, among others. The drug discovery process, founded in error-prone methods, has the attached inconvenient of high costs in time and money.

In recent years, pharmaceutical companies have complemented the conventional drug discovery technologies by **rational drug design**. Thus, trial-and-error chemistry-based discovery methods have been assisted by computer-based techniques, automated assays, and other advanced systems. Advances in combinatorial chemistry, and biotechnologies such as genomics and proteomics, have improved the productivity of the research and development processes. Integrated disciplines are synergistically applied to the drug discovery process in interdisciplinary projects.

Illustratively, **combinatorial chemistry** analyses enormous libraries of millions of compounds by means of **High Throughput Screening (HTS)** methods**.** HTS screens large numbers of compounds selected from a library against a biological target, id est. a protein playing a fundamental role for a particular disease. Afterwards, data mining techniques identify novel valid patterns in the data, potentially useful to analyse the data sets.

In the last decades, thanks to the improvements in computer speed and capacity, computer-based methodologies increased thousand-fold the number of lead compounds available for further research. But not only the number of viable drug candidates increased, but also the costs and time consumed in various drug discovery processes was dramatically reduced, improving the efficiency of the drug development.

However, only one in 5000 early drug candidates makes it through the discovery process. Besides, provided that new chemical entities that may potentially be turned into drugs need to be screened for toxicity and viability in human treatment applications and, most importantly, safety, only one out of ten compounds succeeds in clinical trials. Thus, although emerging technological advances have helped to increase the speed of the early stages of the drug discovery process, they have not been of much help in the trial stages, which are the most time-consuming and costliest parts of the process. For example, in 2003, the estimated costs of bringing one single new drug to market were estimated to have risen to 1200 million dollars [7].

One of the first approaches to reduce these costs were attempted by correlating the biological function of a compound with its chemical structure, expressed in terms of molecular structural descriptors, by means of the so-called QSAR techniques.

Within the QSAR approach, the descriptor variables are not physically measured but computed; hence, they are easy and cheap to generate even for large molecular sets. QSAR analysis attempts to build a mathematical equation that models the behaviour for a series of compounds, in order to provide insight into structure-function or structure-activity relationships. One of the objectives is the comprehension of the factors influencing the behaviour of a set of structurally-related compounds and, once obtained the relationship, the activity for untested compounds of interest is predicted by extrapolation. In particular, in the pharmacological context, QSAR serve to reject and identify the best drug candidates for toxicity and clinical experiments, or, more generally, to convert structural classes of compounds in potential drugs.

QSAR assume that the biological activity of a compound is a consequence of its chemical structure [8-9]. Based on the **similarity principle,** which states that similar molecules possess similar properties, QSAR allow generating descriptors for molecular structure. The similarity principle hypothesis requires a procedure to measure the molecular similarity, provided by **Molecular Quantum Similarity Theory**.

1.1 Molecular Quantum Similarity (MQS)

The notion of similarity is deeply attached to the human knowledge. In the quotidian perception of objects and even situations, unconscious associations in terms of similarity are continuously established by applying hidden criteria. The similarity concept is also tightly attached to science. The establishment of comparative measures between geometrical shapes was already proposed in the ancient Greece. Later, within the field of chemistry, the periodical table of elements [10] was founded in similarity criteria between atomic species.

However, the systematization of cognitive processes leading to the evaluation of similarity is not evident. In the chemistry domain, different proposals have attempted to answer the question: “how similar is a molecule to another?” Therefore, is crucial to have a sound definition of unbiased quantitative measures of molecular similarity. But the arbitrary definition of similarity is necessarily linked to the analysed molecular aspect.

In the nineteenth century, quantitative similarity measures derived from a topological point of view were established using the **chemical graph theory** [11-12]**,** yielding the so-called **Topological Indices (TI)** [13].

More recently, taking advantage of computational advances, molecules could be considered as objects ruled by quantum mechanics. The first quantitative measure of Molecular Quantum Similarity (MQS) between two molecules founded on quantum mechanical concepts was formulated by Carbó [14]. Since any quantum system can be completely characterised by a wave function resulting from Schrödinger equation, Carbó proposed a numerical comparative measure using the superposition of electronic charge densities of two molecules as a comparing source. This significant definition resulted in the so-called **Molecular Quantum Similarity Measures (MQSM)**, calculated as the volume integral between the corresponding **Density Functions (DF)** of the two compared objects, weighted by a non-differential positive definite operator, namely the quantum similarity operator. The global set of quantum similarity measures, which compares all the possible pairs of quantum objects of the system, is expressed in matrix form. MQSM constitute a simple way to obtain relationships between the compared quantum objects, by identifying the characteristics of electronic density that vary from a system to another.

This definition is not only still active but have also been developed in the IQC by Carbó and collaborators, emphasizing its mathematical meaning [15-17], reviewed in several reported works [18-22], and extensively applied to the resolution of problems of different nature [23-25]. This research line has also been pursued by other investigation groups. Among them, the work of Cioslowski [26], Allan and Cooper [27], and Richards [28-29], are particularly relevant. In addition, from a different perspective, Herndon [30-31] substituted the quantum mechanical magnitudes by elements of graph theory and topology. This approach was followed by Mezey [32-33], Ponec [34-41], and other authors [42-44].

Illustratively, the interest in the topic has stimulated the publication of different monographic reviews [45-51] and the biennial international congress *Girona Seminar on Molecular Similarity (GSMS)* [52-54], forum of discussion of the advances in the matter.

1.2 Quantitative Structure-Activity Relationships (QSAR)

The QSAR methods include all the techniques with the objective to establish empirical or theoretical patterns for the behaviour of biologically active families of compounds, with the aim to efficiently acquire the optimal activity. The QSAR analysis presupposes a relationship between the properties of a molecule and its structural characteristics, and attempts to build simple mathematic models to describe the biological or chemical behaviour for a set of compounds [55-57].

QSAR techniques are applied upon analogues of an active principle with an experimentally determined activity. Usually, these compounds are formed by a common pattern and variable substituents or fragments. Then, the molecular descriptors defined to characterise the molecular structural features of the series are stored in a matrix of molecular descriptors. The column vectors of the matrix act as the independent variables in the correlation equation that relates them with the vector of experimental biological properties.

The QSAR analysis comprises as well the definition of molecular descriptors as the statistical techniques used to treat these descriptors. On the other hand, **chemometrics** provides the statistical tools required to build the mathematical models and to enhance their predictive capacity [58-59]. Usually, the model is validated using members excluded from the training series and comparing the predicted values with the experimental ones. The subsequent phase is the prediction of the activity for non-synthesised products with the aim to distinguish the active analogues from the non-active ones, and to improve the effectiveness of the process.

The increasing importance of the QSAR topic has resulted in several specialised journals (*Journal of Medicinal Chemistry*; *Quantitative Structure-Activity Relationships*; *Journal of Computer-Aided Molecular Design*; *Journal of Molecular Modelling*; *SAR and QSAR in Environmental Research*), monographic books [60-62], and congresses (*European QSAR symposium*; *International Workshop on QSAR's in the Human Health and Environmental Sciences*; *Techniques in Pharmacophore Development*), and associations (*The QSAR and Modelling Society*).

1.3 Molecular Similarity in QSAR

Once implemented the QSAR techniques, the molecular similarity was considered as a valid tool to generate molecular descriptors [63-65]. In our lab, the so-called **Similarity Matrices (SM)**, obtained by means of similarity integrals between density functions of the molecules being compared have been used as the source of molecular descriptors. The first application of molecular similarity in QSAR dates back from 1983 [66]. Since then, similarity measures have been used in QSAR studies [67-71], either within the classical formulation of Carbó Indices [72-73], topological similarity indices [74-75], or other [76-80].

In our lab, the application of quantum similarity to QSAR was initially qualitative. To such an extent, the representation of quantum objects was defined [81-82], and physico-chemical properties were associated with spatial groupings [83-85]. Afterwards, the connection between the expected value of a physical observable and quantum similarity measures was described [86]. The practical implementation of this work resulted in the theoretical protocol of application for quantum similarity matrices in QSAR [87-89], together with different illustrative examples, either in QSTR [90-92], QSPR [93-95] or QSAR of pharmacological interest [96-102].

The main drawback of the practical application of MQSM to QSAR analysis is the number and size of compounds. Normally these studies deal with big molecular sets formed by large molecules [104-106] difficult to analyse at an *ab initio* level [107-108]. Indeed, in the MQSM definition there is the implicit alignment of the compounds [109-110]. In particular, the research of the optimal molecular superposition may be the computational bottleneck of the process because it may imply the repeatedly computation of similarity measures [111]. Some measures were adopted to alleviate these problems. On one hand, the **Atomic Shell Approximation (ASA)** [112-117], developed by Constans [120] and optimised by Amat [121-122], computes approximated molecular electronic density functions. On the other hand, optimization techniques used for the pairwise molecular superposition are based either in the **maximal similarity alignment rule**, implemented by Constans [123-124], and the **Topo-Geometrical Superposition Algorithm (TGSA)**, by Gironés [125-126].

2 OVERVIEW

One of the principal research axes of the EMQ group has been devoted to **molecular similarity**, not only from its foundation, but also from the seminal independent work of Professor Ramon Carbó-Dorca, who envisaged the first formal definition of a quantitative measure of molecular similarity [14]. Thus, the research group has been historically implicated in all the aspects related to molecular similarity, from conceptual foundations based on theoretical quantum chemistry [22,127-129] to specific applications developed for particular study cases of interest, going through the development of new methods [130-138], id est., the computational evolvement and the methodological implementation of statistical [139] and mathematical tools [140-149] yielding novel software [22,150], and the systematizing of application procedure protocols using different approaches [25,87,132,151-152]. To such an extent, with the leitmotif of molecular similarity as a wrapping background, this work harnesses several previous contributions of molecular similarity that constitute its precedents, following the initial investigation line opened in the master research project, entitled *Application of Molecular Quantum Similarity Measure to the study of Quantitative Structure-Activity Relationships.*

In particular, several single motivations have impelled the different studies compiled in this memory. But the resolution of specific problems has a unifying objective, focused on the **application of molecular similarity concepts in QSAR analysis**, with the aim to develop QSAR models for specific study cases pharmacologically relevant.

Within the contextualised frame of molecular similarity, the object of study in every case leaded to the differentiation of the general QSAR into the separated branches of QSTR, and QSPR, referred to toxicological or property studies, respectively, depending on the considered investigation case. For the sake of simplicity, these variants will be generally named as QSAR throughout this thesis.

It must be noted that two different representations of molecules have been used, in regard to the treatment of the density function. In the more general case, molecules are described by means of the global density function and similarity measures between all the possible pairs of compounds of the series were computed. In this case, the superposition process and conformational analysis have special relevance. The second approximation is founded on the partition of molecules into fragments. Here, the electronic densities used to define the molecular fragments, were used to calculate similarity measures defined upon the fragments.

In particular, the most relevant contributions of this thesis to all this enginery can be succinctly outlined in chronological order. The track of the outdoors collaborations carried out in temporary research stays contextualises the evolution of the present research memory.

■ **Application of the Classical MQSM protocol,** which had already been elaborated in 2000, at the beginning of the Ph. D., by the preceding EMQ group. Conventional Molecular Quantum Similarity Measures (MQSM) were computed by employing the preexisting software [64-65,89-90], implemented by Robert [117] and Gironés [118], by using the Atomic Shell Approximation (ASA) [119-123] for the electronic densities, and the maximum similarity alignment rule [124] or the Topo-Geometrical Superposition Algorithm (TGSA) [125] for the molecular superposition.

In this initial phase, the work fructified in a QSTR study on polycyclic aromatic hydrocarbons, and two QSAR models for the antimalarial activity on series of 1,2,4-trioxanes, cyclic peroxy ketals, and artemisinin derivatives.

►Gallegos, A.; Robert, D.; Gironés, X.; Carbó-Dorca, R. Structure-Toxicity Relationships of Polycyclic Aromatic Hydrocarbons using Molecular Quantum Similarity. *J. Comput.-Aid. Mol. Des., 15(1),* **2001**, 67-80.

► Gironés, X.; Gallegos, A.; Carbó-Dorca, R. Antimalarial Activity of Synthetic 1,2,4-Trioxanes and Cyclic Peroxy Ketals, a Quantum Similarity Study. *J. Comput.-Aid. Mol. Des., 15(12),* **2001**, 1053-1063.

► Gironés, X.; Gallegos, A.; Carbó-Dorca, R. Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR. *J. Chem. Inf. Comput. Sci, 40,* **2000**, 1400-1407.

■ **Implementation and Computational Development of Topological Quantum Similarity Measures (TQSM)** materialised in the TOPO program. This software was evolved together with Gironés, and taking as precedent the former work of Besalú [153] and Lobato [154153-157]. The definition uses the so-called **Topological Quantum Similarity Indices (TQSI)** resulting from **Atomic Quantum Similarity Measures (AQSM),** as a source of descriptors [158-161].

The calculation of TQSI only depends on the molecule; thus, the alignment process for each molecular pair was avoided. Here, taking into account the study cases used in the previous research line, several contributions were envisaged. QSAR studies for the carcinogenic power, the antimalarial activity, and the aquatic toxicity were performed for the molecular families considered in the preceding section. Also a third study applying TQSI to QSPR, QSTR, and QSAR problems o various sets is in progress.

► Gallegos, A.; Gironés, X.; Carbó-Dorca, R. TOPO. Institute of Computational Chemistry, University of Girona, Girona, **2000**.

► Besalú, E; Gallegos, A.; Carbó-Dorca, R. Topological Quantum Similarity Indices and Their Use in QSAR: Application to Several Families of Antimalarial Compounds. In *MATCH-Communications in Mathematical and in Computer Chemistry (Special issue dedicated to Prof. Balaban)*. Diudea, M.; Ivanciuc, O. (Eds.) *MATCH-Commun. Math. CO, 44,* **2001**, 41-64.

► Gallegos, A.; Gironés, X.; Carbó-Dorca, R. Topological Quantum Similarity Measures: applications in QSAR. In *Proceedings of the 5th GSMS*. Sen, K. (Ed.) Nova Press. *In press.*

► Gallegos, A.; Gironés, X. Topological Quantum Similarity Indices based on Fitted Densities: Theoretical Background and QSPR Application. *To be submitted.*

■ **Novel Application of Quantum Self-Similarity Measures (QS-SM)** in the partitioned study of molecules by **molecular fragments**. The application of this definition for each compound in series of chemicals with a single common backbone and different substituents, also avoided the costly bottleneck superposition process of aligning every two molecules. This research line is headed by the Professor **Robert Ponec** of the *Institute of Chemical Process Fundamentals* of the *Czech Academy of Sciences* in Prague, and has been synergistically developed by Amat [162-166] and Gironés [167-168]. In particular, the collaboration starting from a temporary research stay in the winter 2001 is nowadays still active.

The conjunct work resulted in a contribution applied on the study of several sets of anti-tuberculotic benzoxazines.

► Gallegos, A.; Carbó-Dorca, R.; Ponec, R., Waisser, K. Similarity approach to QSAR. Application to antimycobacterial benzoxazines. *Int. J. Pharm., 269,* **2004**, 51-60.

These families were independently examined by means of combinatorial chemistry procedures. Using COMBINATOR [169], a virtual molecular library of 3D benzoxazines was generated [170]. The computer code calculates the series of all the possible structures formed by the combination of the generation basis with all the various substituents, by placing different selected fragments as substituents at different molecular sites. This molecular generation basis consists of a substitution pattern defined by the substituents, the substitution sites, and the common backbone.

■ **Application of a QSAR study based on both experimental, quantum theoretical parameters, and physicochemical properties, with a pharmacological insight**. This study was carried out under the supervision of Professor **Mark Cronin** [171-172], from the *QSAR and Modelling Research Group* of the *School of Pharmacy and Chemistry*, in the Liverpool John Moores University, in the spring 2002.

In particular, the molecular series of study consisted of a semiquantitative analysis of estrogenic activity, where the experimentally measured biological activity had been classed into two discrete categories.

► Gallegos Saliner, A.; Amat, L.; Carbó-Dorca, R.; Schultz, T.W.; Cronin, M.T.D. Molecular Quantum Similarity Analysis of Estrogenic Activity. *J. Chem. Inf. Comput. Sci., 43*, **2003**, 1166-1176**.**

■ Finally, not specially related to the MQSM research field, the **Electron Localization Function (ELF)** [173-175] was used to investigate maximal probability domains for linear molecules [177-178]. Professor **Andreas Savin** directed this work, carried out in the fall 2002 and beginning 2003, in the *Laboratoire de Chimie Théorique* of the *Centre Nationale de Recherche Scientifique (CNRS)*, of the *Université Pierre et Marie Curie* of Paris.

► Gallegos-Saliner, A.; Carbó-Dorca, R; Lodier, F.; Cancès, E.; Savin, A. Maximal Probability domains in Linear Molecules. In, *Proceedings of the 6th GSMS*. Sen, K. (Ed.) *In press.*

■ As a future project, a collaboration with Professor **Patrick Bultinck** [135,179-180] from the University of Ghent to conceptualise, develop, and implement a new **Chiral Molecular Quantum Similarity Measure** has been envisaged. This new measure would intend to differentiate the similarities and dissimilarities between two chiral molecules [181-184].

REFERENCES

1. Martin, Y.C. *Quantitative Drug Design: A Critical Introduction*. Marcel Dekker: New York, **1978**.
2. Sanz, F.; Martín, M.; Pérez, J.; Turmo, J.; Mitjana, A.; Moreno, V.; Dearden, J.C. (Eds.) *Quantitative Approaches to Drug Design*. Elsevier: Amsterdam, **1983**.
3. Franke, R (Ed.) *Theoretical Drug Design Methods*. Elsevier: Amsterdam, **1984**.
4. Codding, P.W. (Ed.) *Structure-based drug design: experimental and computational approaches*; *Vol. 352*NATO ASI Series: Dordrecht, **1998**.
5. Levy, M.D. The drug discovery and development process in the new millennium. *Journal of the Canadian Association of Gastroenterology, 14 (7),* **2000**.
6. Richards, W.G. *Computer-Aided Molecular Design*. IBC Technical Services: London, **1989**.
7. Of Enabling Technologies and Innovative Strategies in the Pharmaceutical Industry. Industry News in Europe, KHIDI Europe, May 31, **2002**.
8. Klopman, G. Artificial Intelligence Approach to Structure-Activity Studies. Computer Automated Structure Evaluation of Biological Activity of Organic Molecules. *American Chemical Society; Vol. 106(24),* **1984**.
9. Livingstone, D. *Data Analysis for Chemists*. Oxford University Press: **1995**.
10. Mendeleev, D.I. *Principles of Chemistry*; *Vol. 2,* **1868–71,** tr. **1905**.
11. Basak, S.C.; Niemi, G.J.; Veith, G.D. *Computational Chemical Graph Theor.* Rouvray, D.H. (Ed.) Nova Science Publishers: New York, **1990**, 201.
12. Trinajstić, N. *Chemical Graph Theory*. CRC Press: Boca Raton, **1992**.
13. Basak, S.C.; Grunwald, G.D.; Niemi, G.J. Use of graph theoretical and geometrical molecular descriptors in structure-activity relationships. In *From Chemical Topology to Three Dimensional Molecular Geometry.* Balaban, A.T.(Ed.). Plenum Press: **1997,** 73-116.
14. Carbó, R.; Leyda, L.; Arnau, M. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem., 17,* **1980**, 1185-1189.
15. Carbó, R.; Besalú, E. Definition, Mathematical Examples and Quantum Chemical Applications of Nested Summation Symbols and Logical Kronecker Deltas. *Comp. & Chem., 18,* **1994**, 117-126.
16. Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets, Vector Semiespaces and Convex Sets, QSM and ASA density functions, Diagonal Vector Spaces and Quantum Chemistry. In *Advances in Molecular Similarity; Vol. 2.* JAI Press, **1998**, 43-72.
17. Carbó-Dorca, R. Tagged sets, convex sets and quantum similarity measures. *J. Math. Chem., 23,* **1998**, 353-364.
18. Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and Recent Developments on Molecular Quantum Similarity. *Topics Curr. Chem., 173,* **1995**, 31-62.
19. Carbó-Dorca, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum Molecular Similarity Measures: Concepts, Definitions and Applications to Quantitative Structure-Property Relationships. In *Advances in molecular similarity. Vol. 1.* Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: Greenwich, CT, **1996**, 1–41.
20. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity; Vol. 2.* Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: Greenwich, **1998**, 1-42.
21. Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struc. (Theochem), 451,* **1998**, 11-23.
22. Besalú, E.; Gironés, X.; Amat, L.; Carbó-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res., 35,* **2002**, 289-295.
23. Carbó, R.; Calabuig, B. Quantum Molecular Similarity Measures and the Normal-Dimensional Representation of a Molecular Set - Phenyldimethylthiazines. *J. Mol. Struc. (Theochem), 86,* **1992**, 517-531.
24. Carbó, R.; Calabuig, B. Quantum Similarity: Definitions, Computational Details and Applications. In *Computational Chemistry: Structure, Interactions and Reactivity*. Fraga, S. (Ed.) Elsevier: Amsterdam, **1992**.
25. Amat, L.; Besalú, E.; Carbó, R.; Fradera, X. Practical applications of quantum molecular similarity measures (QMSM): programs and examples. *Sci. Gerun., 21,* **1995,** 127-143.

26. Cioslowski, J.; Fleishmann, E.D. Assessing molecular similarity from results of *ab initio* electronic structure calculations. *J. Am. Chem. Soc., 113,* **1991**, 64-67.
27. Cooper, D.L.; and Allan, N.L. Bond formation in momentum space. *J. Chem. Soc. Faraday Trans., 83,* **1987**, 449-460.
28. Bowen-Jenkins, P.E.; Cooper, D.L.; Richards, W.G. Ab initio computation of molecular similarity. *J. Phys. Chem., 89,* **1985**, 2195-2197.
29. Richards, W.G. Molecular similarity and dissimilarity. In *Modelling of biomolecular structures and mechanisms*. Pullman, A. (Ed.) Kluwer: The Netherlands, **1995**.
30. Herndon, W.C.; Bertz, S.H. Linear notations and molecular graph similarity. *J. Comput. Chem., 8,* **1987**, 367-374.
31. Herndon, W.C. Graph codes and a definition of structural similarity. *Comput Math. Applic., 15,* **1988**, 303-309.
32. Mezey, P.G. Shape group studies of molecular similarity: shape groups and shape graphs of molecular contour surfaces. *J. Math. Chem., 2,* **1988**, 299-323.
33. Mezey, P.G. Shape in chemistry: an introduction to molecular shape and topology. VCH: New York, **1993**.
34. Ponec, R. Topological Aspects of Chemical-Reactivity - on the Similarity of Molecular-Structures. *Collect. Czech. Chem. Commun., 52,* **1987**, 555-562.
35. Ponec, R. Similarity Measures, the Least Motion Principle and Selection-Rules in Chemical-Reactivity. *Z. Phys. Chem-Leipzig, 268,* **1987**, 1180-1188.
36. Ponec, R.; Strnad, M. Similarity Approach to Chemical-Reactivity - Specificity of Multibond Processes. *Collect. Czech. Chem. Commun., 55,* **1990**, 2583-2589.
37. Ponec, R.; Strnad, M. Topological Aspects of Chemical-Reactivity - Evans-Dewar Principle in Terms of Molecular Similarity Approach. *J. Phys. Org. Chem., 4,* **1991**, 701-705.
38. Ponec, R.; Strnad, M. Electron Correlation in Pericyclic Reactivity - a Similarity Approach. *Int. J. Quantum Chem., 42,* **1992**, 501-508.
39. Ponec, R.; Strnad, M. Similarity Ideas in the Theory of Pericyclic Reactivity. *J. Chem. Inf. Comp. Sci., 32,* **1992**, 693-699.
40. Ponec, R.; Strnad, M. Position Invariant Index for Assessment of Molecular Similarity. *Croat. Chem. Acta, 66,* **1993**, 123-127.
41. Ponec, R. Similarity Approach to Chemical-Reactivity - a Simple Criterion for Discriminating between One-Step and Stepwise Reaction-Mechanisms in Pericyclic Reactivity. *J. Chem. Inf. Comp. Sci., 33,* **1993**, 805-811.
42. Tsai, C.C.; Johnson, M.A.; Nicholson, V.; Naim, M. A topological approach to molecular similarity analysis and its application. In *Graph Theory and Topology in Chemistry*. King, R.B.; Rouvray, D.H. (Eds.) Elsevier: Amsterdam, **1987**; 231-236.
43. Takahashi, Y.; Sukekawa, M.; Sasaki, S.-I. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 639-643.
44. Uraguchi, R.; Sato, Y.; Nakayama, A.; Sukekawa, M.; Iwataki, I.; Ber, P.; Wakabayashi, K. Molecular Shape Similarity of Cyclic Imides and Protoporphyrinogen-IX. J. *Pesticide Sci., 22,* **1997**, 314-320.
45. Willett, P. *Similarity and Clustering in Chemical Information Systems*. Wiley: New York, **1987**.
46. Johnson, M.A.; Maggiora, G. (Eds.) *Concepts and Applications of Molecular Similarity*. John Wiley & Sons Inc.: New York, **1990**.
47. Carbó, R. (Ed.) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*. Kluwer: Amsterdam, **1995**, 3-30.
48. Dean, P.M. (Ed.) *Molecular Similarity in Drug Design*. Blackie Academic: London, **1995**.
49. Carbó-Dorca, R.; Mezey, P.G. (Eds.) *Advances in Molecular Similarity.* JAI Press: Greenwich, **1996**; *Vol. 1* & **1998**; *Vol. 2*.
50. Kubinyi, H.; Martin, Y.C.; Folkers, G. (Eds.) *3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity (Quantitative Structure-Activity Relationships); Vol.2.* Kluwer: Dordrecht, **1998**.
51. Carbó-Dorca, R.; Gironés, X.; Mezey, P.G. (Eds.) *Fundamentals of molecular similarity*. Kluwer Academic/Plenum Press: New York, **2001.**
52. Webpage of the IV Girona Seminar on Molecular Similarity: http://iqc.udg.es/gsms99/ [last accessed 28 May 2004].
53. Webpage of the IV Girona Seminar on Molecular Similarity: http:// iqc.udg.es/gsms2001/ [last accessed 28 May 2004].

54. Webpage of the IV Girona Seminar on Molecular Similarity: http://stark.udg.es/gsms2003/ [last accessed 28 May 2004].
55. Eriksson, L.; Jaworska, J.; Worth, A.P.; Cronin, M.T.D.; McDowell, R.M.; Gramatica, P. Methods for reliability and uncertainty assessment and applicability evaluations of classification-regression-based and QSARs. *Environ. Health Perspect., 111*, **2003**, 1361-1375.
56. Schultz, T.W.; Cronin, M.T.D.; Walker, J.D.; Aptula, A.O. Quantitative structure-activity relationships (QSARs) in toxicology: a historical perspective. *J. Mol. Struct. (Theochem), 622,* **2003,** 1-22.
57. Schultz, T.W.; Cronin, M.T.D.; Netzeva, T.I. The present status of QSAR in toxicology. *Journal of Molecular Structure (Theochem) 622*, **2003,** 23-38.
58. Cronin, M.T.D. Computational methods for the prediction of drug toxicity. *Curr. Opinion in Drug Discovery and Development*, *3*, **2000,** 292-297.
59. Cronin, M.T.D.; Schultz, T.W. Evaluation of mechanistic and statistical QSARs for toxicity: survival of the (statistically) fittest? In *Designing Drugs and Crop Protectants: Processes, Problems and Solutions*. Ford, M.; Livingstone, D.; Dearden, J.; van de Waterbeemd, H. (Eds.) Blackwell: Oxford, **2003**, 294-296.
60. Kubinyi, H. (Ed.) *3D QSAR in Drug Design*. Kubinyi, H. (Ed.) ESCOM: Leiden, **1993**.
61. Sanz, F.; Giraldo, J.; Manaut, F. (Eds.) *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Aplications. Proceedings of the 10th European Symposium on Structure-Activity Relationships, QSAR and Molecular Modelling***.** Prous Science Publishers: Barcelona, **1995**
62. Bultinck, P.; De Winter, H.; Langenaeker, W.; Tollenaere, J. P. (Eds.) *Computational Medicinal Chemistry for Drug Discovery*. Marcel Dekker: New York, **2004**
63. Gironés, X.; Amat, L.; Carbó-Dorca, R. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res.*, *10*, **1999**, 545-556.
64. Carbó-Dorca, R.; Robert, D.; Amat, L.; Gironés, X.; Besalú, E. Molecular Quantum Similarity in QSAR and Drug Design. In *Springer Lecture Notes in Chemistry; Vol. 7.* Springer: **2000.**
65. Robert, D.; Gironés, X.; Carbó-Dorca, R. Molecular Quantum Similarity Measures as descriptors for Quantum QSAR. *Polycycl. Aromat. Comp., 19,* **2000**, 51-71.
66. Martín, M.; Sanz, F.; Campillo, M.; Pardo, L.; Pérez, J.; Turmo, J.; Aulló, J.M. Quantum chemical structure-activity relationships on b-carbolines as natural monoamine oxidase inhibitors. *Int. J. Quant. Chem., 23*, **1983**, 1643-1652.
67. Sanz, F.; Martín, M.; Lapeña, F. and Manaut, F. Quantitative structure-activity relationships on MAO substrates by means of quantum chemical properties. *Quant. Struct.-Act. Relat., 5*, **1986**, 54-57.
68. Sanz, F.; Manaut, F.; José, J.; Segura, J.; Carbó, M. and De la Torre, R. Automatic determination of MEP patterns of molecules and its application to cafein metabolism inhibitors. *J. Mol. Struct. (Theochem), 170,* **1988**, 171-180.
69. Luque, F.J.; Sanz, F.; Illas, F.; Pouplana, R.; Smeyers, Y.G. Relationships between the activity of some H2-receptor agonists of histamine and their *ab initio* molecular electrostatic potential (MEP) and electron density comparison coefficients. Eur. *J. Med. Chem., 23,* **1988**, 7-10.
70. Sanz, F.; Manaut, F.; Dot, T.; López de Briñas, E. Complete or partial comparison of molecular electrostatic potential distributions? Some tests with 5-HT ligands. *J. Mol. Struct. (Theochem), 256,* **1992**, 287-293.
71. Sanz, F.; Manaut, F.; Rodriguez, J.; Lozoya, E.; Lopez de Briñas, E. MEPSIM: A computacional package for analysis and comparison of molecular electrostatic potentials. *J. Comput.-Aided Mol. Design*, *7,* **1993**, 337-347.
72. Burt, C.; Richards, W.G.; Huxley, P. The application of molecular similarity calculations. *J. Comput. Chem., 11,* **1990**, 1139-1146.
73. Richard, A. M. Quantitative comparison of molecular electrostatic potentials for structure activity studies. *J. Comput. Chem., 12,* **1991**, 959-969.
74. Rum, G.; Herndon, W.C. Molecular similarity concepts. 5. Analysis of steroid-protein binding constants. *J. Am. Chem. Soc., 113*, **1991**, 9055-9060.
75. Herndon, W. C.; Rum, G. Three-Dimensional Topological Descriptors and Similarity of Molecular Structures: Binding Affinities of Corticosteroids. In *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Aplications. Proceedings of the 10th European Symposium on Structure-Activity Relationships, QSAR and Molecular Modelling***.** Sanz, F.; Giraldo, J.; Manaut, F. (Eds.) Prous Science Publishers: Barcelona, **1995**, 380-384.

76. Good, A.C.; Hodgkin, E.E.; Richards, W.G. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 188-191.
77. Good, A.C.; So, S.S.; Richards, W.G. Structure-activity relationships from molecular similarity matrices. *J. Med. Chem., 36*, **1993**, 433-438.
78. Good, A.C.; Peterson, S.J.; Richards, W.G. QSAR's from similarity matrices. Technique validation and application in the comparison of diferent similarity evaluation methods. *J. Med. Chem., 36*, **1993**, 2929-2937.
79. Good, A.C.; Richards, W.G. Rapid Evaluation of Shape Similarity Using Gaussian Functions. *J. Chem. Inf. Comput. Sci., 33,* **1993**, 112-116.
80. Good, A.C.; Richards, W.G. The extension and application of molecular similarity to drug design. *Drug Information Journal, 30*, **1996**, 371-388.
81. Carbó, R.; Calabuig, B. Molecular Quantum Similarity Measures and N-Dimensional Representation of Quantum Objects. 1. Theoretical Foundations. *Int. J. Quantum Chem., 42*, **1992**, 1681-1693.
82. Carbó, R.; Calabuig, B. Molecular Quantum Similarity Measures and N-Dimensional Representation of Quantum Objects. 2. Practical Applications. *Int. J. Quantum Chem.*, *42*, **1992**, 1695-1709.
83. Carbó, R.; Calabuig, B. Quantum Similarity Measures, Molecular Cloud Description; Structure Properties Relationships. *J. Chem. Inf. Comp. Sci.*, *32*, **1992**, 600-606.
84. Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular Quantum Similarity: Theoretical Framework, Ordering Principles and Visualization Techniques. *Adv. Quantum Chem.*, *25*, **1994**, 253-313.
85. Besalú, E.; Amat, L.; Fradera, X.; Carbó, R. An application of the molecular quantum similarity: ordering of some properties of the hexanes. In *QSAR and molecular modelling: concepts, computational tools and biological applications*. Sanz, F.; Manaut, M. (Eds.) Prous Science: Barcelona, **1995**, 396-399.
86. Carbó, R.; Besalú, E.; Amat, L and Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards theoretical foundation of quantitative structureproperties relationship. *J. Math. Chem.*, *18,* **1995**, 237-246.
87. Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.*, *16,* **1997**, 25-32.
88. Fradera, X. *Aplicacions de la semblança quàntica molecular en QSAR*. Master research project. Institute of Computational Chemistry: Girona, **1996.**
89. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum Molecular Similarity: Theory and Applications to the Evaluation of Molecular Properties, Biological Activities and Toxicity. In, *Fundamentals of molecular similarity*. Carbó-Dorca, R.; Gironés, X.; Mezey, P.G. (Eds.) Kluwer Academic/Plenum Press: New York, **2001**, 187-320.
90. Robert, D.; Carbó-Dorca, R. Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR & QSAR Environ. Res., 10,* **1999**, 401-422.
91. Renners, I.; Carbó-Dorca, R.; Grauel, A.; Ludwig, L.A.; Robert, D.; Gironés, X. Toxicity Prediction by Using Genetically Optimized B-Spline Networks Based on Molecular Quantum Similarity. 2nd Int. ICSC Symposium on Neural Computation (CD-ROM). Berlin, **2000**.
92. Renners, I.; Ludwig, L.A.; Grauel, A.; Benfenati, E.; Pelagatti, S.; Robert, D.; Carbó-Dorca, R.; Gironés, X. Modeling toxicity with molecular descriptors and similarity measures via B-Spline networks. IPMU2000, 8th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge Based Systems. Madrid, **2000**, 1021-1026.
93. Robert, D.; Amat, L.; Carbó-Dorca, R. 3D QSAR from tuned molecular quantum similarity measures: Prediction of the CBG binding affinity for a steroids family. *J. Chem. Inf. Comp. Sci., 39,* **1999**, 333-344.
94. Gironés, X.; Carbó-Dorca, R. Molecular Quantum Similarity-based QSARs for Binding Affinities of Several Steroid Sets. *J. Chem. Inf. Comp. Sci., 42,* **2002**, 1185-1193.
95. Gironés, X.; Carbó-Dorca, R. Using Molecular Quantum Similarity Measures under Stochastic Transformation to describe Physical Properties of Molecular Systems. *J. Chem. Inf. Comp. Sci., 42*, **2002**, 317-325.
96. Mestres, J.; Rohrer, D.C.; Maggiora, G.M. A Molecular Field-Based Similarity Approach to Pharmacophoric Pattern Recognition. *J. Mol. Graphics Mod., 15,* **1997**, 114-121.
97. Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned QSAR: An antitumoral family validation study. *J. Chem. Inf. Comp. Sci., 38,* **1998**, 624-631.

98. Robert, D.; Gironés, X.; Carbó-Dorca, R. Quantification of the influence of single-point mutations onthe Haloalkale dehalogenase activity: a molecular quantum similarity study. *J. Chem. Inf. Comp. Sci., 40,* **2000**, 839-846.

99. Amat, L.; Robert, D.; Carbó-Dorca, R Quantum Similarity QSAR: Study of inhibitors binding to thrombin, trypsina and factor XA, including a comparison with CoMFA and CoMSIA methods. *Int. J. Quantum Chem., 80,* **2000**, 265-282.

100. Fradera, X.; Knegel, R.M.A.; Mestres, J. Similarity-Driven Flexible Ligand Docking. *Proteins, 40*, **2000**, 623-636.

101. Fradera, X.; Cruz, X. de la; Silva, C.H.T.P.; Gelpi, J.; Luque, F.J.; Orozco, M. Ligand-induced changes in the binding sites of proteins. *Bioinformatics, 18,* **2002**, 939-948.

102. Rohrer, D.C.; Mestres, J. 3D Molecular similarity methods: Application to modeling HIV-1 reverse transcriptase inhibitor binding. In *Structure-based drug design: Experimental and computational approaches*; *Vol. 352.* Codding, P.W. (Ed.) NATO ASI Series E.: Dordrecht, **1995**, 211-222.

103. Duart, M.J.; Antón-Fos, G.M.; Julian-Ortiz, J.V.; Gozalbes, R.; Gálvez, J.; García-Domenech, R. Use of molecular topology for the prediction of physicochemical, pharmacokinetic and toxicological properties of a group of antihistaminic drugs. *Int. J. Pharm. 246*, **2002**, 111-119.

104. Mestres, J.; Solà, M.; Duran, M.; Carbó, R. On the Calculation of Ab-Initio Quantum Molecular Similarities for Large Systems - Fitting the Electron-Density. *J. Comput. Chem.*, *15*, **1994**, 1113-1120.

105. Gironés, X.; Carbó-Dorca, R.; Mezey, P.G. Application of Promolecular ASA Densities to Graphical Representation of Density Functions of Macromolecular Systems. *J. Mol. Graphics & Mod., 19,* **2001**, 343-348.

106. Gironés, X.; Amat, L.; Carbó-Dorca, R. Modeling Large Macromolecular Structures Using Promolecular Densities. *J. Chem. Inf. Comp. Sci., 42,* **2002**, 847-852.

107. Mestres, J.; Solà, M.; Duran, M.; Carbó, R. General suggestions and applications of quantum molecular similarity measures from ab initio fitted electron densities. In *Molecular similarity and reactivity: From quantum chemical to phenomenological approaches.* Carbó, R. (Ed.) Kluwer Acad.: Dordrecht, **1995**, 89-111.

108. Mestres, J. Solà, M.; Besalú, E.; Duran, M.; Carbó, R. Electron density approximations for the fast evaluation of quantum molecular similarity measures. In *Molecular similarity and reactivity: From quantum chemical to phenomenological approaches.* Carbó, R. (Ed.) Kluwer Acad.: Dordrecht, **1995**, 77-85.

109. Bultinck, P.; Carbó-Dorca, R. Quality of approximate Electron Densities and Internal Consistency of Molecular Alignment Algorithms in Molecular Quantum Similarity. Submitted for publication in *J. Chem. Inf. Comput. Sci*., **2003**.

110. Bultinck, P.; Kuppens, T.; Gironés, X.; Carbó-Dorca, R. A consistent scheme for molecular alignment and Molecular Similarity based on Quantum Chemistry. *J. Chem. Inf. Comput. Sci.*, **2004**, in press.

111. Robert, D.; Carbó-Dorca, R. Anàlisi de procrustes i alineament molecular. *Sci. Gerun.*, *24*, **1999**, 175-181.

112. Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values *J. Chem. Inf. Comput. Sci.*, *35*, **1995**, 1046-1053

113. Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. Quantum Molecular Similarity Measures (QMSM) and the Atomic Shell Approximation (ASA). In *Advances in Molecular Similarity*; *Vol. 1.* Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: London, **1996**, 187-211.

114. Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: first order density fitting using Elementary Jacobi Rotations. *J. Comput. Chem.*, *18*, **1997**, 2023-2039.

115. Gironés, X.; Amat, L.; Carbó-Dorca, R. A comparative study of isodensity surfaces using ab initio and ASA density functions. *J. Mol. Graph. Model.*, *16*, **1998**, 190-196.

116. Amat, L.; Carbó-Dorca, R. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diammine-dichloroplatinum (II) complex as an application example. *J. Comput. Chem.*, *20*, **1999**, 911-920.

117. Robert, D. Semblança Quàntica en QSAR. Nous desenvolupaments i aplicacions. Master research project. Institute of Computational Chemistry, University of Girona, Girona, **1999**.

118. Gironés, X. Funcions Densitat i Semblança Molecular Quàntica: Nous Desenvolupaments i Aplicacions. Doctoral Thesis. Institute of Computational Chemistry, University of Girona, Girona, **2002**.

119. Amat, L.; Carbó-Dorca, R. Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation. *J. Chem. Inf. Comp. Sci., 40*, **2000**, 1188-1198.
120. Constans, P. Desenvolupament computacional de la semblança molecular quàntica. Doctoral Thesis. Institute of Computational Chemistry, University of Girona, Girona, **1997**.
121. Amat, L. Estructura computacional i aplicacions de la semblança molecular quàntica. Doctoral Thesis. Institute of Computational Chemistry, University of Girona, Girona, **2003**.
122. Amat, L. Estructura computacional de les mesures de semblança quàntica: programa MOLSIMIL-96. Master research project. Institute of Computational Chemistry, University of Girona, Girona, **1996**.
123. Amat, L.; Carbó, R.; Constans, P. Algorisme d'optimització global de les mesures de semblança quàntica molecular. *Sci. Gerun., 22*, **1996**, 109-121.
124. Constans, P.; Amat, L.; Carbó-Dorca, R. Toward a global maximization of the molecular similarity function: superposition of two molecules. *J. Comput. Chem., 18*, **1997**, 826-846.
125. Girones, X.; Robert, D.; Carbó-Dorca, R. TGSA: A molecular superposition program based on topo-geometrical considerations. *J. Comput. Chem.*, 22, **2001**, 255-263.
126. Gironés, X.; Carbó-Dorca, R. TGSA-Flex: Extending the capabilities of the Topo-Geometrical Superposition Algorithm to handle rotary bonds. *J. Comp. Chem., 25*, **2004**, 153-159.
127. Carbó, R.; Besalú, E. Theoretical Foundations of Quantum Molecular Similarity. In *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*. Carbó, R. (Ed.) Kluwer: Amsterdam, **1995**, 3-30.
128. Carbó-Dorca, R.; Besalú, E. Quantum Theory of QSAR. *Contribution to Science, 1*, **2000**, 399-422.
129. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum Mechanical Origin of QSAR: Theory and Alications. *J. Mol. Struct. (Theochem), 504*, **2000**, 181-228.
130. Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Triple Density Molecular Quantum Similarity Measures: a General Connection between Theoretical Calculations and Experimental Results. *Molec. Engineer., 2*, **1992**, 43-64.
131. Robert, D.; Carbó-Dorca, R. Analyzing the triple density molecular quantum similarity measures with the INDSCAL model. *J. Chem. Inf. Comp. Sci., 38*, **1998**, 620-623.
132. Robert, D.; Carbó-Dorca, R. On the extension of QS to atomic nuclei: Nuclear QS. *J. Math. Chem., 23*, **1998**, 327-351.
133. Robert, D.; Carbó-Dorca, R. Structure-property relationships in nuclei. Prediction of the binding energy per nucleon using a quantum similarity approach. *Nuovo Cimento A, 111*, **1998**, 1311-1320.
134. Girones, X.; Amat, L.; Robert, D.; Carbó-Dorca, R. Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput. Aid. Mol. Des., 14*, **2000**, 477-485.
135. Bultinck, P.; Carbó-Dorca, R. Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comp. Sci., 43*, **2003**, 170-177.
136. Bultinck, P.; Langenaeker, W.; Carbó-Dorca, R.; Tollenaere, J.P. Fast calculation of Quantum Chemical Molecular Descriptors from the Electronegativity Equalization Method. *J. Chem. Inf. and Comput. Sc., 42*, **2003**, 422-428.
137. Amat, L.; Carbó-Dorca, R.; Cooper, D. L.; Allan, N. L. Classification of reaction pathways via momentum-space and quantum molecular similarity measures. *Chem. Phys. Lett., 367*, **2003**, 207-213.
138. Amat, L.; Carbó-Dorca, R.; Cooper, D.L.; Allan, N.L.; Ponec, R. Structure-property relationships and momentum-space quantities: Hammett σ Constants. *Mol. Phys., 101*, **2003**, 3159-3162.
139. Carbó-Dorca, R. On the statistical interpretation of density functions: Atomic shell approximation, convex sets, discrete quantum chemical molecular representations, diagonal vector spaces and related problems. *J. Math. Chem., 23*, **1998**, 365-375.
140. Besalú, E.; Carbó, R.; Lobato, M. Operator expansions: Definition and molecular applications. *Sci. Gerun., 21*, **1995**, 153-163.
141. Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summation symbols and logical functions: Pedagogical artificial intelligence devices for formulae writing, sequential programming and automatic parallel implementation. *J. Math. Chem., 18*, **1995**, 37-72.
142. Carbó, R.; Besalú, E. Applications of Nested Summation Symbols to Quantum Chemistry: Formalism and Programming Techniques. In *Strategies and Applications in Quantum Chemistry*. Ellinger, Y.; Defranceschi, M. (Eds.) Kluwer Academic Publishers: Dordrecht, **1996,** 229-248.

143. Carbó-Dorca, R. Quantum QSAR and the eigensystems of stochastic quantum similarity matrices. *J. Math. Chem.*, *27*, **2000**, 357-376.

144. Carbó-Dorca, R. Stochastic transformation of quantum similarity matrices and their use in quantum QSAR (QQSAR) models. *Int. J. Quantum Chem.*, *79,* **2000**, 163-177.

145. Carbó-Dorca, R Quantum Quantitative Structure-Activity Relationships (QQSAR): A comprehensive discussion based on Inward Matrix Products, employed as a tool to find approximate solutions of strictly positive linear systems and providing a QSAR-Quantum Similarity Measures. In *Proceedings of Eccomas,* 2000.

146. Besalú, E.; Vera, L. On the optimal selection of principal components in QSPR studies. *J. Math. Chem*., *29,* **2001,** 21-34.

147. Carbó-Dorca, R. Inward Matrix Products:Extensions and Applications to Quantum Mechanical Foundations of QSAR. *J. Mol. Struct.Theochem, 537,* **2001**, 41-54.

148. Besalú, E.; Carbó-Dorca, R.; Karwowski, J. Generalized one-electron spin functions and self-similarity measures. *J. Math. Chem., 29,* **2001**, 41-45.

149. Carbó-Dorca, R.; Besalú, E. Fundamental Quantum QSAR (Q2SAR) Equation: Extensions, Nonlinear Terms and Generalizations within Extended Hilbert-Sobolev Spaces. *Int. J. Quantum Chem., 88,* **2002,** 167-182.

150. Besalú, E.; Carbó, R.; Duran, M.; Mestres, J.; Solà, M. MESSEM: A Quantum molecular similarity system of programs. In *Methods and techniques in computational chemistry (METECC-95*) Clementi, E. Corongiu, G. (Eds.) STEF: Cagliari, **1995**, 491-508.

151. Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des.*, *13,* **1999**, 597-610.

152. Robert, D.; Carbó-Dorca, R. General trends in atomic and nuclear quantum similarity measures. *Int. J. Quantum Chem*., *77*, **2000**, 685-692.

153. Besalú, E. Fonaments Teòrics i Aplicacions de la Semblança Quàntica. Doctoral Thesis. Institute of Computational Chemistry, University of Girona, Girona, **1996**.

154. Lobato, M. Definició i Aplicació a les Relacions Estructura-Activitat de nous Índexs Topològics derivats de la Semblança Molecular Quàntica. Master research project. Institute of Computational Chemistry, University of Girona, Girona, **1997**.

155. Lobato, M.; Besalú, E.; Carbó, R. Relacions estructura-propietat per un conjunt d'hidrocarburs a partir de nous descriptors tridimensionals de la Semblança Molecular. *Sci. Gerun., 22,* **1996,** 79-86.

156. Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, 16, 465-472.

157. Lobato, M.; Amat, L.: Besalú, E.: Carbó-Dorca, R. Estudi QSAR d'una familia de quinolones utilitzant índexs de semblança i índexs topològics de semblança. *Sci. Gerun*., *23*, **1998**, 17-27.

158. Besalú, E.; Carbó, R. Quantum Similarity Topological indices: definition, analysis and comparison with classical molecular topological indices. *Sci. Gerun*., *21*, **1995,** 145-152.

159. Cioslowski, J.; Stefanov, B.B.; Constans, P. Efficient Algorithm for Quantitative Assessment of Similarities among Atoms in Molecules. *J. Comput. Chem.*, *17,* **1996**, 1352-1358.

160. Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. On quantum molecular similarity measures (QMSM) and indices (QMSI). *J. Math. Chem.*, *19,* **1996**, 47-56.

161. Robert, D.; Carbó-Dorca, R. A formal comparison between molecular quantum similarity measures and indices. *J. Chem. Inf. Comp. Sci*., *38*, **1998**, 469-475.

162. Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular quantum similarity measures as an alternative to log P values in QSAR studies. *J. Comput. Chem.*, *19*, **1998**, 1575-1583.

163. Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. *J. Comput. Aid. Mol. Des*., *13*, **1999**, 259-270.

164. Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to LFER: substituent and solvent effects on the acidities of Carbóxylic acids. *J. Phys. Org. Chem.*, *12,* **1999**, 447-454.

165. Amat, L.; Carbó-Dorca, R.; Ponec, R. Simple linear QSAR models based on Quantum Similarity Measures. *J. Med. Chem., 42,* **1999**, 5169-5180.

166. Amat, L.; Besalú, E.; Carbó-Dorca, R.; Ponec, R. Identification of active molecular sites using quantum-self-similarity measures. *J. Chem. Inf. Comp. Sci., 41,* **2001**, 978-991.

167. Ponec, R.; Gironés, X.; Carbó-Dorca, R. Molecular basis of LFER. The nature of inductive effects in aliphatic series. *J. Chem. Inf. Comp. Sci., 42,* **2002**, 564-570.

168. Gironés, X.; Carbó-Dorca, R.; Ponec, R. Molecular basis of LFER. Modeling of the electronic substituent effect using fragment quantum self-similarity measures. *J. Chem. Inf. Comp. Sci.*, *43*, **2003**, 2033-2038.
169. Besalú, E. Combinator v1.3. Institute of Computational Chemistry, University of Girona, **2003**.
170. Besalú, E.; Ponec, R.; Julián-Ortiz, J.V. Virtual Generation of Agents Against Mycobacterium tuberculosis. A QSAR study. *Molecular Diversity, 6,* **2003**, 107-120.
171. Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Effect of substituent size and dimensionality on potency of phenolic xenoestrogens evaluated with a recombinant yeast assay. *Environ. Toxicol.*, *19*, **2000**, 2637-2642.
172. Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Structure-Activity Relationships for Gene Activation Oestrogenicity. Evaluation of a Diverse Set of Aromatic Chemicals. *Environ. Toxicol. Chem.*, *17,* **2002**, 14- 23.
173. Becke, A.D.; Edgecombe, K.E. A simple measure of electron localization in atomic and molecular systems. *Chem. Phys*., *92,* **1990**, 5397-5403.
174. Savin, A.; Becke, A.D.; Flad, J.; Nesper, R.; Preuss, H.; von Schnering, H.G.. Einneuer Blick auf die Elektronenlokalisierung. *Angew. Chem*., *103,* **1991**, 421.
175. Silvi, B.; Savin, A. Classification of Chemical Bonds on Topological Analysis of Electron Localization of the Si(100) Surface. *Nature, 371,* **1994**, 683.
176. Savin, A.; Nesper, R.; Wengert, S.; Fässler, T.F. ELF: The Electron Localization Function. *Angew. Chem. Int. Ed. Engl., 36,* **1997**, 1808-1832.
177. Savin, A. *Reviews of modern quantum chemistry: A celebration of the contributions of Robert G. Parr*. Sen, K.D. (Ed.) World Scientific: Singapore, **2002**, 43.
178. Cancès, E.; Keriven, R.; Lodier, F.; Savin, A. How electrons guard the space: shape optimization with probability distribution criteria. *Theor. Chem. Acc*. In press.
179. Bultinck, P.; Kuppens, T.; Gironés, X.; Carbó-Dorca, R. Quantum Similarity Superposition Algorithm (QSSA): A Consistent Scheme for Molecular Alignment and Molecular Similarity Based on Quantum Chemistry. *J. Chem. Inf. Comp. Sci., 43,* **2003,** 1143-1150.
180. Bultinck, P.; Carbó-Dorca, R.; Van Alsenoy, C. Quality of Approximate Electron Densities and Internal Consistency of Molecular Alignment Algorithms in Molecular Quantum Similarity. *J. Chem. Inf. Comp. Sci., 43,* **2003,** 1208-1217.
181. Mestres, J.; Solà, M.; Carbó, R.First-order molecular descriptors for molecular steric similarity. *Sci. Gerun., 21,* **1995**, 165-173.
182. Carbó, R.; Besalú, E. Extending Molecular Similarity to Energy Surfaces: Boltzmann Similarity Measures and Indices. *J. Math. Chem., 20,* **1996**, 247-261.
183. Mezey, P.G.; Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to the characterization of molecular chirality. *Enantiomer, 4,* **1999**, 371–378.
184. Bultinck, P.; Augustynen, S.; Hilbers, H.W.; Moret, E.; Tollenaere, J.P. Generate: A program for 3-D structure generation and conformational analysis of peptides and peptidomimetics. *J. Comp. Chem.*, *23,* 746-754.

Quantum Similarity Theory

The underlying physical laws necessary for the mathematical theory
of a large part of physics and the whole of chemistry
are thus completely known, and the difficulty is only that
the exact application of these laws
leads to equations much too complicated to be soluble.
It therefore becomes desireable
that approximate practical methods
of applying quantum mechanics should be developed,
which can lead to an explanation
of the main features of complex atomic systems
without too much computation.

Quantum Mechanics of Many-Electron Systems
Proc. R. Soc. Lond. Ser. A, 123,1929, 714
P.A.M. Dirac

1 INTRODUCTION

The characteristics and behaviour of substances are partially conditioned by their structure, that is, compounds have different functions because they have different structures. Thus, it can be considered that in chemistry, function follows form. Quantum mechanics principles postulate that the geometric and electronic structure of a molecule contains the features responsible for its physicochemical properties and biological activity. However, it is not obvious that these features can be discerned in a simple way. In addition, in quantitative structure-function studies, there is the conceptual difficulty of relating structures, which cannot be simply depicted by a number, with properties, most of which are represented by numbers. By using a set of well-understood mathematical parameters as descriptors of molecular structure, complex physicochemical and biological behaviour of molecules can be described. Such approaches differ from the traditional QSAR methodology, where selected simpler physicochemical properties are employed to predict the function of molecules. Indeed, mathematical descriptors have a direct structural interpretation and they offer a deeper insight into the structural factors governing molecular properties.

In particular, the characterization of **chemical structure** has been forever of great interest, although the term was not properly described until 1861 by the Russian chemist Butlerov [1]. Butlerov defined chemical structure as the type and manner of the mutual binding of atoms in a compound, without specifying the nature of bonding. The links existing between atoms in molecules were depicted as dotted or continuous lines [2], solid rods [3], or even as tubes of force [4]. Structural formulas drawn with straight lines connecting the bonded atoms were first published in 1858 by the Scottish chemists **Couper** [2], and in 1864 by **Crum Brown** [5-7]. From those times, several characterization levels of molecular structures have been described, from the simple enumeration of atoms to complex metabolic simulations.

Thus, the characterization of a structure is represented by an ordered set of components with some information on the relationship between the components. Such information can be given in a form of a list of the components that imply the labelling of atoms and bonds (molecular codes), or in the form of the count of components of various types, describing the mathematical properties of a structure (structural invariant). Different structural molecular description levels, ordered by the increasing amount of provided information, are listed below:

a) List of **type of atoms** that constitute the molecule.

b) **Empirical formula**, that is, the simplest stechiometric formula indicative of the proportion of different atoms.

c) **Molecular formula**, indicative of the number of atoms of each type. It corresponds to the formula necessary to calculate the exact molecular mass.

d) In contrast to the monodimensional constitutional information provided by the preceeding formulas, the **bidimensional structural formula** represents the arrangement of atoms using the topology of the molecule, and the connectivity of the constituting atoms.

d.1) The **graph**, a variant of the structural formula, omits the type of atom and nature of bonding.

d.2) Also, the simplified **hydrogen-supressed graphs**, employed in organic chemistry for structure representation, are widely used.

d.3) It has to be remarked that alternative representations at a similar level have been designed, for example Simplified Molecular Input Line Entry Specification (SMILES) [8].

e) **Three-dimensional structure** describes the structure of the molecule as a three-dimensional entity with the atoms situated in specific positions in the space (*x*,*y*,*z*, coordinates), thus providing geometrical and spatial information.

f) Resolutions of **Schrödinger equation** independent or dependent of the time, which include the description of charge distribution. They can constitute the most accurate description, depending on the level of theory used to solve them.

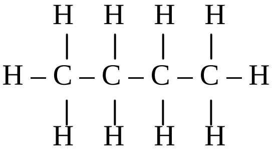
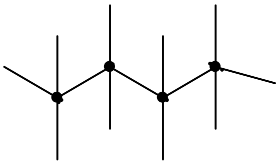
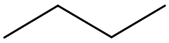
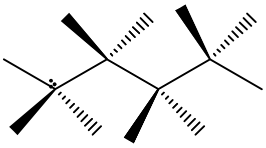
a) Type of atoms	C and H
b) Empirical formula	C_nH_{2n+2}
c) Molecular formula	C_4H_{10}
d) 2D Structural formula	
d.1) Graph	
d.2) Hydrogen-suppressed graph	
d.3) SMILES	CCCC
e) 3D Structure	
f) Schrödinger equation	$\hat{H}\Psi = E\Psi$ $\hat{H}\Psi = i\hbar\frac{d\Psi}{dt}$

Figure 1. Levels of characterization for butane.

For the formulation of structure-function relationships, different descriptors can be employed, depending on the theoretical basis adopted for the description of the structure of molecules:

- **Quantum similarity descriptors**, based on quantum mechanical concepts
- **Topological descriptors**, based on graph-theoretical concepts
- **Quantum similarity-based topological descriptors**, founded on the intersection of quantum similarity and classical graph theory

The first approach considers the three-dimensional geometrical structure of molecules, derived from quantum mechanical calculations. Schrödinger equation provides the theoretical tools to calculate density functions, which describe the electronic characteristics of molecules at different levels of theory, i.e. semiempirical, and *ab initio*, in order of increasing precision. Then, quantum similarity theory, presented in detail in *Section 3,* employs density functions to construct quantum similarity measures, used as source for descriptors, as exposed in *Section 4*. The second approach, treated in *Section 5*, pictures molecules as planar graph structures, and is focused on the topological description of molecules. Graph-theoretical descriptors are based on binary connectivity or adjacency tables, which account for the presence or absence of connections between atoms. Finally, a connection between classical graph theory and the general theory of quantum similarity can be also envisaged, as shown in *Section 6,* leading to the quantum similarity-based topological indices, which have into account as well the 3D spatial disposition as the structure-based description of the system.

2 INTRODUCTION TO SIMILARITY

2.1 The concept of similarity

In human conscience, the **intuitive concept** of similarity is deeply attached to knowledge. In everyday life, several unconscious associative mechanisms based on experience allow to establish common characteristics and differences among the perception of several objects, events, or situations, i.e., the usual contexts for similarity. In an instinctive way, the human mind continuously compares new knowledge with background knowledge, using criteria founded on experience. Therefore, a new concept is acquired when some similarities and/or dissimilarities are processed among the new information received and the previous one [9-10].

The first scientific-like contributions to the similarity concept date back to the ancient Greek philosophy. Concretely, the logical inference process of **analogy** consists of a relational process based on the systematic comparison between structures, in order to analyze their common and distinctive features. Using the properties and relations between objects as a kind of thinking involved in various cognitive tasks, analogy allows making inferences based on the correspondences found between objects. In a philosophical interpretation of science, similarity attempted to explain the **characterization of matter** from the basic elements: fire, air, water, and earth. In chemistry, incidentally, the principle of analogy is the basis to assume that similar molecules possess similar properties, and this is the foundation of empirical relationships between structure and activity.

In the frame of mathematics, similarity had been rationally used by **Pythagoras** to formulate the well-known theorem, based on the similarity of triangles. Indeed, similarity is undoubtedly an important geometrical and spatial concept. Particularly, some models for rational number concepts are based on similarity; for example, **ratio and proportion** [11].

Mathematicians use the term similar to describe objects that have the same shape but not necessarily the same size, that is, proportional objects with the same ratio. However, this is not a precise definition for similarity; to delimit this definition it has to be taken into account that, on one hand, the qualifier “not necessarily” includes congruent figures as a special case of similar figures and, on the other hand, the necessary conditions for objects to have the same shape need to be defined.

For example, two geometrical figures are considered to have the same shape if their corresponding angles are equal and the corresponding sides are proportional. To state the matter differently, two objects are similar if they can be transformed (by translations, reflections or rotations) only changing their position in a plane, so that they can be enlarged or reduced. The enlargement refers to similarity transformations that make a figure bigger, whereas the reduction is therefore regarded as an enlargement with a scale factor between 0 and 1.

In chemistry, the **periodical table of the elements** was predicted in 1869 by **Mendeleev** [12], from the observation and comparison of the similar chemical behaviour and reactivity of elements. From the systematization of atomic properties, the elements were classified into a table, with empty gaps corresponding to the substances still unknown. Mendeleev was able to predict not only the non-discovered elements but also their physico-chemical properties, by noting patterns of the combination ratio of well-classified elements.

Summarizing, the concept of similarity can be directly related to the **relative comparison** between different systems. Notwithstanding, although in the human's intuitive concept of similarity there is a notion of degree of similarity, the process to establish similarities and analogies is often heuristic and subjective, and hence difficult to systematize. Thus, it is only meaningful to compare two objects with respect to a third one. Similarity does not exist in absolute terms but it is a relative term. Besides, similarity can be performed between the representations of molecular structures but also between numerical descriptors obtained by conversion of those structures. In particular, for scientific purposes, the **quantification of similarity**, univocally calculated from well-defined parameters, has been forever of great interest.

2.2 Molecular Quantum Similarity

The definition of a similarity measure between two chemical systems is a crucial question in theoretical chemistry. Specially, the description of quantitative measures for molecular similarity has been carefully examined in the bibliography 13]. **Molecular similarity** attempts to answer the question: "how similar is a molecule to another?". There is not a unique answer for this question, but it depends on the molecular aspect to be analyzed, such as functional groups or common substructures. Considering that atoms and molecules can be regarded as **Quantum Objects (QO)**, a rigorous definition of similarity is based on quantum mechanics theory, which deals with the information regarding such microscopic systems. Thus, **molecular similarity measures** can be naturally based on well-defined theoretical **quantum mechanical descriptors,** derived from the molecular electronic structure.

Quantum mechanics provides a feasible way to attach a descriptive function to each molecule. According to quantum mechanics postulates, since all the information that can be extracted from a quantum system is contained in its wavefunction, **density functions** constitute a suitable source for similarity descriptors. Thus, from the subsequent definition of a similarity measure, founded on **quantum similarity theory**, quantum objects can be described in a quantitative way, and, therefore, their degree of similarity can be evaluated. The ability to measure the similarity between a pair of items provides the capacity to construct new objects with better characteristics than the first ones, as will be shown afterwards. Moreover, quantum similarity theory can be also based as a frame to compare general shapes, conformations of points, distance functions on graphs, matrix eigenspectra, distributions, etc. [14].

Historically, the quantification of the similarity between two molecular structures based on quantum chemistry was firstly proposed by **Carbó** et al. in 1980 [15]. In this seminal paper, Carbó provided a general numerical definition for the measure of similarity between two molecules based on the comparison of their molecular electronic densities, hence founding quantum similarity theory. Afterwards, the same research group has devoted to its theoretical development [16-24], and comprehension of the quantum mechanical meaning of similarity measures [25], soundly based on mathematical foundations [26-28]. Concerning to the implementation of novel procedures, it must be noted the deduction of adjusted electronic density functions [18,21,29-34], and the design of new algorithms for molecular superposition [35-36]. In the application frame, the definition of new measures, and their practical applications [16,37] have also been extensively developed.

Simultaneously, other research groups adopted quantum similarity, extending the implementation of the theory to various fields and envisaging new applications, therefore producing a great deal of interesting results. Among the most relevant contributions, the work of **Richards** [38-49], who quantified similarity measures substituting density functions by electrostatic potentials [51-52], the measures proposed by **Cioslowsky** [53-56], the similarity measures defined upon momentum-space density functions of **Allan** and **Cooper** [57-64], and the work of other authors [65-74] must be noted.

Besides, from another point of view, **Herndon** substituted quantum mechanical descriptors by topological indices derived from graph theory. Herndon studied the quantification of a measure of similarity synthesizing graph theory and molecular similarity [75-80]. This topological approach was also pursued by **Mezey** [81-91], and **Ponec** [92-99], who studied the electronic effects in several organic reactions, among others.

Since then, molecular similarity has had a great success, entailing the publication of several monographs [10,100-107], and joining scientists of all over the world in the biannual Girona Seminar on Molecular Similarity, which has recently celebrated its sixth edition [108-110].

3 QUANTUM MECHANICS AND THE ROLE OF DENSITY FUNCTION

3.1 The Wavefunction: First postulate of Quantum Mechanics

The first postulate of quantum mechanics [111-113] states that every state of a quantum object, i.e. a physical system formed by a numerable assembly of microscopic particles, can be described by means of a function depending on the variables of the system: time and coordinates of position. This complex mathematic function, from which all the information of the system can be extracted, is the so-called wavefunction, $\mathbf{\Psi}\left(\mathbf{r};t\right)$. In a stationary state, the time-dependent wavefunction can be separated by means of the variable separation technique in the product of a time-independent wavefunction by a time-dependent function:

$$\mathbf{\Psi}\left(\mathbf{r};t\right)=\mathbf{\Psi}(\mathbf{r})\cdot f(t) \qquad (1)$$

In the particular case of molecular systems, the **Born-Oppenheimer approximation** [114] qualitatively assumes that nuclei are much more massive than electrons. This allows considering nuclear charges nearly fixed with respect to the motion of electrons. Working within this approach, the nuclear and electronic motion can be separated. Thus, the configuration of nuclei is fixed, so that the system's nuclear positions are composed of a set of constant nuclear coordinates, and the electronic wavefunction depends only parametrically on the spatial nuclear coordinates.

The state time-independent electronic wavefunction for *N*-particles can be defined as:

$$\mathbf{\Psi}\left(\mathbf{r},\mathbf{R}\right)=\mathbf{\Psi}(\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_N,\mathbf{R}) \qquad (2)$$

where the vector **r** collects the coordinates for *N* particles, while **R** describes the wavefunction dependence upon a parameter set.

3.2 Born's interpretation: Probability distribution Density Function

However, although the wavefunction gives probabilistic information about the position of the particles in the system, it has no physical meaning by itself. Even more, in systems with a large number of particles, it is not easy to treat. Instead, the probability distribution of the **Density Function** (**DF)**, expressed in terms of the wavefunction, $\mathbf{\Psi}$, and its complex conjugated, $\mathbf{\Psi}^*$, has the advantage of gathering all the information in a more handy way.

$$\rho = \mathbf{\Psi}^* \mathbf{\Psi} = \left| \mathbf{\Psi} \right|^2 \tag{3}$$

In addition, the squared module of the wavefunction is positive definite by construction, and has an attached physical significance. The first description of the physical sense of the squared module of the wavefunction can be attributed to **Born** [116], and is gathered in the seminal quantum mechanics books of **von Neumann** [117], and **Dirac** [118]. The Born's interpretation of the wavefunction states that, in a stationary state, the probability of finding a particle in a given infinitesimal volume element *dV* can be expressed as:

$$dP = \mathbf{\Psi}^* \mathbf{\Psi} dV = \left| \mathbf{\Psi} \right|^2 dV \tag{4}$$

Then, the **probability density function** or the probability distribution function, also called charge density, is expressed by the differential equation:

$$\frac{dP}{dV} = \left| \mathbf{\Psi} \right|^2 \tag{5}$$

The numerical probability value to find a given particle in a specific region of the space, ω, can be calculated by integration of the previous equation within the domain of interest:

$$P(\omega) = \int_{\omega} dP = \int_{\omega} \mathbf{\Psi}^* \mathbf{\Psi} dV \tag{6}$$

Therefore, from the classical quantum mechanics point of view, **Born** endowed the solution of the Schrödinger equation [119] with a statistical meaning of physical significance. Thereby, the wavefunction set of a microscopic system conveniently transformed into its squared module produces a set of probability density functions, giving the probability of finding the electron in a volume element *d***r** at a given position **r**, and a fixed time *t*:

$$dP = \mathbf{\Psi}^*(\mathbf{r}) \mathbf{\Psi}(\mathbf{r}) d\mathbf{r} = \left| \mathbf{\Psi}(\mathbf{r}) \right|^2 d\mathbf{r} = \rho(\mathbf{r}) d\mathbf{r} \tag{7}$$

The probability distribution function results in:

$$\frac{dP}{d\mathbf{r}} = \rho(\mathbf{r}) \tag{8}$$

Thus, consistent with the physical interpretation of $\rho(\mathbf{r})\,d\mathbf{r}$, the **distribution of probability** of encountering a particle within this infinitesimal volume is comprised between 0 and 1. Hence, the probability is normalized to the unit, that is, the probability that an electron exists in the whole space is complete, and so the integral over the entire space, Ω, yields the total probability of presence:

$$P(\Omega)=\int_{\Omega}\rho(\mathbf{r})\,d\mathbf{r}=\int_{\Omega}\mathbf{\Psi}^{*}(\mathbf{r})\,\mathbf{\Psi}(\mathbf{r})\,d\mathbf{r}=1 \tag{9}$$

Besides, the wavefunction applied over an electronic system can also define the distribution of electronic charge density, which accounts for the electronic charge concentration of the system in the space. When this definition is adopted, the integral over the whole space gives the total number of electrons in the molecule, *N*:

$$P(\Omega)=\int_{\Omega}\rho(\mathbf{r})\,d\mathbf{r}=N \tag{10}$$

This interpretation was consistently put on a firm theoretical basis by Von Neumann [117], who set the whole quantum theory into the frame of his operator algebra.

3.3 Experimentally measurable properties: Fifth postulate of Quantum Mechanics

From the canonical interpretation of quantum mechanics, any microscopic system wavefunction set, conveniently transformed into a square module, produces a set of probability density functions. DF is the adequate tool that has to be used for interpreting the experimental behaviour of particle systems, such as atoms and molecules. Accordingly, DF plays a leading role in quantum mechanical systems description.

In particular, the fifth postulate of quantum mechanics states that once known the DF of the system in a precise internal energy state, all the compatible observable property values of the system, ω, which are experimentally measurable, can be formally extracted from it as expectation values, $\langle\omega\rangle$, of its associated hermitian operator, Ω, acting over the corresponding wavefunction, $\mathbf{\Psi}$:

$$\langle\omega\rangle=\frac{\langle\mathbf{\Psi}|\Omega|\mathbf{\Psi}\rangle}{\langle\mathbf{\Psi}|\mathbf{\Psi}\rangle} \tag{11}$$

Assuming a normalized wavefunction, the denominator can be cancelled:

$$\langle\omega\rangle = \langle\mathbf{\Psi}|\Omega|\mathbf{\Psi}\rangle = \int \mathbf{\Psi}^*(\mathbf{r})\Omega(\mathbf{r})\mathbf{\Psi}(\mathbf{r})d\mathbf{r} \tag{12}$$

In the case of physical observables associated to non-differential hermitic operators, the expected value of the operator can be expressed as a function of the electronic density, ρ:

$$\langle\omega\rangle = \langle\Omega|\rho\rangle = \int \Omega(\mathbf{r})\rho(\mathbf{r})d\mathbf{r} \tag{13}$$

Thus, the obtaining of the similarity degree between two compared systems in a unique system state can be formally performed like a **statistical expectation value technique**.

3.4 Density Function construction

In consequence, in order to define and interpret **Quantum Similarity Measures** (**QSM**) according to classical quantum mechanical principles, the description of microscopic systems can be set up with the following algorithm:

1) Construction of the mathematical **Hamiltonian operator**, $\widehat{\mathrm{H}}$, representing the set of different interactions acting upon the quantum object.

2) Computation of the state energy-wavefunction pairs, $\{E,\mathbf{\Psi}\}$, by solving the **Schrödinger equation** independent of the time: $\widehat{\mathrm{H}}\mathbf{\Psi} = \varepsilon\mathbf{\Psi}$

3) Evaluation of the **state DF**, computed from the squared module of the wavefunction: $\rho = \mathbf{\Psi}^*\mathbf{\Psi} = |\mathbf{\Psi}|^2$

The formation process of DF starting from the original system's wavefunction can be expressed by way of the **generating rule** [2424], $\Re(\Psi \to \rho)$, which summarises the three steps of the quantum mechanics algorithm:

$$\Re(\mathbf{\Psi} \to \rho) = \left\{\forall\mathbf{\Psi} \in H(\mathbf{C}) \to \exists\rho = \mathbf{\Psi}^*\mathbf{\Psi} = |\mathbf{\Psi}|^2 \in H(\mathbf{R}^+)\right\} \tag{14}$$

where the wavefunction and the Hamiltonian are explicitly defined over the Hilbert complex field $H(\mathbf{C})$, whereas the DF is defined over the real field $H(\mathbf{R}^+)$. Thus, quantum mechanical density functions are elements of a Hilbert semispace, positive definite and normalized in the usual sense, that is, submitted to **convexity conditions**:

$$\rho(\mathbf{r}) \in \mathbf{R}^+ \wedge \int \rho(\mathbf{r})d\mathbf{r} = 1 \tag{15}$$

In the particular case where the quantum system is a molecule, the real density function contains all the information of the distribution of electrons. Once known the electronic wavefunction for the stationary system, $\mathbf{\Psi}\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\right)$, where $\mathbf{x}_i$ includes all the variables of the particle *i*, its electronic DF is expressed as the product of the wavefunction and its conjugated:

$$\rho\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\right)d\mathbf{x}_1 d\mathbf{x}_2...d\mathbf{x}_N=\mathbf{\Psi}^*\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\right)\mathbf{\Psi}\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\right)d\mathbf{x}_1 d\mathbf{x}_2...d\mathbf{x}_N \qquad (16)$$

The previous equation corresponds to the probability of finding the electron *1* in a *d*$\mathbf{x}_1$ configuration, simultaneously with the electron *2* in a *d*$\mathbf{x}_2$ configuration, and successively with the *N* electrons. Explicitly, the density function depends on the spatial and spin coordinates of the particles:

$$\boldsymbol{\rho}\left(\mathbf{r}_1,...,\mathbf{r}_N;s_1,...,s_N\right)=\mathbf{\Psi}^*\left(\mathbf{r}_1,...,\mathbf{r}_N;s_1,...,s_N\right)\mathbf{\Psi}\left(\mathbf{r}_1,...,\mathbf{r}_N;s_1,...,s_N\right)=\left|\mathbf{\Psi}\left(\mathbf{r}_1,...,\mathbf{r}_N;s_1,...,s_N\right)\right|^2 \qquad (17)$$

3.4.1 First-order Density Function

For practical purpose, i.e. for use in molecular comparison, the dimensionality of the density function can be reduced using the **McWeeny and Löwdin** theoretical development [120-122]. A *p*-th order density matrix element can be defined by integrating ρ over the entire system particle coordinates except *p* of them.

$$\rho^{(p)}\left(\mathbf{x},\mathbf{x}'\right)=\binom{N}{p}\int...\int\mathbf{\Psi}^*\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_p,\mathbf{x}_{p+1},...,\mathbf{x}_N\right)\Psi\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_p,\mathbf{x}_{p+1},...,\mathbf{x}_N\right)d\mathbf{x}_{p+1}d\mathbf{x}_{p+2}...d\mathbf{x}_N \qquad (18)$$

The diagonal elements of the density matrix allow the derivation of the *p*-th order DF element. As a useful particular case, the integration over the *N-1* position coordinates yields the first-order electronic density function**,** $\rho^{(1)}\left(\mathbf{x}_1\right)$**.** The first-order DF is defined as the probability of finding one electron, indistinguishable from the *N-1* remaining electrons in the molecule, with the *d*$\mathbf{x}_1$ configuration.

$$\rho^{(1)}\left(\mathbf{x}_1\right)=N\int...\int\mathbf{\Psi}^*\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\right)\mathbf{\Psi}\left(\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\right)d\mathbf{x}_2...d\mathbf{x}_N \qquad (19)$$

From the previous expression, the first-order spatial electronic density function, which indicates the probability to find an electron with an independent spin at *d*$\mathbf{r}_1$, results from integrating all the *N* spin coordinates and the *N-1* spatial coordinates:

$$\rho^{(1)}\left(\mathbf{r}_1\right)=N\int...\int\left|\mathbf{\Psi}\left(\mathbf{r}_1,\mathbf{r}_2,...,\mathbf{r}_N;s_1,s_2,...,s_N\right)\right|^2\,\mathrm{d}\mathbf{r}_2...\mathrm{d}\mathbf{r}_N\ \mathrm{d}s_1...\mathrm{d}s_N \qquad (20)$$

Certainly, for a given molecular structure, its corresponding wavefunction can be computed for a chosen system's state and, from this one, the first-order density function can be easily evaluated. Although any density function order could be used, the first-order state electronic density functions have been chosen for being sufficiently well behaved. Thus, in the construction of QSM for practical purpose, that is, for use in molecular comparison, first-order DF are considered to be good candidates to be used, even though higher order DF can be employed as well.

For the sake of simplicity, the first-order density function, $\rho^{(1)}\left(\mathbf{r}_1\right)$, will be expressed omitting the superscript and subscript, i.e. $\rho\left(\mathbf{r}\right)$**.**

3.4.2 *Ab initio* Density Function

In particular, in molecules and electronic systems, many-electron wavefunctions are constructed by multiplying monoelectronic wavefunctions that describe the molecular electronic structure. The monoelectronic wavefunctions describing the particles of the system are constructed by means of **Molecular Orbitals (MO)**, $\Psi_i\left(\mathbf{x}\right)$, where **x** depends on the spatial and spin coordinates. A **spatial molecular orbital** $\psi_i\left(\mathbf{r}\right)$ is a function of the position vector $\mathbf{r}$ describing the spatial distribution of an electron. This spatial wavefunction is constructed by means of a basis function set that, for the sake of simplicity in the calculation of the associated integrals, is usually formed by Gaussian functions.

Within the **Linear Combination of Atomic Orbitals – Molecular Orbital (LCAO-MO)** approximation [123], each molecular orbital is expressed as a linear expansion of basis functions, that is, **Atomic Orbitals (AO),** $\left\{\phi_\mu(\mathbf{r})\right\}$. If the set of spatial orbitals was complete, then, any monoelectronic wavefunction could be exactly expanded as a linear combination of spatial basis functions:

$$\psi_i\left(\mathbf{r}\right) = \sum_{\mu=1}^{\infty} c_{\mu i}\phi_\mu\left(\mathbf{r}\right) \qquad (21)$$

where the subscript *μ* is for the basis functions and *i* for the molecular orbitals.

Unfortunately, in practice, it is not possible to work with an expansion with an infinite number of basis functions**.** Therefore, the set of spatial molecular orbitals is restricted to a finite set of *k* known basis functions $\left\{\phi_\mu(\mathbf{r})\middle|\mu = 1, 2, ...k\right\}$.

$$\psi_i(\mathbf{r}) = \sum_{\mu=1}^{k} c_{\mu i}\phi_\mu(\mathbf{r}) \quad i = 1,2,...k \tag{22}$$

Thus, the finite set of basis functions only spans a certain region of the complete space, so that the results can be only considered as exact within the subspace spanned by the orbital set used. For this reason, it is important to choose a basis set that describes the molecular orbitals with a reasonable accuracy and computationally affordable at the same time.

The total **charge density** of a closed-shell molecule can be described at the **Hartree-Fock level** within the molecular orbital theory [124] as a sum of charge densities, expressed by means of the mentioned spatial molecular orbitals**,** where each occupied molecular orbital contains two electrons:

$$\rho(\mathbf{r}) = 2\sum_{i}^{N/2} |\psi_i(\mathbf{r})|^2 = 2\sum_{i}^{N/2} \psi_i^*(\mathbf{r})\psi_i(\mathbf{r}) \tag{23}$$

As pointed before, $\rho(\mathbf{r})d\mathbf{r}$ is the probability of finding any electron in $d\mathbf{r}$ at $\mathbf{r}$. Considering normalized molecular orbitals, the probability of finding an electron within the whole space is just the number of electrons involved in the system:

$$\int \rho(\mathbf{r})d\mathbf{r} = 2\sum_{i}^{N/2} \int |\psi_i(\mathbf{r})|^2 d\mathbf{r} = 2\sum_{i}^{N/2} 1 = N \tag{24}$$

where *N* is the total number of electrons and *N/2* is the number of occupied molecular orbitals.

If the molecular orbital expansion (22) is inserted into the expression for the charge density (23), then the first-order electronic density function can be expressed as a double sum upon all the basis function pairs:

$$\rho(\mathbf{r}) = 2\sum_{i}^{N/2} \sum_{\mu} c_{\mu i}^* \phi_\mu^*(\mathbf{r}) \sum_{\nu} c_{\nu i}\phi_\nu(\mathbf{r}) = \sum_{\mu\nu} P_{\mu\nu}\phi_\mu^*(\mathbf{r})\phi_\nu(\mathbf{r}) \tag{25}$$

Where $c_{\mu i}$, $c_{\nu i}$ are coefficients of atomic orbitals, $\{\phi_\mu(\mathbf{r})\}$ is the basis function set of spatial atomic orbitals and $\{P_{\mu\nu}\}$ are the elements of the **density matrix** or charge and bond order matrix, defined from the AO coefficients as:

$$P_{\mu\nu} = 2\sum_{i}^{N/2} c_{\mu i}^{*} c_{\nu i} \tag{26}$$

Given a set of known basis functions, $\{\phi_\mu\}$, the **P** matrix specifies completely the charge density $\rho(\mathbf{r})$, directly related to the expansion coefficients **c** by (26).

3.4.3 Density Function of a molecular fragment

Alternatively, the **holographic electron density theorem** [125] states that all the information contained in the total electronic density of a molecule is also included in the local density cloud of any nonzero volume fragment.

Using the LCAO-MO approximation, the DF of a fragment *X* belonging to the molecule *A* is defined as:

$$\rho_A^X(\mathbf{r}) = \sum_{\mu\in X}\sum_{\nu\in A} P_{\mu\nu}\phi_\mu^*(\mathbf{r})\phi_\nu(\mathbf{r}) \tag{27}$$

where ν is calculated over all the basis function of the *A* molecule, whereas µ is only used for the basis functions centred in the atoms belonging to the studied fragment *X*.

This definition of **fragment density** provides an additive partition of the total molecular electronic density [126]. For example, when the density is divided into all the possible fragments formed by a single atom, the total density of the whole molecule, $\rho_A(\mathbf{r})$, can be generated by adding all the atomic contributions, $\rho_a(\mathbf{r})$:

$$\rho_A(\mathbf{r}) = \sum_a \rho_a(\mathbf{r}) \quad \wedge \quad \rho_a(\mathbf{r}) = \sum_{\mu\in a}\sum_{\nu\in A} P_{\mu\nu}\phi_\mu^*(\mathbf{r})\phi_\nu(\mathbf{r}) \tag{28}$$

3.5 Fitted Density Function

Nowadays, precise theoretical *ab initio* studies of large molecular systems or transition metal complexes are usually limited by the number and kind of atoms involved, due to the high computational requirements. Thus, in spite of the progressive growth in the capacity of calculation of current computers, the systematic study of such complex molecular systems has motivated the development of approximated DF sufficiently accurate so as to be able to replace the *ab initio* DF [127-133]

Taking into consideration the generic expression of the density function at the **Hartree-Fock** level (25), the main *ab initio* calculation hindrance is generally located in the computation of cumbersome four-center integrals, which appear in the calculation of quantum similarity measures of two molecular systems. Specially, when the optimization procedure for alignment requires multiple evaluations of many-centre similarity integrals, the calculations are highly time-consuming, thus making the computations at *ab initio* level unaffordable and limiting the use of *ab initio* QSM to small molecular systems, due to the current computational limits.

Therefore, the development of simplified electron densities, symbolized as $\tilde{\rho}(\mathbf{r})$, has been widely used not only to overcome the problem of the bottleneck superposition process in the computation of quantum similarity measures, but also to accelerate any related DF calculation.

For free atomic systems, the wavefunction can be expressed as a sum of squares corresponding to a spherically symmetric subshell [134]. Consequently, the electron density function of closed-shell atoms can be modelled by means of an integral **Gaussian transform** over the radial coordinate:

$$\widetilde{\rho}(\mathbf{r}) = \int_0^\infty f(\zeta) e^{-\zeta r^2} d\mathbf{r} \tag{29}$$

which can be approximated by truncating the previous expression as a finite sum

$$\widetilde{\rho}(\mathbf{r}) \approx \sum_i^k w_i f_i e^{-\zeta r_i^2} \tag{30}$$

that is, a linear combination of exponential *1S* Gaussian functions, where the coefficients of the functions, if they are positive-valued, indicate non-negative occupancies.

Thus, the basis function set of atomic orbitals can be built as a linear combination of Gaussian functions φ_a:

$$\phi_\mu(\mathbf{r}) = \sum_{a=1}^k g_{a\mu} \varphi_a(\mathbf{r}) \tag{31}$$

where $\{\varphi_a\}$ is the normalized *1S*-type **Gaussian Type Orbital (GTO)** basis set, and $\{g_{ai}\}$ are the coefficients. Then, the first-order *ab initio* DF involves products of two centres or indices:

$$\rho(\mathbf{r}) = \sum_{\mu\in A}\sum_{\nu\in A} P_{\mu\nu}\phi_{\mu}^{*}(\mathbf{r})\phi_{\nu}(\mathbf{r}) = \sum_{\mu}\sum_{\nu} P_{\mu\nu}\sum_{a}\sum_{b} g_{a\mu}^{*} g_{b\nu}\varphi_{a}^{*}(\mathbf{r})\varphi_{b}(\mathbf{r}) \tag{32}$$

After a simple diagonalization plus a unitary transformation of the basis set, the initial expression of DF can be transformed into the linear combination:

$$\widetilde{\rho}(\mathbf{r}) = \sum_{i} w_i \left|S_i(\mathbf{r})\right|^2 \tag{33}$$

In the literature, different fitting algorithms for first-order electronic density functions have been reported [130, 135], but not all of them take into account the conditions needed to obtain a definite positive density.

In the first MQSM calculations [15,17,136-143] a **Complete Neglect of Differential Overlap** (**CNDO)** approximation was used. The electronic density was described only by means of valence shell spherical functions.

3.5.1 Atomic Shell Approximation (ASA)

In order to avoid expensive computational calculations, the molecular DF has been adjusted using the **Atomic Shell Approximation (ASA)**. This approximation can reduce the computation time several orders of magnitude without a significant loss of accuracy. Moreover, in order to assure the probabilistic meaning of the original electron density functions, the expansion coefficients are restricted to be positive-valued.

The ASA has been developed as a theoretic model of adjustment of density functions, widely implemented, for the calculation of molecular electronic densities used in the computation of MQSM [144-145]. This electron density fitting algorithm constitutes a way to adjust the first-order molecular electronic density functions at the Hartree-Fock level to linear combinations of spherically symmetric functions. In addition, the ASA DF must fulfill two conditions: it must be normalized to one or to the number of electrons, and possess positive definite coefficients. Provided that the electronic density is previously calculated at a given computational level and with a specific basis function set, and then the *ab initio* densities are fitted to the *1S* functions, this methodology can be considered to be based on *ab initio* calculations.

The molecular electron density is expressed as a linear expansion of spherical *1S* functions:

$$\rho_A^{ASA}(\mathbf{r}) = \sum_{i\in A} w_i \left|S_i(\mathbf{r};\alpha_i)\right|^2 \tag{34}$$

where *i* refers to the atomic shells, $\left\{w_i\right\}$ are the positive definite ASA coefficients, and $\left\{S_i\right\}$ is the set of normalized *1S* Gaussian type orbitals:

$$S_i\left(\mathbf{r};\alpha_i\right)=\left(\frac{\alpha_i}{\pi}\right)^{3/4}e^{-\alpha_i r^2} \tag{35}$$

so that $\int\rho_A^{ASA}\left(\mathbf{r}\right)=1$.

Alternatively, the spherical *1S* functions, {S_i}, can be atom-centred, that is, positioned
at the atomic coordinates {$\mathbf{r}_a$}.

$$\rho^{ASA}\left(\mathbf{r}\right)=\sum_{i\in A}w_i\left|S_i\left(\mathbf{r}-\mathbf{r}_a;\alpha_i\right)\right|^2 \tag{36}$$

where the basis functions are defined as: $S_i\left(\mathbf{r}-\mathbf{r}_a;\alpha_i\right)=\left(\frac{\alpha_i}{\pi}\right)^{3/4}e^{-\alpha_i\left(\mathbf{r}-\mathbf{r}_a\right)^2}$

The ASA coefficients, w_i, are restricted to accomplish the **convexity conditions** in order to obtain positive definite values. These constraints preserve the statistical meaning of DF, providing a fitted DF with the suitable features of a probability distribution, which has a physical quantum mechanical meaning.

$$\left\{w_i\in\mathbf{R}^+\ \forall i\ \wedge\ \sum_{i\in A}w_i=1\right\} \tag{37}$$

To ensure a positive definite DF in the whole domain, the ASA coefficients must be positive definite: $w_i\geq 0\ \forall i$, so that $w_i\in\mathbf{R}^+,\forall i\in\mathbf{A}$.

Furthermore, the coefficients can be also normalised to the number of electrons of the molecule, *N*:

$$\sum_i w_i=N \tag{38}$$

When the latter normalization condition is imposed, the integral of the DF in the whole space gives the total number of electrons. Also, the expansion coefficients are the occupation numbers, n_i, for the corresponding atomic shells:

$$\sum_{i\in A} w_i \int \left|S_i\left(\mathbf{r}\right)\right|^2 d\mathbf{r} = \sum_{i\in A} n_i = N \tag{39}$$

Therefore, the molecular electron density can be pictured as a superposition of spherical atomic shells, whose occupations are the variational coefficients.

The simplified electron density clouds, constructed as linear combinations of Gaussian type functions can be obtained from a fitting procedure and readily evaluated. The procedure consists on the optimization of the set of coefficients of the linear expansion $\left\{w_i\right\}$ by minimizing the **quadratic error integral function** between the *ab initio* and the approximated density function, $\varepsilon^{(2)}$, while conserving the weighting coefficients positive definite.

The set of coefficients that minimizes $\varepsilon^{(2)}$ is obtained by solving the linear equation system:

$$\varepsilon^{(2)} = \int \left|\rho_A(\mathbf{r}) - \rho_A^{\text{ASA}}(\mathbf{r})\right|^2 d\mathbf{r} \tag{40}$$

Nevertheless, this approximation has some disadvantages. First, the number of atoms to be calculated is restricted by the limit of the *ab initio* calculation. Thus, the previous computations make the study of macromolecules cumbersome. Then, the adjustment of the *ab initio* density function to the ASA one is computationally expensive, especially with increasing molecular size or number of basis functions. Finally, the calculations cannot be used several times; instead, the whole adjustment for each molecule must be done each time.

3.5.2 Promolecular Atomic Shell Approximation (PASA)

The main limitation in any adjustment method of DF is the *ab initio* DF calculation. This problem is not especially significant for atoms; however, for big molecular systems with a large number of particles the capacity of computation is critical. In molecular quantum similarity studies, in addition to the size of the molecules studied, the aforementioned optimization of the pairwise relative position of all the compounds studied results in nonviable *ab initio* calculations. In order to avoid costly molecular *ab initio* calculations, a promolecular approximation [146-148] has been implemented to the previous development.

The **Promolecular ASA (PASA)** density function represents the atoms in a molecule as neutral entities of spherical shape, with a radial dependence equal to the isolated atoms, providing a precise three-dimensional electron distribution. The promolecular description of the charge density distribution employed is based on the sum of atomic ASA densities, previously fitted to an *ab initio* atomic basis set. Although the schemes providing a partition of a given molecule into its atomic components are not possible to derive from the quantum mechanical postulates, the concept of a promolecule has been used in many theoretical electron density distribution analyses. These theoretical schemes have been useful to obtain chemical information belonging to bonding interactions between atoms.

The PASA approximation [30-31,33,107,149], considers molecular densities as a sum of discrete contributions formed by atomic densities. The independent atomic contributions, $\rho_a^{ASA}(\mathbf{r})$, are generated with parameterized ASA atomic densities:

$$\rho_A^{PASA}(\mathbf{r}) = \sum_{a \in A} P_a \rho_a^{ASA}(\mathbf{r}) \tag{41}$$

where $\rho_A^{PASA}(\mathbf{r})$ is the promolecular density function, and P_a represents the total charge over the atom *a,* and is usually approximated by the atomic number, Z_a.

The atomic density functions fitted to an *ab initio* basis set, are built by means of a linear combination of *1S* Gaussian functions:

$$\rho_a^{ASA}(\mathbf{r}) = \sum_{i \in a} w_i \left| S_i(\mathbf{r}) \right|^2 \tag{42}$$

where the subindex *i* symbolizes the functions in atoms, *a* and *A* represent atoms, and molecules, respectively, $\{w_i\}$ are the adjusted PD coefficients of the linear expansion, and $\{S_i\}$ the gaussian *1S* functions.

The coefficients and exponents of the adjusted atomic DF are also obtained by minimizing the measure of the integral of the quadratic error functions in relation to the *ab initio* atomic densities. Once calculated the atomic densities, they are stored in a database to construct the molecular functions.

In this case, the convexity conditions are:

$$\left\{ w_i \in \mathbf{R}^+ \;\forall i \;\wedge\; \sum_{i \in a} w_i = 1 \right\} \tag{43}$$

The coefficients are normalized to the unity, $\sum_{i \in A} w_i = 1$, due to the fact that each DF is weighted by P_a, preserving the statistical meaning of the charge distribution definition, normalized to the total number of electrons of the molecule. Substituting the atomic DF in the molecular DF definition and integrating over the whole space:

$$\int \rho_A^{PASA}\left(\mathbf{r}\right) d\mathbf{r} = \sum_{a \in A} P_a \sum_{i \in a} w_i \int \left| s_i \left(\mathbf{r} - \mathbf{r}_a\right)\right|^2 d\mathbf{r} = \sum_{a \in A} P_a \sum_{i \in a} w_i = \sum_{a \in A} P_a = N \tag{44}$$

Accordingly, the normalization condition is accomplished, whereas in the atomic adjustments the ASA functions have been normalized to the unit.

The promolecular approach considers molecular densities like an ensemble of superposed atomic densities. This approximation is confirmed by the accuracy of PASA densities, which have been demonstrated to be sufficiently efficient for QSM purposes. The main advantage of this approximation is that only atomic density functions are adjusted to an *ab initio* calculation, instead of fitting the whole density function. This allows constructing the molecular density simply by adding the adjusted atomic density functions in the given geometry. For each atom, density functions are retrieved form a database, so that coefficients and exponents are calculated once and, afterwards, to obtain the PASA DF, only the molecular coordinates and the data for the atom are needed. In this sense, the PASA approximation resembles the semiempiric methods; as long as the latter use previously calculated parameters, the former takes the previous fitting to build density functions. So the PASA can be considered as a semiempiric method that provides adjusted density functions.

Furthermore, the PASA DF can be also used as a starting point for molecular adjustments [34]. In this way, only a refinement of the coefficients to be adapted to the molecular environment is needed, and the procedure is effectively accelerated.

Using the PASA approximation, the calculations for atomic systems are affordable, even using complex methods and large basis sets. Only the atom type, the atomic coordinates and an adjusted basis of atomic functions are needed. However, the obtained ASA density functions still have some **limitations**, for instance, the incomplete description of the charge migration in the regions surrounding the nucleus. Hence, most of the atomic density is concentrated in the nuclei, with low diffusion. The collapse of the DF upon the atomic nuclei influences the description of the bond formation in the molecules: at some levels, the *ab initio* methods allow the visualization of the interatomic bonds, while the ASA densities cannot. This is an intrinsic problem of using only spherical functions. The assumption that atoms in molecules are of spherical shape, as free atoms in gas phase, presupposes some limitations in the description of the molecular electronic charge distribution. Essentially, the deformations produced around bonded atoms cannot be modulated using only spherical functions centred on atomic coordinates. However, if the procedure is adapted for spherically symmetrical and nuclear centred fitting functions improved by locating additional fitted spherical functions outside atomic nuclei, a better description of the detailed structure of molecular electron density can be obtained. To such an extent, supplementary functions, with centres that differ from the atomic coordinates, have been aggregated in order to improve ASA densities. The so-called **Multicenter ASA (MASA)** densities try to simulate the electron cloud deformation of atoms in a molecule and to remove some of the deficiencies of the spherical representation of ASA densities.

In addition, the promolecular density presents some lacks in the description of bond regions. These deficiencies have been evidenced by means of the density deformation maps, formulated as the difference between the molecular electron density and the superposition of the ground-state densities of the atoms in a molecule.

3.5.3 ASA adjustment method

In the **ASA fitting algorithm,** the parameters to be optimized are the coefficients of the expansion $\left\{w_i\right\}$, and the exponents $\left\{\alpha_i\right\}$. The k number of functions, i.e., atomic shells, is arbitrarily defined.

$$\rho_A^{ASA}\left(\mathbf{r}\right)=\sum_{i\in a}^{k}w_i\left|S_i\left(\mathbf{r},\alpha_i\right)\right|^2 \tag{45}$$

The first adjustment technique to fit the ASA DF was a variant of the **least squares method**, using a Lagrange multiplier to optimize the coefficients, and keeping the normalization condition in order to obtain positive definite coefficients [30,144]. The proposed method assigns a saturated basis of ASA functions [150] to the considered system. Then, the algorithm selects the shells or functions with the optimal exponents and the PD coefficients.

The initial ASA exponents are generated by means of a geometric even-tempered sequence [151-153]. The use of even-tempered basis functions in the least squares fitting provides truncated basis sets with positive coefficients and, thus, the fitted densities belong to a subset of all the functions generated by the even-tempered series.

Also, the initial ASA coefficients of basis functions are required to minimize the $\varepsilon^{(2)}$ function. With such purpose, the weighting coefficients are optimized using the **Elementary Jacobi Rotations (EJR)** technique [154-155]. The EJR technique is a norm-conserving procedure based in orthogonal transformations that transforms a vector conserving its norm. It was initially developed in the adjustment of atoms [30,33,107], and, afterwards, it has also been applied to molecules [34]. The distinctive features of the designed algorithm is the definition of a new vector of coefficients, **x**, that generates the ASA coefficients: $w_i = |x_i|^2 \; \forall i$. Consequently, the elements of the **w** vector are positive and, moreover, the norm of **x** vector is the unit: $\sum_i x_i^2 = \sum_i w_i = 1$, so that the two convexity conditions are fulfilled. In order to accelerate the optimization algorithm of the ASA coefficients an alternative development of the Taylor *truncated* series has been implemented when applying the EJR transformation to get trigonometric functions. Finally, the exponents of basis functions are refined and optimized using a Newton-type method [156].

Pnce the atomic basis function has been parameterized employing the promolecular approximation, only the atomic coordinates need to be known and the parametrized ASA function set must be specified to generate automatically the electronic density of any molecule. Different atomic ASA basis functions have been adjusted for several basis sets; ASA exponents $\{\alpha_{i;a}\}$ and expansion coefficients $\{w_{i;a}\}$ for several atomic basis sets can be found in a WWW page [145]. Among the calculated databases, some of them can be mentioned:

- ASA functions for atoms H to Kr, fitted to an *ab initio* HF calculation with the **3-21G basis set** [31,157-160]. The atoms have been described using more than a set of ASA functions: a single *1S* Gaussian function for H and He, 3,4, and 5 functions for the series Li-Ar, and 5 functions for the series K-Kr.

- ASA adjusted coefficients and exponents for atoms H to Rn, adjusted to a **432-Huzinaga basis**, specially indicated for heavy atoms [33,161,162-164].
- ASA functions for H to Ar adjusted to the **6-21G basis set** [158-159].
- ASA functions for H to Ar adjusted to the **6-311G basis** set [165-166].

The main objective of the ASA adjustment is the description of molecular densities with a low computational cost and good accuracy, to obtain an efficient calculation of MQSM. In order to achieve their suitability, comparative studies between MQSM generated from *ab initio* DF, and MQSM derived from ASA DF have been performed with different basis sets, and compared by means of isodensity surfaces at different levels [21,31-32,145-149]. Taking into account the reduction in the complexity of the expression and in the computational cost, in all the studied cases, the differences between the similarity measures calculated using *ab initio* and ASA DF were less than 2% [34], and so the ASA electron density can be considered sufficiently accurate for the practical implementation of quantum similarity measures.

4 MOLECULAR QUANTUM SIMILARITY MEASURES (MQSM)

4.1 General Definition of MQSM

Quantum similarity measures are based on the psychological perception of similarity and they are founded in the obvious **similarity principle**: “the more similar are two molecules, the more similar are their properties”. This affirmation requires a procedure to compare the two molecules: quantum similarity attempts to give a quantitative measure of the degree of similarity between two quantum objects, basing on the comparison of their densities [167].

To quantify the degree of similarity between the compared systems, the previously mentioned statistical expectation value technique can be employed. Thus, a general **Molecular Quantum Similarity Measure (MQSM)** can be defined by means of an integral measure computation between the DFs attached to the involved molecular systems [15]. The DFs are multiplied and integrated over the electronic coordinates in a convenient domain. So, MQSM can be also defined as the scalar product between the first order molecular DFs associated to the compared molecules, and weighted by a bielectronic non-differential and positive definite operator.

$$Z_{AB}(\Omega) = \langle \rho_A | \Omega | \rho_B \rangle = \iint \rho_A(\mathbf{r}_1)\ \Omega(\mathbf{r}_1,\mathbf{r}_2)\, \rho_B(\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \in \mathbf{R}^+ \tag{46}$$

where *A* and *B* are the two quantum objects of study, $\{r_1, r_2\}$ are the sets of electron coordinates associated with the corresponding wavefunctions, $\{\rho_A, \rho_B\}$ the corresponding first-order density functions or electron probability densities, and $\Omega(\mathbf{r}_1,\mathbf{r}_2)$ the positive definite weighting operator, depending on two-electron coordinates.

Within the LCAO-MO approach, the MQSM at the Hartree-Fock level is expressed as the cumbersome four-center integral:

$$Z_{AB}(\Omega) = \sum_{\mu\in A}\sum_{\nu\in A}\sum_{\lambda\in B}\sum_{\sigma\in B} P_{\mu\nu}P_{\lambda\sigma}\int\int \phi_\mu^*(\mathbf{r}_1)\phi_\nu(\mathbf{r}_1)\Omega(\mathbf{r}_1,\mathbf{r}_2)\phi_\lambda^*(\mathbf{r}_2)\phi_\sigma(\mathbf{r}_2)\ d\mathbf{r}_1 d\mathbf{r}_2 \tag{47}$$

These integrals can be readily evaluated if approximated electronic density functions are used instead of *ab initio* ones, as has been shown in the previous section. If an ASA adjusted density function is used, the quantum similarity measure is reduced to:

$$Z_{AB}\left(\Omega\right)=\sum_{i\in A}\sum_{j\in B}w_i w_j\int\int\left|S_i\left(\mathbf{r}_1\right)\right|^2\Omega\left(\mathbf{r}_1,\mathbf{r}_2\right)\left|S_j\left(\mathbf{r}_2\right)\right|^2 d\mathbf{r}_1 d\mathbf{r}_2 \qquad (48)$$

By construction, due to the presence of the positive definite integrands, i.e. the operator and DF, the values of the integral are always real and positive definite: $Z_{AB}\left(\Omega\right)\in\mathbf{R}^+$

From the practical point of view, given two molecules and supposing that the respective wavefunctions $\left\{\Psi_A,\Psi_B\right\}$ are known, the Schrödinger equation is solved at an arbitrary level for both molecules, and the density matrix connected with the wavefunction pair can also be computed. Then, given a set of n objects M and their corresponding density functions, there is always the possibility of computing the whole array of QSMs between molecular pairs. The global set of QSM, which compares all the possible pairs of quantum objects, is also expressed in matrix form, by means of the quantum **Similarity Matrix (SM)**: $\mathbf{Z}=\left\{Z_{ij}\right\}$, where $i,j\in\left[1,n\right]$. The similarity matrix of dimension $\left(n\times n\right)$ is defined as:

$$\mathbf{Z}=\left\{Z_{ij}\left(\Omega\right)\middle|\forall i,j\in M\right\} \qquad (49)$$

This matrix can be also considered as an hypervector formed by a set of column vectors: $\mathbf{Z}=\left\{\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_n\right\}$, where each column (or row) vector, $\mathbf{z}_a$, is formed by the collection of all the QSM related to the quantum object a, that is all the QSMs between the a-th molecule and each element of the set, including itself. Consequently, every vector $\mathbf{z}_a$ is interpreted as a discrete N-dimensional representation of the a-th structure.

$$\mathbf{Z}=\begin{array}{c|ccccccc|}
 & 1 & 2 & 3 & \ldots & i & \ldots & n\\
1 & z_{11} & z_{12} & z_{13} & \ldots & z_{1i} & \ldots & z_{1n}\\
2 & & z_{22} & z_{23} & \ldots & z_{2i} & \ldots & z_{2n}\\
3 & & & z_{33} & \ldots & z_{3i} & \ldots & z_{3n}\\
\vdots & & & & \ddots & & & \vdots\\
j & & & & & z_{ii} & & z_{jn}\\
\vdots & & & & & & \ddots & \vdots\\
n & & & & & & & z_{nn}\\
 & \downarrow & \downarrow & \downarrow & & \downarrow & & \downarrow\\
 & \mathbf{z}_1 & \mathbf{z}_2 & \mathbf{z}_3 & & \mathbf{z}_i & & \mathbf{z}_n
\end{array}$$

Figure 2. Generic quantum similarity matrix, ***Z***, for n quantum objects, with the $n\left(n+1\right)/2$ pairwise calculations for the upper triangle.

The elements of the squared SM are the similarity measures between all the possible pairs of density functions of the considered molecular data set. Any homogeneous SM is symmetric, so the QSM between two molecules is identical independently of the order of the comparison of the QO. This fact is used to compress the information only in the upper or the lower triangle of the matrix. The order of magnitude of the different types of QSM is highly connected to the structural form of the molecule, and to the presence of heavy atoms. Due to the particular construction of the SM, the diagonal elements of SM bring out information on the size of the compound.

Quantum similarity matrices contain all the information of the system, act as discrete representations of quantum objects, and subsequently their elements can be used for the generation of molecular quantum descriptors, where every descriptor, $\mathbf{z}_i$, is collected in the columns of the SM. Similarity matrices are universal in the sense that it can be obtained from any molecular set and for any molecule in the set, and unbiased, because in the building process, there are no other choices than those provided by the knowledge of the involved DFs and the QSMs.

4.2 Types of Molecular Quantum Similarity Measures

Depending on the information being requested, several weight operators can be chosen, producing different types of MQSM:

4.2.1 Overlap QSM

The simplest and most intuitive usual election for the positive definite weight operator is the **Dirac's delta distribution**, $\Omega(\mathbf{r}_1,\mathbf{r}_2)=\delta(\mathbf{r}_1,\mathbf{r}_2)$. This choice transforms the general QSM definition into the so-called overlap-like QSM [15], which gives a measure of the volume enclosed in the superposition of both electronic density functions:

$$Z_{AB}(\Omega)=\iint \rho_A(\mathbf{r}_1)\ \delta(\mathbf{r}_1-\mathbf{r}_2)\rho_B(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2=\int \rho_A(\mathbf{r})\ \rho_B(\mathbf{r})d\mathbf{r} \tag{50}$$

The Dirac's delta function provides a physically intuitive definition, and it is computationally affordable. The overlap-like MQSM provides information on the concentration of electrons in the molecule, and indicates the degree of overlap between the compared molecules. For heavy elements and small interatomic distances, the overlap-like QSM increases.

4.2.2 Coulomb QSM

If the Ω operator is adopted to be the **Coulomb operator**, $\Omega\left(\mathbf{r}_1,\mathbf{r}_2\right)=\left|\mathbf{r}_1-\mathbf{r}_2\right|^{-1}$, this yields the Coulomb-like MQSM [136], which represents the electronic repulsive Coulomb energy between the two charge densities:

$$Z_{AB}\left(\Omega\right)=\iint\rho_A\left(\mathbf{r}_1\right)\ \frac{1}{\mathbf{r}_1-\mathbf{r}_2}\rho_B\left(\mathbf{r}_2\right)d\mathbf{r}_1 d\mathbf{r}_2 \tag{51}$$

The Coulomb operator acts as a weight for the density functions overlapping. Considering the molecular density function as an electron distribution in the space, this expression is anything but the extension of the Coulomb law for continuous charge distributions, and therefore, it can be considered, in some sense, as an electrostatic potential descriptor. This operator gives a measure on the Coulomb bielectronic repulsion between electronic distributions, and it is associated to electrostatic interactions.

4.2.3 Kinetic Energy QSM

This definition employs the **Kinetic Energy Density Function (KE DF)**, instead of the electronic DF [168]:

$$\kappa\left(\mathbf{r}\right)=\sum_i w_i\left|\nabla\varphi_i\left(\mathbf{r}\right)\right|^2 \tag{52}$$

The KE DF can be derived from the statistical expectation value technique:

$$2\left\langle K\right\rangle=-\int\Psi^*\nabla^2\Psi d\mathbf{r}=\int\left(\nabla\Psi\right)^*\left(\nabla\Psi\right)d\mathbf{r}=\int\kappa\left(\mathbf{r}\right)d\mathbf{r} \tag{53}$$

Using KE DF, similarity measures are expressed as:

$$Z_{AB}=\int\kappa_A\left(\mathbf{r}_1\right)\Omega\left(\mathbf{r}_1,\mathbf{r}_2\right)\ \kappa_B\left(\mathbf{r}_2\right)d\mathbf{r}_1 d\mathbf{r}_2 \tag{54}$$

When the overlap operator is used, the measure also brings information on the shared volume between both distributions:

$$Z_{AB}=\int\kappa_A\left(\mathbf{r}\right)\kappa_B\left(\mathbf{r}\right)d\mathbf{r} \tag{55}$$

Although it has been successfully applied in QSAR studies [169], this QSM is still in development phase.

4.2.4 Molecular Quantum Self-SM

The **Molecular Quantum Self-Similarity Measure (MQS-SM)** is obtained, independently of the operator, when comparing a system with itself. It is related with the electronic charge density occupation in the space, that is, it provides information on the charge concentration of the considered QO. So, this type of measure is useful to distinguish local molecular differences from a charge density concentration point of view. Taking $\rho_A = \rho_B$ in the general definition,

$$Z_{AA}(\Omega) = \iint \rho_A(\mathbf{r}_1)\ \Omega(\mathbf{r}_1, \mathbf{r}_2)\, \rho_A(\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \qquad (56)$$

When the selected operator is the overlap operator, self-similarities can be considered as the square of the norm of density function in the chosen metric, and they can be extracted from the diagonal of quantum similarity matrices.

$$Z_{AA}(\Omega) = \langle \rho_A | \rho_A \rangle = \|\rho_A\|^2 \qquad (57)$$

Z_{AA} quantities have been used as simple molecular descriptors in QSAR analysis, in order to describe certain molecular properties such as hydrophobicity, and the electronic effects produced by substituents [170-173].

4.2.5 Fragment Quantum Self-SM

This is a particular case of QS-SM, obtained from the comparison of particular substructures of two identical objects. That is to say, this measure is calculated taking the corresponding part of the squared norm of the density function belonging to the fragment of interest. Self-similarities occupy the diagonal of similarity matrices, as in the preceeding case:

$$Z_{AA}^{X}(\Omega) = \iint \rho_A^{X}(\mathbf{r}_1)\ \Omega(\mathbf{r}_1, \mathbf{r}_2)\, \rho_A^{X}(\mathbf{r}_2)\, d\mathbf{r}_1 d\mathbf{r}_2 \qquad (58)$$

The use of molecular descriptors defined upon molecular fragments [174-176] is partly based on the holographic theorem of the electronic density, which assures that the information contained in the total electronic density of a molecule is also present in the local density of any molecular fragment [125]. Provided that a wealth of molecular properties is usually closely associated with the presence of certain molecular fragments or functional groups, which can be regarded as pharmacophores, it can be assumed that focusing on the active fragment rather than on the whole molecule can improve the structure-function relationship.

In particular, fragment QS-SM can aid to the rationalization of the effect of systematic structural variation of substituents on the observed biological activity in the series of structurally related molecules. Indeed, they provide the detection and localisation of the fragment most likely to be responsible for the observed activity.

4.2.6 Triple-density QSM

In this particular case, the Ω operator is replaced by a third density function, $\Omega = \rho_C(\mathbf{r}_3)$:

$$Z_{AB,C} = \iiint \rho_A(\mathbf{r}_1)\rho_C(\mathbf{r}_3)\rho_B(\mathbf{r}_2)d\mathbf{r}_1 d\mathbf{r}_2 d\mathbf{r}_3 \quad (59)$$

This measure relates the volume shared among three density functions, allowing in this way to measure the similarity between two objects *A* and *B*, taking a third object *C* as a reference. Given a set of *n* objects, there are *n* possible triple density similarity matrices, where each member of the set acts as an operator [142,177].

4.2.7 Tuned QSM

This QSM simultaneously combines different operators in the calculation. Thus, taking linear combinations of different measures, a new similarity matrix can be obtained, enhancing the characterization of the system [178-179]:

$$\mathbf{Z} = \sum_{i=1}^{m} c_i \mathbf{Z}(\Omega_i) \quad (60)$$

where $\mathbf{Z}(\Omega_i)$ are the quantum similarity matrices obtained with the previous definitions, c_i are the expansion coefficients, and *m* the number of similarity matrices. In order to preserve the properties of similarity measures and with the aim to clarify the interpretation of the resulting matrix, the convexity conditions are imposed to the coefficients $\{c_i\}: c_i \geq 0, \forall i \ \wedge \ \sum_{i=1}^{m} c_i = 1$. These constraints allow the direct association of the coefficient to the unitary percentage of contribution of each matrix, thus clarifying the interpretation of the resulting matrix.

Given *m* initial similarity matrices, there are infinite combinations of coefficients to obtain infinite new similarity matrices. However, the calculation of coefficients in tuned similarity matrices maximizes the capacity of prediction of the model.

4.2.8 Boltzmann QSM

Another problem related with molecular structures is the conformational degree of freedom. QSM depend on molecular geometries and this means that, for compounds with conformational freedom, a particular conformer must be chosen. Different conformers of the same type can yield different results when performing a QSM calculation because a small variation in the geometry of the molecules compared may substantially vary the similarity measure. Usually, as an arbitrary decision criterion, the conformer used is the one with the lowest energy, unless the studied property is explicitly requires a specific active conformer.

A rational alternative would be taking several representative conformers for each molecule, obtaining an amplified similarity matrix. Another possibility more rigorous would be performing a **combined similarity measure with a Boltzmann term** [180], depending on the energy:

$$Z_{AB}^{(c)} = Z_{AB} e^{-\left(c_A \Delta E_A / RT \right)} e^{-\left(c_B \Delta E_B / RT \right)} \tag{61}$$

where Z_{AB} is a QSM between two conformers of molecules A and B, and c_A and c_B are the weights of each conformer, ΔE_A and ΔE_B the differences of energy of the two conformers in relation to the conformer of minimum energy, R the gases constant, and T the temperature. This would yield a global similarity measure for both molecules, optimizing simultaneously the position and the energy. The only disadvantage of these calculations is their expensive computational cost.

4.2.9 Other QSM

In the bibliography, other types of MQSM have also been defined and applied for the generation of QSAR descriptors. Some of them are the **electrostatic potential** QSM [181-182], **interelectronic repulsion energy** MQSM [183-185], and **gravitational** QSM [186] and **Cioslowski** QSM [20,53,187], defined respectively as the squared Coulomb and Overlap QSM.

Recently, self-similarity measures on **Fermi hole densities** [188], and similarity measures using **momentum-space electron densities** [60-63,189-191], defined from momentum-space wavefunctions instead of spatial wavefunctions have also been defined. In this case, the position wavefunction is transformed into the momentum-space one by means of a Fourier transformation:

$$\Psi(\mathbf{p}) = \frac{1}{(2\pi)^{3/2}} \int \Psi(\mathbf{r}) e^{-i\mathbf{pr}} d\mathbf{r} \tag{62}$$

The electronic density, the individual orbitals, and the basis functions are related in the space *p* in an analogous way as in the space *r*.

4.2.10 QSM defined over other QO

In addition, as a methodology based on quantum mechanics, quantum similarity can be applied to any microscopic system, i.e. quantum object, described by a Hamiltonian. Apart from molecules, QSM have been extended to other quantum objects, for instance, quantum similarity measures have been described between **atoms** [192, 257], **atomic nuclei** [23,193-194], **atomic** and **molecular orbitals** [195], and **second order** (intracule and extracule) **densities** [196-197].

The comparison between the values of similarity measures among different types of operators has been reported in the bibliography [20,187, 202].

4.3 Normalization and mathematical transformations of QSM

Once the quantum object set and the weight operator have been defined, the MQSM becomes unique. Nevertheless, the value of the similarity measure between two objects does not report conceptual information on the degree of the similarity of the compared objects. For that reason, the similarity matrix elements can be transformed and combined to obtain normalized or scaled values. These mathematical transformations on the QSM yield the so-called **Molecular Quantum Similarity Indices** (**MQSI**), or more generally, **Quantum Similarity Indices** (**QSI**), which can be numerically manipulated in an easy way, and intuitively interpreted.

QSI can also be used as molecular descriptors in QSAR studies, in the same way as QSM. In fact, QSI do not produce new information on molecular similarity relationships, so the election of any one of them should not be decisive for the derived results. Several transformations of similarity measures can be described, but the most common are the Carbó index and the Euclidean distance index.

4.3.1 Carbó Index

This is the most used index**,** defined in the seminal paper [15]. This index is equivalent to the normalization of the similarity measure value Z_{AB} in relation to the self-similarities of A, Z_{AA}, and B, Z_{BB}. Mathematically, it can be expressed as:

$$C_{AB} = \frac{Z_{AB}}{\sqrt{Z_{AA}Z_{BB}}} \tag{63}$$

C_{AB} varies in the interval (0,1]. The nearer to the unit, the more similar are the compared objects, while a value approaching to zero indicates that the two objects are dissimilar. The exact unity value is only obtained when both compared objects are the same, that is, in the case of self similarity measures, where $C_{AB} = 1$, that is, an object is identical to itself.

Geometrically, the Carbó index can be interpreted as the cosine of the angle subtended by the involved density functions, considered in turn as vectors. The Carbó index is a correlation-like or cosinus index, also called **C-class** index.

$$C_{AB} = \frac{\langle \rho_A | \rho_B \rangle}{\|\rho_A\| \|\rho_B\|} = \cos \alpha_{AB} \tag{64}$$

4.3.2 Euclidean distance index

This is another typical transformation [136] that can be defined according to the classical definition of distance:

$$d_{ab} = \left[\sum_{j=1}^{p} \left(\Delta x_j \right)^k \right]^{1/k} \tag{65}$$

where $\Delta x_j = x_{a_j} - x_{b_j}$ is the distance between the objects *a* and *b*, and $k = 2$ for the Euclidean distance definition. So, the Euclidean distance between two any QO *A* and B is defined as:

$$d_{ab} = \sqrt{\left(x_a - x_b \right)^2} \tag{66}$$

Therefore, the Euclidean distance index is expressed as:

$$D_{AB} = \sqrt{Z_{AA} + Z_{BB} - 2Z_{AB}} \tag{67}$$

D_{AB} is comprised within the interval [0,∞) but, conversely to the previous case, values close to zero imply a greater similarity between the compared objects. So, if the two compared objects are identical, D_{AB}=0.

Geometrically, this index can be interpreted as the norm of the difference between the density functions of the compared objects. The Euclidean distance index can be defined as a distance or dissimilarity index, also called a **D-class** index.

$$D_{AB} = \left\| \rho_A - \rho_B \right\| = \sqrt{\left(\rho_A - \rho_B \right)^2} \tag{68}$$

4.3.3 Stochastic Transformation

Besides fromtransformations of QSM into QSI, another possible scaling can be performed by means of a stochastic transformation [198-199]:

$$S_{AB} = Z_{AB} \left(\sum_{C=1}^{N} z_{AC} \right)^{-1} \tag{69}$$

Such transform provides a stochastic SM, $\mathbf{S} = \left\{ s_{AB} \right\}$, where the sum of elements of each row has been used as a scale factor. This procedure creates an alternative uniform non-symmetric SM, whose columns can also be used as new descriptors for a given molecular set, and be interpreted as discrete probability distributions.

4.3.4 Other normalisation indices

Other common similarity indices defined in the bibliography are the **Hodgkin-Richards** [41], the **Tanimoto** [200], and the **Petke** [201] indices. Some studies comparing the QSM generated by different operators and several QSI have been reported in the literature [20,187 ,202].

4.4 Dependence of QSM on the relative orientation of the objects

Similarity measures have a strong dependence with the relative orientation of the objects compared. So, it is necessary to establish an alignment criterion to superimpose the structures in an appropriate way. The three-dimensional superposition implies the positioning of an object in relation to another object. In the particular frame of molecular similarity, this alignment can help to interpret and understand the molecular data.

The energy of a chemical system that allows the molecular recognition with a biological target is given by nuclear and electronic forces, and interaction energies of solvation. The energy also depends on the molecular conformation and weak interactions with the environment. These properties are intimately linked with the spatial disposition of the molecule. Then, if atoms with similar characteristics are located in the same positions, it can be assumed that they will elicit similar effects.

Those molecular regions that share the same space for a molecular series is generalized as the pharmacophore, that is, the common molecular region that due to its interactive characteristics is supposed to be the responsible for the activity of the molecules. Consequently, several techniques require systematized, effective, fast overlapping methods for the molecular alignment, in order to interpret the data.

The determination of the optimal molecular alignment has become a widely studied problem with application in several chemistry fields. Some of the most common uses of molecular superposition are applied to X-ray crystallography, and protein structural research. In particular, several applications are the quantification of the similarity between a given structure optimized with different methods, the quantitative comparison of molecular stereochemistry [203], the computer-based search for the determination of the distortion caused by chemical substitution in crystals [203-205], and pattern recognition for the determination of pharmacophores in three-dimensional structural databases [206-208]. Within the QSAR field [209], the COmparative Molecular Field Analysis (CoMFA) [210], and the techniques based on three-dimensional similarity [211-212], also demand efficient alignment methods.

The molecular superposition is conceptually simple and even intuitively evident when two similar molecules are compared. However, the practical computational implementation faces the combinatorial problem, provided that the number of possible ways to overlap two molecules may be high. Indeed, if different conformations are considered, the problem dealing with flexible bonds becomes more complicated, and even more when the size and the number of molecules increases. The molecular superposition techniques can be mainly classified into **atom-based techniques**, founded on the alignment of atoms, fragments, or common substructures, and **field-based techniques**, which superpose molecular force fields (i.e. electrostatic or steric), volumes or surfaces. Other employed techniques are genetic algorithms [213], techniques based on symmetry [214], molecular skins [215], and local overlaps [216].

The first works in molecular alignment were iniciated in the seventies, with the contributions of Gavuzzo [217], McLahan [218], Gerber and Müller [219], and Redington [220]. These works are based in the minimization or maximization of a function, which depends on the relative position of molecules in the space. For instance, the minimization of the weighted sum of distances between atoms [221], or the maximization of the molecular similarity measure, considered here.

The use of similarity indices as a basis for the alignment has also been proposed by means of the Monte Carlo method [222] and the gradient method [223]. Other possibilities are the flexible superposition of molecules [224], the superposition of dissimilar molecules [225], or the approaches related to graph theory that align the molecule according to subgraph coincidences [226-229].

In contrast, other methods take the values of common physicochemical properties as the basis for the superposition. These properties may be different molecular fields simulating steric, hydrophobic, shape or electrostatic effects [230-232], calculated in three-dimensional grids [216,233-234], or fields analytically adjusted [235-236].

The computation of similarity measures obviously requires the superposition of the two molecules being compared. The dependence of MQSM on the relative position of the molecular structure can be included in the general definition:

$$Z_{AB}\left(\Omega;\Theta\right)=\iint \rho_A\left(\mathbf{r}_1\right)\ \Omega\left(\mathbf{r}_1,\mathbf{r}_2\right)\rho_B\left(\mathbf{r}_2;\Theta\right)d\mathbf{r}_1 d\mathbf{r}_2 \tag{70}$$

where *A* and *B* are the studied molecules; $\rho_A\left(\mathbf{r}_1\right)$, $\rho_B\left(\mathbf{r}_2\right)$ the corresponding density functions, $\Omega\left(\mathbf{r}_1,\mathbf{r}_2\right)$ a positive definite bielectronic operator, and Z_{AB} is the resulting quantum similarity measure. The Θ operator represents the transformation of the coordinates of *B* in relation to the *A* coordinates, due to the similarity measure dependance on the relative position of both molecules in the space.

In the IQC, several methods for molecular alignment have been proposed. Initially, in the first works with polyatomic molecules, the structures were superposed in basis to the alignment of dipole moments [235-236]. However, this procedure, only valid for polar molecules, had serious drawbacks. The most relevant alignment algorithms used for molecular superposition in the calculation of QSM are the so-called maximum similarity superposition algorithm, and the topo-geometrical superposition algorithm.

4.4.1 Maximum Similarity Superposition Algorithm

The maximum similarity superposition algorithm was proposed by **Constans** [35], and implemented in the **MOLSIMIL** program [237]. This **field-based** method considers that the optimal alignment provides the maximum value of the similarity measure for a given similarity operator. Thus, the algorithm maximizes the value of the similarity integral to find the optimal superposition between each pair of molecules.

$$Z_{AB}(\Omega;\Theta) = \max_{\Theta} \iint \rho_A(\mathbf{r}_1)\ \Omega(\mathbf{r}_1,\mathbf{r}_2)\,\rho_B(\mathbf{r}_2;\Theta)\,d\mathbf{r}_1 d\mathbf{r}_2 \qquad (71)$$

being *A* and *B* the two compared molecules, and Θ the set of translations and rotations needed to superimpose in an optimal way the object *B* with the object *A*, considered as fixed in the space. Thus, Θ symbolizes the transformation of *B* coordinates in relation to the *A* coordinates, that is, the relative orientation between the two objects. This procedure provides a rigorous superposition criterion between two molecules, yielding different alternative alignments when different operators are used.

The method orientates the molecule *B* by means of translations and rotations, in a way that maximizes the similarity integral. Being $\{a,a',a''\}$ and $\{b,b',b''\}$ triads of the molecules *A* and *B*, respectively, first, the pair $\overline{ab}$ is exactly aligned in such a way that the axis described by the segments $\overline{aa'}$ and $\overline{bb'}$ coincide. Finally, the b'' atom is rotated so that the planes described by $\{a,a',a''\}$ and $\{b,b',b''\}$ coincide. This process is systematically repeated for all the possible triads of atoms in both molecules, keeping the value that maximizes the similarity measure. The algorithm includes an atomic similarity threshold that must be surpassed as a constraint to restrict the atoms to be compared.

To find the maximum of similarity an intensive search is performed, involving a large number of intermediate superpositions, where the similarity measure is evaluated. The superposition providing the maximum similarity is taken as the optimal. Provided that similarity measures are products between densities, the contribution at each point is highly sensitive to the existence of atomic nuclei and to the type of atom. As a consequence, this method can superimpose weighty atoms, due to their high concentration of the total molecular density. These elements act as density wells that compel the superposition in basis to them, putting aside possible common skeletons formed by several light atoms. So, when molecules contain heavy atoms up to the third period (such as Br or I), the optimization algorithm of molecular orientations might find as an optimal alignment the one that superimposes a heavy atom with the common skeleton, thus precluding the correct alignment of the common backbone. Thus, depending on the composition of the molecules being compared, this superposition does not provide intuitive alignments, since the density of the common molecular substructure is not important enough in comparison with the corresponding density of heavy atoms. Hence, in these cases, there may be a loss of chemical sense, reflected in low intuitive alignments. Some possible inconsistencies may be avoided by considering local similarities [238].

Besides, the algorithm performs an exhaustive search that consists of analyzing different positions and evaluating the similarity measure at each step. As a consequence, the method is computationally quite expensive. However, the computational cost can be reduced by simplifying the levels of the optimization cycles.

Alternatively, similarity can also be evaluated form other field types, apart from electron densities. Thus, to diminish the cost of the process, electronic densities can be approximated by expansions centered in the atoms -the statregy employed in the present case, with the calculation of ASA DF-, electrostatic potentials, steric surfaces or the volume, or representations based on grids [236]. All these methods overlap the coincident zones.

In order to avoid these superpositions chemically not relevant, an alternative algorithm has been proposed.

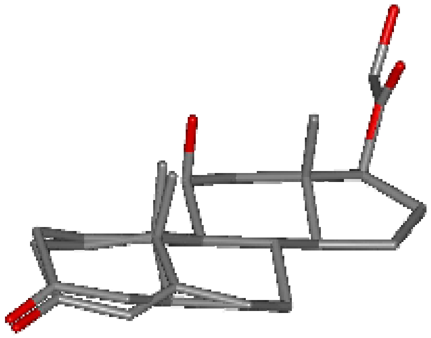


Figure 3. Alignment of steroids by the maximum similarity superposition algorithm.

4.4.2 Topo-Geometrical Superposition Algorithm (TGSA)

The **Topo-Geometrical Superposition Algorithm** (**TGSA**) was proposed by **Gironés** [36**],** and programmed and implemented by the same author [239].

In contrast to the maximum alignment superposition algorithm, this **atom-based** method considers that the optimal superposition aligns the molecules in basis to a common skeleton, taking only into account atomic types and interatomic bond lengths, that is, atomic numbers and coordinates. With such purpose, the algorithm examines all the atomic pairs within the molecule and aligns the largest molecular substructure common to a series of molecules, orienting the molecular matching towards the matching of such substructure. The method is only based on topological and geometrical considerations, where the molecular topology is embedded in the form of comparisons between bond distances. Given two molecules, the superposition is unique and it does not depend on the type of operator chosen to perform the similarity measures.

First, the molecular coordinates and the atomic numbers are read from the output from a pool of commercial programs. Once stored, the molecular coordinates are reordered in basis to the decreasing atomic number, in order to determine in a simple way the number of non-hydrogen atoms of the molecule. Considering that the superposition of hydrogens is not significant, and with the aim to save computational requirements, hydrogen atoms are not included in the process.

The next step consists of the definition of the atomic pairs, the so-called diads. A diad is defined only if the involved atom pairs are bonded and, it is thus determined by the number of bonds. Once the diads have been defined, within each molecule, every diad is compared to all diads of the other molecules by means of their interatomic distances, using a given length threshold. The threshold takes into account the fluctuations in backbone atom conformations produced by the presence of different substitutions within a molecule. This procedure allows discarding the bonds that do not belong to the common skeleton.

Once all the diads have been compared, the algorithm creates atomic triads by adding a third atom to the selected diads. This supplementary atom must be bonded to at least one atom of the considered diad. In geometrical terms, this step generates a triangle or a plane, where atoms occupy the vertices of the triangle, and the sides correspond to effective chemical bonds. The triangles obtained for one molecule are compared to all the triangles obtained for the second molecule by means of their interatomic distances, within the same distance threshold adopted in the diads comparison. If the three distances of both compared triangles are similar, both triads also are considered similar and stored. The triads that do not fulfill this criterion are automatically discarded. After having completed this comparison, the selected triads are superposed, and the resulting molecular alignment is univocally determined.

The process is repeated for all the atoms and the algorithm chooses the alignment that maximizes the number of atoms superimposed, minimizing the **fit index**. The fit index, C_{AB}, is used as the criterion to compare two interatomic distances, and it is calculated contrasting the absolute value of their difference with a fixed threshold:

$$C_{AB} = \sqrt{\frac{\sqrt{d_{AA} d_{BB}}}{d_{AB}}} \qquad (72)$$

where $d_{AB} = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} \left| \mathbf{x}_{i,A} - \mathbf{x}_{j,B} \right|^2$, *n* is the number of atoms and **x** the molecular coordinates.

C_{AB},, comprised within [0,1], evaluates the quality of superposition. It indicates a better alignment when C_{AB} approaches the unit, with the ideal case of two identical structures $C_{AB}=1$.

The TGSA method considers the molecules as rigid bodies, and does not allow flexibility in the structures (neither rotation nor variations in distances and bond angles). It is designed to operate with homogeneous sets of molecules, and does not yield good results with structurally diverse molecules, provided that the pairwise alignments are restricted to common skeleton recognition. In contrast, it is able to recognise a common substructure, thus providing a coherent alignment with chemical intuitive sense. The procedure is simple and has low computational requirements. To solve such drawbacks presented in the stuy of non-rigid, flexible conformations, the TGSA algorithm has been recently improved to enable handling rotatable bonds [240].

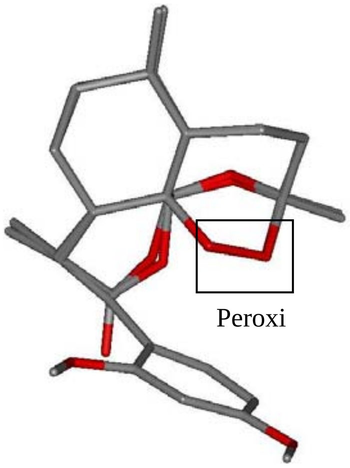


Figure 4. Pairwise molecular alignment between artemisinin and a derivative.

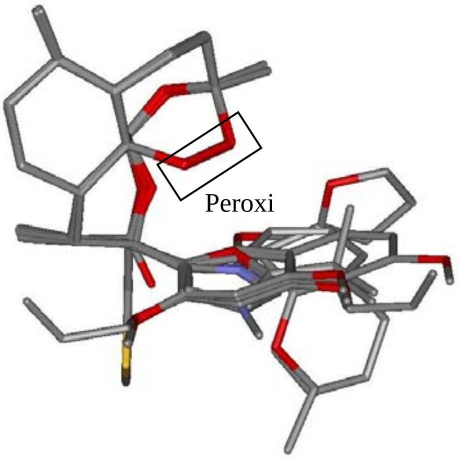


Figure 5. Molecular alignment solution for a molecular set of 17 synthetic 1,2,4-trioxanes over artemisinin.

As observed in *Figure 4,* both structures are aligned around the peroxy bridge present in artemisinin and the rest of the molecular structures belonging to the set. This fact is again evidenced observing *Figure 5,* where the entire molecular set is overlaid according to this substructure. Hence, the molecular skeleton has been detected and superposed [169].

4.5 Applications

Several applications of quantum similarity theory have been developed in different computational fields since the publication of the seminal paper by Carbó et al. [15]. Some of the most rellevant among them are outlined next:

Analysis of the charge distributions. The comparison and analysis of the charge concentration either in different species or in the same specie calculated at different levels allows the determination of the effect of solvation [243], and the comparative analysis between different methods of calculation [84,250-252].

Determination of the quality of a basis function set [253], study of the effect of the basis functions on similarity measures [254], and quantification and improvement of Density Functional Theory (DFT) parameters [255-256].

Determination of the optimal geometry optimization method. Given a molecular structure experimentally determined, its geometry can be optimized at different theoretic levels (*ab initio*, DFT, semi-empirical, or Molecular Mechanics (MM) level) with different basis. Afterwards, the similarity between densities obtained by different approaches and *ab initio* densities are calculated on a single structure, used as a template. Those methods that provide a density that resembles the more to the *ab initio* density of the original structure can be considered as suitable models to model compounds structurally related to the template [169].

Evaluation of atomic properties within a molecule [188192,257] from fragment self-similarity calculations between the different atoms constituting the molecule, and **Modelization of properties in atomic nuclei.** [23,193-194] from similarity measures between nuclear density functions [258-259].

Determination of transition states and the behaviour in intramolecular reactions [260-261] using similarity measures between reagents and products of reactions and the respective transition state. These calculations elucidate if a reaction is product-like or reactant-like, thus analyzing and quantifying the **Hammond postulate** [262]. More generally, also the assistance for the **Classification of reaction paths** [191], and the classification of other quantum objects [263] is possible.

Applications related with molecular reactivity [264], and **Quantification of the effects of substituents in chemical reactivity** [172, 174-175] from self-similarity measures of the functional group involved in the process, as well as the **Determination of molecular chirality** [265-266].

Priorization of orbitals for the calculation of truncated configuration interactions [267-268] from similarity calculations between atomic and molecular orbitals, and **Reduction of the number of cycles in SCF calculations** [269].

Finally, the most used application in our lab is the **Generation of descriptors for their use in QSAR**, which will be extensively revised in the following section.

4.6 Molecular Quantum Similarity in QSAR

Once established the theoretical basis of quantum similarity theory, quantum similarity was considered a valuable tool to generate molecular descriptors for their use in QSAR. The first work where similarity measures were applied to QSAR dates from the 1983, when a work of **Martin** et al was published [270]; afterwards, this work was extended to a qualitative study of molecular activity using **electrostatic potential distributions** [180-182,271-273]. In a similar way, **Richards** used Carbó indices obtained by comparison of electrostatic potentials as molecular descriptors for the construction of predictive models [51]. Richard also used electrostatic potentials as a basis for the calculation of similarities for their use in QSAR [52]. Besides, **graph theory** has also been included in the molecular similarity frame. Thus, **Rum and Herndon** [275] built a matrix of molecular similarity indices, comprised between 0 (completely dissimilar objects) and 1 (identical objects), whose columns were used as molecular descriptors in a multilinear regression model and provided a definition of molecular similarity based on graph theory [80]. **Good** described a protocol of application of similarity measures in QSAR [276-278] using reduction of dimensions and statistical validation of the results for the prediction models and the transformation of results into Carbó indices. Chemometric tools will be considered and revieed in more detail in a following section.

Following the research line headed by professor Carbó, initially, molecular quantum similarity was applied in a **qualitative** way in QSAR studies, associating the spatial groupings of molecules with the value of physico-chemical properties, and interpreting the different groupings as a function of molecular activities [143,279].

Afterwards, the existence of an operator relating the similarity measures with the molecular activity, that is, a physical observable, was postulated [241]. The practical implementation of these concepts in a theoretic protocol of application of quantum similarity matrices in QSAR has been afterwards established [149, 242-245]. Also, within the QSAR frame, the **connection between MQS and graph teory** has been theoretically examined and extensively applied [202], yielding the so-called **Topological Quantum Similarity Indices** (TQSI).

Finally, according to the application field of interest, MQS has been applied in QSAR problems with pharmacological insight, oriented to the rationalization and prediction of activity of drugs [169,178-179,241,280-281], in the evaluation of toxicity, in the so-called QSTR [24,187,185,282-283]. In particular, in QSPR studies, the constants associated to carboxylic acids have been described [174], as well as the stability of proteins [174].

5 CLASSICAL TOPOLOGICAL APPROACH

5.1 Introduction to Graph Theory

Graph theory is a subdiscipline of discrete mathematics closely related to both topology and combinatorics, which is concerned with the study of the mathematical properties of a structure and its components. In fact, graph theory provides a way to represent entities by means of **graphs**, schemes of connections between points by way of lines, in such a way that each line connects two points, thus providing the natural mathematical framework for the quantitative codification of classical chemical bonding ideas [284-288]. An account of the historical development of such ideas can be found in a review by Rouvray [289-290].

The application of graph theory to chemistry [291] results in **chemical graph theory** [292]**,** which can be considered as a branch of theoretical chemistry. It is concerned with molecular representation and handling **chemical graphs** [293], that is, structural models representing the topological structure of chemical systems. Hence, chemical graph theory analyses the connectivity in a chemical system and can be used to characterize structurally a compound by applying the topological method.

Topology is the part of algebra that studies the connections of elements within a set and their mutual position. Topology deals with all the various pieces of an object identified by breaking up its constitutive parts. Applied to molecules, topology has evolved to a new discipline called **molecular topology**. Molecular topology collects structural information concerning connectivity and ramification, mainly derived from structural intuitive information embedded in chemical systems. Therefore, what it is really important is how many atoms are in the molecule and their disposition and arrangement: if they form a linear or branched chain, a ring or a combination of them, and the number and type of connections between them, thus providing the description of the manner in which atoms are bonded in a molecule.

In chemistry, the representation of molecules by means of graph theory only provides information about the topological planar structure of the molecules. Hence, the topological analysis regards molecules as topological entities rather than geometric ones, so that the real three-dimensional morphology of the molecule is not important; the nature and length of chemical bonds and angles between bonds are of minor interest. Also, it is usually insensible to the type of atoms forming part of the molecule, although heteroatoms and multiple bonds can be also represented as vertex and edge-weighted molecular graphs [293-295] by means of the so-called topochemical parameters. These parameters are properly weighted graphs that represent the heterogeneity of atom types and bonding pattern.

The **classical topological approach** [296-297] relates the chemical structure constitution (the two dimensional model of a molecule represented by a structural formulae) with a non-dimensional quantitative scalar numerical entity characteristic of the compound. This univocally calculated mathematical invariant is the so-called **topological index (TI)**. To derive topological indices**,** the topological structure of molecules is represented by graphs. Characterizing a molecule by a single number represents a considerable loss of information: a three-dimensional object (molecule) is described by a one-dimensional object (topological index). However, relevant structural information is still retained in the topological index. To translate chemical structures into a single number, graph theory visualizes chemical structures as mathematical object sets [298] consisting of vertices, which symbolize atoms, and edges, linking a pair of vertices, which represent covalent bonds or shared electron pairs. In this notation, adjacent vertices stand for pairs of covalently linked atoms situated at a topological distance of one.

5.1.1 Historical Revision

The beginnings of classical graph theory date back from the first half of the eighteenth century, when **Euler** solved the so-called problem of the seven bridges of Königsberg. Euler published a discussion [299] on the probability of strolling around the city of Konigsberg, crossing each of its seven bridges that connect two islands and the mainland once and only once, without retracing their steps.

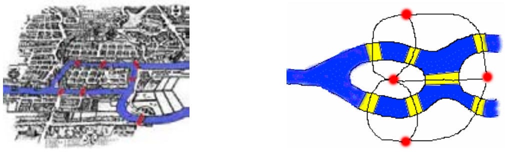


Figure 6. Illustration of the problem of the seven bridges of Königsberg [300].

Euler realized that all problems of this form could be represented by symbolizing the **areas of land** by **vertices**, and the **bridges** by **arcs** or connecting lines. From a diagrammatic graph representation, he proved that the solution was related to the geometry of position and that crossing the seven bridges in a single journey was impossible.

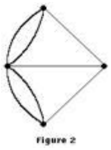


Figure 7. Diagrammatic representation of the problem of the seven bridges of Königsberg.

Independently, the scientist and philosopher Rugjer Josip **Bošković** [301] introduced the idea of representing atoms as points in space [302]. Bošković's fundamental idea was that substances have different properties because they have different structures. He estimated unknown quantities from emipirical equations, founding the origin of the **structure-property concept**. In this way, he was able to account for the existence of different substances. Bošković's model may be considered as the forerunner of a topological model for the structure of matter, although the word topology was not used until 1836 by **Listing,** who reported a paper [303] where he described the fundamentals of topology for quantification and modelling purposes.

The considerable convenience of having a two-dimensional depiction of molecular species readily available introduced chemists into two-dimensional thinking and neglecting three-dimensional implications. The structural formula was transformed into the mathematical device known as chemical graph. Indeed, mathematicians began taking interest in the structural formula because they recognized such formulas as examples of topological graphs.

Specially, the term graph, referring to diagrams showing analogies between chemical bonds in molecules and graphical representations of **mathematical invariants**, was coined by the mathematician **Sylvester** [304]. Sylvester introduced the term graph into both chemical and mathematical literature [305], and proposed the key idea of representing chemical graphs by mathematical invariants that could in turn be employed to characterize the chemical species from which they were originally derived. The invariants usually described as topological indices [306] have been employed for the prediction and interpretation of a wide range of physical and other properties of chemical species [307-308]. Thus, he demonstrated that a molecule can be represented by a connectivity or adjacency table [309].

Also, important applications of chemical graphs derive from the work of **Cayley**. Cayley had the idea of representing members of homologous series of molecules by graphs. This kind of representation enabled Cayley to enumerate the number of structural isomers for several chemical series [310-311].

Since the initial development of the seminal bases of topology, founded by Euler, graph theory evolved from the classical topological approach, firstly introduced by Boŝković, to nowadays. The work of several individuals from several different fields contributed to the development of graph theory, i.e. the applications of Gauss and Kirchoff in electrical circuits [312], the publication of Listing describing the Möbius band, among others. From that point, graph theory began to be used as a calculation tool to solve different and varied problems: Lord Kelvin was influenced by a work of Helmholtz on vortices, and the connectivity of surfaces was studied by Riemann. The connectivity in the three dimensional Euclidean space extended to *n* dimensions was examined by Betti, whereas Poincaré put the idea of connectivity on a rigorous basis and introduced the concept of homology in a series of papers in 1895.

Today, topology is an active field in modern mathematics. As an illustrative example, a curious topological problem that was recently solved was to determine how many colours are needed to colour an ordinary map so that no two regions that share the same border have the same colour. In 1976, Kenneth Appel and Wolfgang Haken used a computer to prove that four colours are sufficient, no matter how large the map is, or how many regions are in the map.

5.1.2 Applications

Graph theory is a field of mathematics that has a lot of applications in several fields, in addition to physics and chemistry. Within chemistry, graph theory has been applied to problems from a wide range of research areas. By describing the most general geometric features of molecules, topology and graph theory provide a suitable basis for answering the old question of how to derive the properties of chemical compounds from their structures. As a result, the construction of abstract chemical graph theory has become a very powerful tool for the topological characterization of chemical structures in such diverse areas as drug and new materials design, modelling of surface phenomena crystals and polymers, and assessing toxicities of chemicals in the environment.

In **organic chemistry**, the structure of molecules has been traditionally represented by the use of schematized diagrams representing only the backbone of carbons and heteroatoms. Graph theory allows the analysis of these graphs and the derivation of numerical quantities known as topological indices.

The interest of developing new graph descriptors for organic compounds revived in recent years, when topological indices found new applications in molecular similarity and diversity assessment [314], database mining, and virtual screening of combinatorial libraries [315-317]. In **drug design**, similarity and dissimilarity based methods have been very useful in the rational selection of candidate chemicals [318], database screening [319], and risk assessment [320].

The use of molecular similarity methods is based on the structure-property similarity principle. This notion states that **similar structures usually have similar properties** [10,321-322]. Intermolecular similarity can be defined in terms of the number of structural features and their mutual arrangements common to two chemical species [323]. The structural features used to quantify similarity vary with the level of organization to which the chemical species belong, atomic, molecular, macromolecular, etc. It is also dependent on the mode of representation of the species, choice of the set of structural descriptors, and the selection of the particular mathematical function used to quantify similarity from the chosen set of descriptors. Methods for quantitative molecular similarity analysis of chemicals give an ordered set of molecules (analogs) structurally related to the chemical of interest. The properties of the related analogs can then be used to estimate properties of the candidate chemical [318]. Similarity has been quantified using empirical and non-empirical properties or parameters. In particular, graph theoretical parameters such as topological indices have been used in the quantification of molecular similarity [10,321,324-329], with the aim to select efficiently some new active drugs from the hundred of thousands of compounds available from the data sources at the disposal of the pharmaceutical chemists.

There have been many attempts to design effective molecular structural indices, within the fields of QSPR, QSAR, and quantitative drug design. Topological indices derived from graph theory have been used as structural descriptors in QSPR and QSAR models [330-333]. Most applications have been in pharmacology and toxicology [334-336], but also many other physical and chemical molecular properties have also been modelled and predicted. Indeed, the selection of the adequate set of topological indices is not evident since numerous TIs have been described in the literature.

Various physicochemical parameters have been used jointly with connectivity indices, topological charge indices, electrotopological indices and geometrical indices in order to get solid models able to predict the pharmacological, biological, physical or chemical activity. Topological indexes have demonstrated their utility in the prediction of the diverse physical, chemical and biological and even analytical properties for different types of compounds, and on the design of new lead drugs.

Thus, one of the oldest and most successful QSPR approaches relies on the topological paradigm. Within this field, numerical correlation between quantities derived from topological graphs and physicochemical or biological properties usually produces acceptable results. The success of topological models relies in the possibility of estimating the properties of new chemical compounds from the molecular structure, encoded in a numerical form with the aid of various descriptors, i.e. molecular graph descriptors and topological indices [337-341].

The first applications of graph theory to QSPR models were introduced in 1947, by Wiener, which determined the boiling point of a set of aliphatic hydrocarbons [353351]. Afterwards, a number of well-known indices have been gradually introduced. The variety of these graph-theoretical descriptors has spectacularly increased in last decade; nowadays, in the literature, hundreds of topological indices, suitable to describe different properties, are reported. In the last few years, also the necessity of describing the three-dimensional character of molecular structures has contributed to the development of three-dimensional indices [381]. The main application of topological descriptors is to quantitatively correlate structures and properties of biologically active compounds [382]. But it has to be taken into account that, whereas chemical structures are discrete entities, their properties show a continuous variation, expressed within a certain numerical range. Since then, many new TIs have been added for Quantitative Structure-Property Relationship (QSPR) and Quantitative Structure-Activity Relationship (QSAR) studies [296].

5.2 Characterization of structures: representation by means of graphs

Mathematically, a graph is the application of a set on itself, that is, a collection of elements of the set and of binary relations between these elements. Graphs are two-dimensional objects, but they can be embedded or realized in spaces of higher dimensions. In terms of its pictorical representation, a graph is a mathematical structure formed by a set of points and a set of lines that join some or all pairs of points. Points are also referred as vertices, nodes, and junctions, and lines as edges, axis, segments, arcs, and branches. A graph edge symbolizes a binary relation between the vertices that connects. A graph is a topological rather than a geometrical concept, and hence metric lengths, angles, and three-dimensional spatial configurations have no meaning.

Formally, a Graph $G\{V,E\}$ is formed by a non-empty and finite set of vertices $V=\{v_1,v_2,...,v_k\}$, and edges $E=\{e_1,e_2,...,e_q\}$, in such a way that each line joins two points.

The two-dimensional realization of a graph is a set of vertices (points) and of edges (lines) joining these vertices. A graph can be visualized by a diagram where the vertices are drawn as small circles or dots, and the edges as lines or curves connecting the appropriate circles. Mainly due to their diagrammatic representation, graphs are used as structural models in science, and, in particular, in chemistry.

In chemistry, graphs can be used to represent a variety of chemical objects such as molecules, reactions, crystals, polymers and clusters. The common feature of chemical systems is the presence of sites and connections between them. Sites can be atoms, electrons, molecules, molecular fragments, intermediates, etc., while the connections between sites can represent bonds, reaction steps, van der Waals forces, etc. Chemical systems can be represented by chemical graphs using a simple conversion rule: sites are replaced by vertices and connections by edges. Thus, chemical structures can be represented by a special class of structural graphs, the so-called chemical graphs, molecular graphs or constitutional graphs.

Molecular graphs are non-directed planar chemical graphs that represent the constitution of organic compounds. In the representation of chemical species by means of molecular graphs, individual atoms are represented by vertices, whereas covalent bonds or shared electron pairs are depicted by edges linking a pair of vertices.

The representation is the so-called symmetric tree, that is, a symmetric connected graph without circuits (acyclic) and non-directed connecting lines, i.e. there is a single path between each pair of vertices. A graph is connected when between each pair of vertices exists at least one path connecting them, where a path is an alternating sequence of vertices and edges, with each edge being incident to the adjacent vertices, and with no repeated vertices. Simple connected graphs express the connectedness of atoms in molecules, with a single edge between any pair of neighbouring, adjacent atoms. The chemical nature of atoms is neglected, thus dealing with a reference molecular skeleton. To simplify the manipulation of molecular graphs hydrogen-suppressed or hydrogen-depleted graphs are often used. Such graphs represent only the molecular skeleton, omitting hydrogen atoms and their bonds, and leaving only the non-hydrogen atoms, i.e., second or higher-row atoms, whose principal quantum number is $n \geq 2$. In this notation, adjacent vertices linked by a connecting axis stand for pairs of covalently linked atoms situated at a topological distance equal to one. Then, the different atoms or vertices are assigned an arbitrary number.

Virtually, all molecules, whether arbitrarily defined as rigid or non-rigid, can be represented by a chemical planar graph [342]. For this reason, it is meaningful to define **molecular structure** as an equivalent class of chemical graphs [343]. Such a definition associates an extended, fuzzy, vibrating and rotating molecular entity with an unchangeable, static mathematical structure of well-defined connectivity.

As previously commented, in graph representation, the geometrical features of organic compounds, such as bond lengths or bond angles, are not taken into account and the chemical bonding of atoms is regarded as the most important characteristic. However, this kind of representation has as a disadvantage the loss of information from the information reported by the real three-dimensional molecular geometry to the knowledge provided by the graphs, represented in a bidimensional plane. In the particular cases where the three dimensional geometry is crucial for their properties, geometrical parameters can be included in the description of the structure.

5.3 Associating graphs with matrices

Hence, in order to establish structure-function relationships, the chemical structure of a molecule visualized as a graph must be codified into a numerical form. In computing, graphs can be associated into the so-called **Topological Matrices (TM).** From the topological matrix elements derived from a molecular graph, topological indices are mathematically derived in a direct and unambiguous manner. Graph invariants can be used as sets of molecular descriptors to perform a comparative regression analysis and study how different properties of a set of molecules depend in the same structural factors.

The association of graphs with matrices is the link between the graphical description and the numerical description by means of invariants, which are calculated from connectivity matrices. These are squared symmetric matrices of order *n* being *n* the number of vertices, symmetrical in relation to the principal diagonal. The rows and columns labeling the TM elements correspond to the numeration of vertices that, in turn, correspond to the enumeration of atoms. It has to be noted that the numbering is arbitrary.

A labelled chemical connected graph may be associated with several matrices that account for connectivity, adjacency, and distance. Particularly, the most commonly used indices in graph theoretical representation can be coded by means of an attached **topological matrix,** and the **topological distance matrix,** which have been the source of generation of many **topological indices**.

5.3.1 Topological matrix

The **topological** or **adjacency** matrix (**T**), used by Cayley, codifies the vertices and the edges of a graph, and contains information on the connectivity between atoms, independently of the nature of bonds and the type of atoms. The elements of the **topological matrix** are composed by the unity if the associated vertices are directly connected, and by zero otherwise. Self-connections are not allowed in the adjacency matrix of a graph, so that it is a diagonal-zero matrix, (T_{ii} = 0). From the chemical point of view, the topological matrix can be regarded as a table of connections, where T_{ij} non-null entries indicate that atoms *i* and *j* are bonded.

5.3.2 Topological distance matrix

The **topological** or **graph distance matrix** (**D**), introduced in graph theory by Harary [286], accounts for the topological length, namely the topological distance, or the number of edges or bonds in the sequence defining the shortest path between two vertices or atoms. It is constructed taking the integer value of the number of bonds separating the considered vertices or zero otherwise, in the case of non-bonded atoms. The distance of an atom to itself is considered to be null (D_{ii}=0). The *ij* position value of the *k*-th potency of **T** gives the number of paths of length *k* from vertex v_i to vertex v_j. So, the distance matrix of a graph can be generated using powers of the corresponding adjacency matrix and the distance matrix [344].

5.3.3 Valence vector

Finally, the topological valence of a vertex is defined as the number of axis of incident to the vertex. The **valence vector** (**v**) is calculated as the sum of entries in *i*-th row or *j*-th column of the topological matrix, which indicates the coordination of an atom, that is, if an atom is primary, secondary, tertiary or quaternary. The excess or default of valence can be obtained by directly comparing the topological valence with the chemical valence of elements.

Table 1 shows the definition of the above-mentioned matrices, being *n* the number of atoms in the molecule and n_b the length of the shortest path between the vertices v_i and v_j.

Table 1. Definition and symbols of classical topological arrays used in chemical graph theory.

Definition	**Symbol**
$\mathbf{T}_{ij} = \begin{cases} 1 \text{ if atoms i and j are bonded} \\ 0 \text{ if atoms i and j are not bonded} \end{cases}$	**T**(nxn)
$\mathbf{D}_{ij} = \begin{cases} 0 \text{ if i=j} \\ \mathrm{n_b} \text{ if } \mathrm{i} \neq \mathrm{j} \end{cases}$	**D**(nxn)
$\mathbf{v}_i = \sum_{j=1}^{n} T_{ij}$	**v**(n)

5.4 From matrices to indices

Topological matrices can be used as a source to derive univocally calculated molecular descriptors that contain topological information embedded in the molecular structure. These structural invariants, called **Topological Indices (TI)**, can be collected in an *ad hoc* manner in the matrix form.

Thus, topological indices are scalar numerical descriptors mathematically derived in a direct and unambiguous manner from structural graphs. These univocal numerical quantities can be used for the structural characterization of molecular graphs of chemical structures, and its calculation must be independent of the arbitrary numeration chosen for the graph. In fact, the term graph theoretical index would be more accurate than topological index, although traditionally these invariants are referred to as topological indices. Since isomorphic graphs possess identical values for any given topological index, these indices are referred to as graph invariants. Two graphs are isomorphic if there is a one-to-one correspondence between the vertices so that preserves the adjacencies between axes. Isomorphic graphs can give the appearance of being graphically very different.

Alternatively, TI can be regarded as mathematical relationships related to the count of components of various types accounting for the properties of the structure they characterize. In this correspondence, each structure has a single descriptor associated, but not vice versa; one index may correspond to more than a graph. So it is desirable that the indices present low degeneracy. Given a list of invariants of structures, in a general case it is not possible to reconstruct the structure.

TQSI contain tridimensional information of the molecular structure (Euclidean distances between atoms) and, also, chemical information due to the fact that the overlap matrix elements have been obtained using a basis set depending on the nature of the atoms entering into the molecule. Thus, TQSI could be able to distinguish between rotamers, conformers contrarily to the classical TI.

Geometry-dependent and three-dimensional structural invariants for the characterization of molecular structures comprise the three-dimensional geometric information content present in chemical structures, and consequently, they can discriminate geometric isomers as well as conformational isomers. Graph-theoretically derived topological indices have found numerous applications to the prediction of physico-chemical properties and structure-activity relations.

There is an always-increasing proliferation of topological indices in the literature. For that reason, a list of **desirable properties** has been proposed by Randić [351-352]:

- Direct structural interpretation
- Good correlation with at least one molecular property
- Good discrimination of isomers
- Locally defined
- Generalizable
- Linearly independent
- Simplicity
- Not based on physical or chemical properties
- Not trivially related to tother indices
- Efficientcy of construction
- Based on familiar structural concepts
- Correct size dependence
- Gradual change with gradual change in structures

The first TIs, able to characterize the ramification of a graph, were introduced in the late 40s by Wiener [353-357], Platt [358-359], and Gordon et al. [360]. Various definitions of topological indices have been used in order to obtain molecular descriptors. The most relevant are the indices formulated by **Wiener** [353], **Hosoya** [361], **Randić** [362], **Kier and Hall** [337-338,363], **Balaban** [364], **Schultz** [366-367] and **Harary** [368]. Also, Zagrev indices (M1 and M2) [369], the Largest Eigenvalue [370] and Xu index [371] have been formulated. More recently, information theoretic J index, shape kappa indices, hyper Wiener, electrotopological state indices, and three-dimensional analogs of TI by Trinajstic and Todeschini have also been formulated.

5.4.1 Wiener Index

In 1947, Wiener introduced an index to relate the structure of hydrocarbons with phsycochemical properties such as boiling points [354], molar refraction, heat of formation [355], steam pressure as a function of temperature [356], and superficial tension [357]. The Wiener Path Number (*WPN*) [353] can be defined as the total number of bonds among all the pairs of atoms in a graph. The number of paths can be calculated from the topological distance matrix as the half-sum of the elements of this matrix:

$$WPN = \sum_{i=1}^{n} \sum_{j=i+1}^{n-1} D_{ij} \tag{73}$$

where n is the number of atoms, and D_{ij} are the elements of distance matrix, that is, the number of bonds in the shortest path between i and j.

Wiener also defined a number of polarity, p_3, obtained in the computation of the number of paths of length three. The Wiener Index (W) is defined as the summation of Wiener path number and the polarity number:

$$W = WPN + p_3 \tag{74}$$

Both Wiener path number and Wiener Index increase with the size of the molecule, but tend to diminish with molecular ramification.

5.4.2 Hosoya index

The Hosoya index (Z^A) was defined in 1971 [361] for non-directed graphs, as follows:

$$Z^A = \sum_{i=0}^{n/2} p^A(k) \tag{75}$$

where $p(k)$ is the number of ways in which such k bonds are chosen from the graph that no two of them are connected. By definition, $p^A(0)=1$ i $p^A(1)=n_b$.

The Hosoya index was firstly used to correlate with several of the thermodynamic quantitites of saturated hydrocarbons, such as the boiling point.

5.4.3 Randić Index of molecular ramification

The molecular ramification index was introduced by Randić in 1975, as the connectivity index [362]. Based in the classification of bonds in molecular graphs, is one of the most widely used topological indices in QSAR analysis. Randić classified the kind of bond between atoms, depending on the number of atoms bonded to each terminal vertex. The contribution to each type to the index is the inverse of the product of the square root of both valence vectors. The sum of all contributions for all *k* axis (within a total of *m* axis) constitute the Ramificaton Index of Randić (*R*), which classifies molecules attending to their ramifications:

$$R = \sum_{k=1}^{m} \frac{1}{\sqrt{\left(v_i v_j\right)_k}} \tag{76}$$

The expression of calculation of Randić index can also be modified to work directly with the adjacency matrix:

$$R = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{T_{ij}}{\sqrt{\left(v_i v_j\right)_k}} \tag{77}$$

For molecules with the same number of atoms, the index decreases when ramification increases. Randić index was satisfactorily correlated in alkanes with the boiling point, heat of formation and steam pressure. Besides, valences can be modified to include the effect of heteroatoms.

5.4.4 Generalized Connectivity Indices

Introduced and posteriorly developed by Kier and Hall [337-338,363], this kind of connecting graph can be divided into five types: the trivial, the path, the cluster, the path-cluster and the chain graph. Connectivity indices are calculated adding terms corresponding to all the connected subgraphs of the main graph, attending to a given order, where the subgraph order is the number of vertices that form it. A generalized connectivity index of order *m* and type *t*, ${}^{m}\chi_t$, is defined as:

$${}^{m}\chi_t = \sum_{i}^{n_t} \prod_{j=1}^{m+1} \frac{1}{\sqrt{\left(v_i\right)_s}} \tag{78}$$

where n_t is the number of connected subgraphs of type *t*, with *m* vertices.

5.4.5 Balaban index

Balaban index (*B*) was introduced in 1982 [364] as one of the less degenerated indices. It calculates the average distance sum connectivity index, according to the equation:

$$B = \frac{n_e}{\mu + 1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{\sqrt{(D)_i (D)_j}} \tag{79}$$

where n_e is the total number of axes (bonds) of the molecule, and $(D)_i$ represents the sum of topological distances from the vertex i to all the other vertices of the graph, extended for all the possible axes. μ is the number of cycles of the molecule, also called cyclomatic number, which expresses the deficiency of hydrogen in hydrocarbons, and it can be calculated using:

$$\mu = m - n + 1 \tag{80}$$

being m the number of axes and n the number of vertices. Balaban index measures the ramification and it tends to increase with molecular ramification. It has been satisfactorily correlated with octane numbers of alkanes [365].

5.4.6 Schultz Index

The Schultz index (MTI) was introduced by Schultz in 1989, as the molecular topological index [366-367]. It takes into account the effect of adjacency and distance matrices and the valence vector, and it is computed as:

$$MTI = \sum_{i=1}^{n} e_i \tag{81}$$

where e_i represent the elements of the row matrix of order n, calculated as follows:

$$e_i = [\mathbf{v}(\mathbf{T} + \mathbf{D})]_i \tag{82}$$

Hence, in this way,

$$MTI = \sum_{i=1}^{n} [\mathbf{v}(\mathbf{T} + \mathbf{D})] \tag{83}$$

5.4.7 Harary Number

The Harary number (H) was introduced in 1991 by Plavšić et al. [368] in honour of professor Frank Harary, due to his influence in the development of graph theory and, especially, to its application in chemistry [286]. This index is defined from the inverse of the squared elements of the distance matrix according to the expression:

$$H = \sum_{i=}^{n-1} \sum_{j=i+1}^{n} D_{ij}^{-2} \tag{84}$$

where D^{-2} is the squared inverse distance matrix.

5.4.8 Other indices

Other indices are Zagrev indices (M_1 and M_2), the largest eigenvalues index (x_1), and Xu index (Xu), defined as follows:

$$\begin{aligned} M_1 &= \sum_i \partial_i^T \\ M_2 &= \sum_i \sum_j \partial_i^T \partial_j^T \end{aligned} \tag{85}$$

$$x_1 = \max\left(EigenValues(T)\right) \tag{86}$$

$$Xu = n^{1/2} \log\left(\frac{\sum_i \left(\left(\sum_j T_{ij} \right) \left(\sum_j D_{\mathrm{int},ij} \right)^2 \right)}{\sum_i \left(\left(\sum_j T_{ij} \right) \left(\sum_j D_{\mathrm{int},ij} \right) \right)} \right) \tag{87}$$

6 FROM CLASSICAL TOPOLOGICAL APPROACH TO QUANTUM SIMILARITY THEORY

The graph-theoretical approach to QSPR is based on the use of topological indices for encoding the structural information [291,296,372-376]. The term topological index [361] indicates a characterization of a molecule (or a corresponding molecular graph [296]) by a single number. The need to represent molecular structure by a single number arises from the fact that most molecular properties are recorded as single numbers. The ordinary connectivity indices derived from a graph theory suffice for the prediction of many physical properties. However, geometry-dependent three-dimensional characterization of molecular structures is an important topic in several areas, such as computer enumeration, construction and representation of stereoisomers of molecules. So, for the prediction of properties of other molecules, i.e. molecules containing heteroatoms, purely graph theoretical indices may not suffice. The pharmaceutical and chemical community needs extended tools capable to extract complex information, derived from the three-dimensional (3D) molecular structure, provided that classical topological indices derived from topological matrices only provide part of the spatial information of a molecule.

Three-dimensional topological indices derived from molecular graphs have been defined as topographic indices, which include the 3D structural characterization of the molecules and can difference cis/trans, gauche/anti isomers [345-348]. Those geometry dependent structural invariants derived from geometry-based matrices are based on the comparison of shape and three-dimensional topology of molecules.

Some 3D variants of well-known TI can be also defined; such indices are the 3D Wiener path number (^{3D}W), 3D Shultz index (^{3D}MTI) and the 3D Harary number (^{3D}H). Their definition is the same as the related ones appearing in *Table 2*, but the **matrix of distances** entering into the index computation is the one collecting all the **euclidean distances** between pairs of atoms present in the molecule.

Also, a link between classical topological approach and the general quantum similarity theory can be sketched [21]. The same techniques for construction of TI have been generalised not only from classical topological matrices to geometry matrices [349-350], but they have also been extended to the use of QS matrices.

As a result, a new kind of three-dimensional molecular indices have been described: the so-called **TQSI** [21,199,247,377-378]. These novel descriptors, which also account for further three-dimensional information, have been computed using the program TOPO, completely developed at the Institute of Computational Chemistry [202]. TQSI can be defined using the classical construction according to the theoretical framework, but replacing the classical topological matrices by matrices derived from quantum similarity calculations [246].

6.1 Topological Quantum Similarity Matrices

In contrast to integer classical topological matrices, the three-dimensional structure of molecules can be included in classical graph theory by codifying molecules with real matrices. Geometry-based real matrices are sensitive to details of molecular architecture and, besides from connectivities between atoms, they also describe the kind of atoms and nature of bonds. For example, it may happen that two atoms are not directly bonded but they are close enough to present chemical interactions. Also, the effect of heteroatoms, as well as the 3D structure of the molecules, is considered. The inclusion of atomic distinction and 3D structure can be introduced using QS tools, which describe atoms by means of functions and take into account the optimized 3D structure of molecules. Hence, classical TM are replaced by QS matrices, which codify and describe the 3D molecular structure. Also, the topological distance matrix is replaced by the Euclidean distance matrix. Consequently, each matrix has its own associated real valence vector.

6.1.1 QS Matrix

The previously defined integer topological matrix can be substituted by similarity matrices resulting from an interatomic quantum similarity measure, and calculated between each pair of atoms of a given molecule:

$$Z_{ij}(\Omega) = \iint \rho_i(\mathbf{r}_1)\ \Omega(\mathbf{r}_1, \mathbf{r}_2)\ \rho_j(\mathbf{r}_2)\ d\mathbf{r}_1 d\mathbf{r}_2 \qquad (88)$$

where the non-differential positive definite operator, Ω, can be replaced by the Overlap or anyother similarity operator. This matrix accounts for the three-dimensional structural information and provides information on the strength of the interaction between atoms. In order to obtain comparable values to the classic topological matrix, the diagonal elements are set to 0.

$$Z_{ij} = \begin{cases} \iint \rho_i(\mathbf{r}_1)\rho_j(\mathbf{r}_2) d\mathbf{r}_1 d\mathbf{r}_2 & \text{if } i \neq j \\ 0 \ \ \text{if } i = j \end{cases} \qquad (89)$$

The entries of this matrix are defined as the interatomic overlap QS integrals between two Gaussian functions, for atoms *i* and *j*. The values of these normalized integrals are real numbers, comprised between 0 and 1, that substitute the integer values of the TM. The superposition of an element with itself gives the unity; however, the diagonal is composed by zeros, whereas the other elements are zero or close to zero in the case of far atoms.

Analogously to the overlap QS Matrix (**S)**, other QS definitions can be employed, i.e. the Cioslowski matrix [53] (**C')**, where $C'_{ij} = S^2_{ij}$, the electronic repulsion or Coulomb matrix (**C)**, the gravitational matrix (**G)**, etc.

6.1.2 Euclidean Distance Matrix

In addition to the use of QSM, the topological distance has been also replaced by the **Euclidean distance** between every pair of atoms [202]. The distinction between these two kinds of distances has been made from the index definition itself, clearly separating the 2D indices, calculated with the integer matrix, from the 3D ones, obtained with the real matrix, which characterises the 3D geometry of a molecule by including spatial information.

Whereas the topological DM only considers the shortest path equivalent to the number of bonds between two atoms, independently of the geometry of the molecule, the Euclidean DM regards the effect of non-bonded close atoms. This matrix of interatomic distances is simply derived from the X-ray crystallographic data or from the optimized coordinates of output file of geometry optimization programs.

The Euclidean distance matrix associated to a molecular structure is calculated as:

$$\begin{aligned} D_{ij} &= d_{ij}, \text{ if } i \neq j \\ D_{ij} &= 0, \text{ otherwise} \end{aligned} \tag{90}$$

where

$$d_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2 + \left(z_i - z_j\right)^2} \tag{91}$$

where $\left(x_i, y_i, z_i\right)$ and $\left(x_j, y_j, z_j\right)$are the coordinates of the nuclei i and j, respectively. In a connex graph, the elements of the distance matric, d_{ij}, belong to the Euclidean metric:

$$\begin{aligned} &d_{ij} \geq 0 \;\; \wedge \;\; d_{ij} = 0 \;\Leftrightarrow i = j \\ &d_{ij} = d_{ji} \\ &d_{ij} + d_{jk} \geq d_{ik} \end{aligned} \tag{92}$$

From the computational point of view, the diagonal elements are directly defined as null, and only the upper or lower triangle elements are computed, due to the symmetry of these matrices.

The Euclidean matrix is a generalization of the topological distance matrix in the sense that the classical topological matrix only contains the distances between the chemically bonded atoms. In contrast, the geometry-based matrix contains the distances between all the nuclei in the molecular structure, and depends on bond angles and dihedral angles. Thus, the distance matrix associated to the molecular graphs gives information about of chemical bonding (adjacency) but not on the geometry. Geometrical distance matrix reflects interactions through space, while topological and distance matrix reflects interactions through bonding.

6.1.3 Valence Vector

Similarly, the valence vector also is computed from the entries of the QS matrix. In some previous studies, the concept of similarity, embedded with topological indices, was successfully applied to the calculation of physicochemical properties [378] and QSAR studies [247].

6.2 Topological Quantum Similarity Indices (TQSI)

As in the classical topological approach, appropriate manipulations of the elements embedded in the TQSMs permit to generate TQSI [24]. Besides, the structural invariants derived from the classical topological matrix lead to degeneracies since many entries in the adjacency matrices are zeros. As a result, isomers which differ geometrically cannot be differentiated by the indices derived from the adjacency matrix.

In contrast to geometry-dependent structural invariants, graph-theoretically derived indices depend only on the topological matrix and thus they do not correlate directly to with the three-dimensional structure or the compactness of the molecule in question in the three-dimensional space. Graph-theoretically derived indices are, however, easier to compute as they depend only on the connectivity.

TQSI can be calculated from a modified formulation analogous to the classical formulation [247,380]. The indices used in this work are: Wiener (*W*) and Wiener Path Number (*WPN*), Randić (*R*), Schultz (*MTI*), Balaban (*B*) and Hosoya (*Z*) indices, Harary Number (*H*), and the generalised connectivity indices ($^{m}\chi_{t}$) of Kier and Hall.

In order to obtain the respective TQSI, it is only necessary to invoke the same mathematical generating rules for classical indices but replacing the numbers coming from the TM by the ones arising from TQSM. The valence vectors are replaced by the ones generated from the respective TQSM, and the topological distance matrices are substituted in the TQSM case by the three dimensional Euclidean distances [24,247].

Table 2 shows the definition of several TQSI, where the summations run over the same integer indices (thus, the discretised molecular bond structure is also considered) but new indices include, in some way, information about the molecular 3D structure. *Z* represents any choice of the operator Ω providing different forms of TQSI.

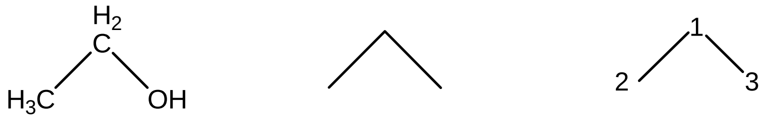
Table 2. Definition of TQSI.

[†]3D analogous of classical TI; they are not properly TQSI, provided that the definition does not include the interatomic similarity matrix, Z.

Index	Definition
Wiener Path Number[†]	$WPN = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} D_{ij}$
Wiener index[†]	$W = WPN + p_3$
Hosoya index[†]	$Z^A = \sum_{i=0}^{n/2} p^A(k)$
Randić index	$R^Z = \sum_{bonds} \frac{Z_{ij}}{\sqrt{\left(v_i^Z v_j^Z\right)_k}}$
Generalised connectivity indices of order *m* and type *t*	${}^m\chi_t^Z = \sum_{s=1}^{n_t}\prod_{i=1}^{m+1}\left(v_i^Z\right)^{-\frac{1}{2}}$
Balaban index[†]	$B = \frac{n_e}{\mu+1}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\left(D_i\right)\left(D_j\right)\right]^{-\frac{1}{2}}$
Schultz index	$MTI = \sum_{i=1}^{n}\left[v^{\mathbf{Z}}\left(\mathbf{Z}+\mathbf{D}\right)\right]_i$
Harary number[†]	$H = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} D_{ij}^{-2}$
Zagrev indices	$M_1 = \sum_i \partial_i^Z$ $M_2 = \sum_i\sum_j \partial_i^Z \partial_j^Z$
Largest eigenvalue	$x_1 = \max(EigenValues(Z))$
Xu index	$Xu = n^{1/2}\log\left(\frac{\sum_i\left(\left(\sum_j Z_{ij}\right)\left(\sum_j D_{ij}\right)^2\right)}{\sum_i\left(\left(\sum_j Z_{ij}\right)\left(\sum_j D_{ij}\right)\right)}\right)$

Where p_3 is the number of atoms separated by three bonds in the molecule, the symbol $[B]_i$ stands for the *i*-th row of matrix *B*, μ is the number of cycles, n_e is the number of edges of the related graph, $(D)_i$ stands for the sum of distances from vertex *i* and n_t is the number of connected subgraphs of type *t*. Within the classical approach, $p^T(i)$ is the number of ways to draw *i* non-adjacent bonds in the molecular graph. As a particular case, it is defined $p^T(0)=1$.

6.3 Molecular example



Structural formula | Hidrogen Suppressed graph | Numbering of atoms

Figure 8. Molecular characterizations for etanol molecule.

$$T=\begin{pmatrix}0&1&1\\1&0&0\\1&0&0\end{pmatrix}\qquad V=\begin{pmatrix}2\\1\\1\end{pmatrix}\qquad D=\begin{pmatrix}0&1&1\\1&0&2\\1&2&0\end{pmatrix}$$

Topological Matrix | Valence Vector | Topological Distance Matrix

Figure 9. Classical Topological Matrices for ethanol molecule

In the hydrogen-suppressed form, only *n*=3 atoms are relevant. In *Table 3* the employed cartesian coordinates and the full topological valences attached to every atom have been indicated.

Table 3. Optimised cartesian coordinates for the molecular skeleton C2-C1-O. The topological valence is referred to the original graph considering the hydrogen atoms.

Atom	Topological	Cartesian coordinates / a.u.		
	valence	x	y	z
C_2	4	-0.85096	2.7269	0.0000
C_1	4	0.0000	0.0000	0.0000
O	2	2.6827	0.0000	0.0000

TQSM and the corresponding valence vectors are shown together with the euclidian distance matrix.

Table 4. Lower triangles of the TQSM and euclidian distance matrix attached to the ethanol molecule.

Overlap	Cioslowski	Distance (a.u.)
$\boldsymbol{S}=\begin{pmatrix}0 & & \\ 0.54503 & 0 & \\ 0.57662 & 0.08389 & 0\end{pmatrix}$	$\boldsymbol{C}=\begin{pmatrix}0 & & \\ 0.29706 & 0 & \\ 0.33249 & 0.00704 & 0\end{pmatrix}$	$\boldsymbol{D}=\begin{pmatrix}0 & & \\ 2.8566 & 0 & \\ 2.6827 & 4.4635 & 0\end{pmatrix}$
$v^S=\begin{pmatrix}1.1217\\ 0.62892\\ 0.66051\end{pmatrix}$	$v^C=\begin{pmatrix}0.62955\\ 0.3041\\ 0.33953\end{pmatrix}$	

The Randić index coming from each of the matrices *T*, *S* and *C* can be calculated as follows:

$$\chi^T=\frac{T_{12}}{\left(v_1^T v_2^T\right)^{\frac{1}{2}}}+\frac{T_{13}}{\left(v_1^T v_3^T\right)^{\frac{1}{2}}}=\frac{1}{(2\times 1)^{\frac{1}{2}}}+\frac{1}{(2\times 1)^{\frac{1}{2}}}=\sqrt{2}$$

$$\chi^S=\frac{S_{12}}{\left(v_1^S v_2^S\right)^{\frac{1}{2}}}+\frac{S_{13}}{\left(v_1^S v_3^S\right)^{\frac{1}{2}}}=\frac{0.54503}{(1.1217\times 0.62892)^{\frac{1}{2}}}+\frac{0.57662}{(1.1217\times 0.66051)^{\frac{1}{2}}}=1.3188$$

$$\chi^C=\frac{C_{12}}{\left(v_1^C v_2^C\right)^{\frac{1}{2}}}+\frac{C_{13}}{\left(v_1^C v_3^C\right)^{\frac{1}{2}}}=\frac{0.29706}{(0.62955\times 0.3041)^{\frac{1}{2}}}+\frac{0.33249}{(0.62955\times 0.33953)^{\frac{1}{2}}}=1.3981$$

Within the classical topological formulation, the Randić index is equivalent to the connectivity path one [337]: $\chi^T={}^1\chi_P^T$. In the TQS-based approach, it can be realised that these indices may have different values. This is due to the contribution of the T_{ij} terms into the first one, while in the later only topological valences are present. For the ethanol molecule, the connectivity path ${}^1\chi_P^A$ indices are

$${}^1\chi_P^T=\chi^T=\sqrt{2}$$

$${}^1\chi_P^S=\frac{1}{\left(v_1^S v_2^S\right)^{\frac{1}{2}}}+\frac{1}{\left(v_1^S v_3^S\right)^{\frac{1}{2}}}=\frac{1}{(1.1217\times 0.62892)^{\frac{1}{2}}}+\frac{1}{(1.1217\times 0.66051)^{\frac{1}{2}}}=2.3524$$

$${}^1\chi_P^C=\frac{1}{\left(v_1^C v_2^C\right)^{\frac{1}{2}}}+\frac{1}{\left(v_1^C v_3^C\right)^{\frac{1}{2}}}=\frac{1}{(0.62955\times 0.3041)^{\frac{1}{2}}}+\frac{1}{(0.62955\times 0.33953)^{\frac{1}{2}}}=4.4485$$

and from here it can be seen how different TQSM can really lead to different index values.

Concerning the Hosoya index, the $p^T(0)$ and $p^Z(1)$ contributions can be computed as:

$$Z^T=1+2=3$$

$$Z^S=-(\ln S_{12}+\ln S_{13})=-\ln(S_{12}S_{13})=1.1575$$

$$Z^C=-(\ln C_{12}+\ln C_{13})=-\left(\ln S_{12}^2+\ln S_{12}^2\right)=-2(\ln S_{12}+\ln C_{13})=2Z^S$$

REFERENCES

1. Butlerov, A.M. *Z. Chem., 4,* **1861**, 549. (Tr. Kluge, F.F.; Larder, D.F.*J. Chem. Educ., 48*, **1971**, 289).
2. Couper, A.S. *Ann. Chim. Phys., 53*, **1858**, 469.
3. Hofmann, A.W. *Proc. R. Inst. G.B. London, 4,* **1865**, 414.
4. Stark, J. *Prinzipien der Atomdynamik, Part III. Die Elek- trizitat im Chemischen Atom*. Hirzel: Leipzig, **1915**, 81.
5. Crum Brown, A. *The Theory of Chemical Combination*. M.D. Thesis, University of Edinburgh, **1861**.
6. Crum Brown, A. *Trans. R. Sot. Edinburgh, 23,* **1864**, 707.
7. Crum-Brown, A. and Fraser, T.R. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from strychia, brucia, thebaia, codeia, morphia and nicotia. *Trans. Royal Soc. Edinburgh, 25*, **1968-9**, 257-274.
8. SMILES: Simplified Molecular Input Line Entry Specification. Daylight Chemical Information Systems, Inc. 18500 Von KArman Ave 450, Irvine. CA 92715, USA.
9. Sen, K. (Ed.) *Topics in Currrent Chemistry; Vol. 173.* Springer-Verlag: Berlin, **1995**.
10. Johnson, M.A.; Maggiora, G. (Eds.) *Concepts and Applications of Molecular Similarity*. John Wiley & Sons Inc.: New York, **1990**.
11. Fuson, K.C. Analysis of research needs in projective, affine and similarity geometries, including an evaluation of Piaget's results in this area. In *Recent Research concerning the Development of Spatial and Geometric Concepts.* R. Lesh (Ed.) ERIC/SMEAC: Columbus, Ohio, **1978**, 243-260.
12. Mendeleev, D.I. *Principles of Chemistry*; *Vol. 2,* **1868–71,** tr. **1905**.
13. Rouvray, D.H. Similarity in chemistry: past, present and future. In *Topics in Current Chemistry; Vol 173*. Sen, K. (Ed.) Springer-Verlag: Berlin, **1995**, 1-30.
14. Klein, D.J. Similarity and Dissimilarity in posets. *J. Math. Chem., 18,* **1995,** 321-348.
15. Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem., 17,* **1980**, 1185-1189.
16. Carbó, R.; Calabuig, B. Quantum Similarity: Definitions, Computational Details and Applications. In *Computational Chemistry: Structure, Interactions and Reactivity*. Fraga, S. (Ed.) Elsevier: Amsterdam, **1992**.
17. Carbó, R.; Calabuig, B.; Vera, L.; Besalú, E. Molecular quantum similarity: Theoretical framework, ordering principles, and visualization techniques. *Adv. Quantum Chem., 25*, **1994**, 253-313.
18. Besalú, E.; Carbó, R.; Mestres, J.; Solà, M. Foundations and Recent Developments on Molecular Quantum Similarity. *Topics Curr. Chem., 173,* **1995**, 31-62. *Molecular similarity and reactivity: From quantum chemical to phenomenological approaches.*
19. Carbó, R.; Besalú, E. Theoretical Foundations of Quantum Molecular Similarity. In *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Approaches*. Carbó, R. (Ed.) Kluwer: Amsterdam, **1995**, 3-30.
20. Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. On quantum molecular similarity measures (QMSM) and indices (QMSI). *J. Math. Chem., 19,* **1996**, 47-56.
21. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. Quantum Similarity. In *Advances in Molecular Similarity; Vol. 2*. Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: Greenwich, **1998**, 1-42.
22. Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struc. (Theochem), 451,* **1998**, 11-23.
23. Robert, D.; Carbó-Dorca, R. General trends in atomic and nuclear quantum similarity measures. *Int. J. Quantum Chem., 77,* **2000**, 685-692.
24. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum Molecular Similarity: Theory and Applications to the Evaluation of Molecular Properties, Biological Activities and Toxicity. In, *Fundamentals of molecular similarity*. Carbó-Dorca, R.; Gironés, X.; Mezey, P.G. (Eds.) Kluwer Academic/Plenum Press: New York, **2001**, 187-320.
25. Carbó-Dorca, R. On the statistical interpretation of density functions: Atomic shell approximation, convex sets, discrete quantum chemical molecular representations, diagonal vector spaces and related problems. *J. Math. Chem., 23,* **1998**, 365-375.
26. Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets. *J. Math. Chem., 22,* **1997**, 143-147.

27. Carbó-Dorca, R. Fuzzy sets and Boolean tagged sets, Vector Semiespaces and Convex Sets, QSM and ASA density functions, Diagonal Vector Spaces and Quantum Chemistry. In *Advances in Molecular Similarity; Vol. 2.* JAI Press, **1998**, 43-72.
28. Carbó-Dorca, R. Tagged sets, convex sets and quantum similarity measures. *J. Math. Chem.*, *23*, **1998**, 353-364.
29. Mestres, J.; Solà, M.; Duran, M.; Carbó, R. On the Calculation of Ab-Initio Quantum Molecular Similarities for Large Systems - Fitting the Electron-Density. *J. Comput. Chem.*, *15*, **1994**, 1113-1120.
30. Constans, P.; Amat, L.; Fradera, X.; Carbó-Dorca, R. Quantum Molecular Similarity Measures (QMSM) and the Atomic Shell Approximation (ASA). In *Advances in Molecular Similarity*; *Vol. 1.* Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: London, **1996**, 187-211.
31. Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: first order density fitting using Elementary Jacobi Rotations. *J. Comput. Chem., 18*, **1997**, 2023-2039.
32. Gironés, X.; Amat, L.; Carbó-Dorca, R. A comparative study of isodensity surfaces using ab initio and ASA density functions. *J. Mol. Graph. Model.*, *16*, **1998**, 190-196.
33. Amat, L.; Carbó-Dorca, R. Fitted electronic density functions from H to Rn for use in quantum similarity measures: cis-diammine-dichloroplatinum (II) complex as an application example. *J. Comput. Chem.*, *20*, **1999**, 911-920.
34. Amat, L.; Carbó-Dorca, R. Molecular electronic density fitting using elementary Jacobi rotations under atomic shell approximation. *J. Chem. Inf. Comp. Sci.*, *40*, **2000**, 1188-1198.
35. Constans, P.; Amat, L.; Carbó-Dorca, R. Toward a global maximization of the molecular similarity function: superposition of two molecules. *J. Comput. Chem.*, *18*, **1997**, 826-846.
36. Girones, X.; Robert, D.; Carbó-Dorca, R. TGSA: A molecular superposition program based on topo-geometrical considerations. *J. Comput. Chem.*, 22, **2001**, 255-263.
37. Amat, L.; Besalú, E.; Carbó, R.; Fradera, X. Practical applications of quantum molecular similarity measures (QMSM): programs and examples. *Sci. Gerun.*, *21*, **1995,** 127-143.
38. Bowen-Jenkins, P.E.; Cooper, D.L.; Richards, W.G. Ab initio computation of molecular similarity. *J. Phys. Chem.*, *89*, **1985**, 2195-2197.
39. Bowen-Jenkins, P.E.; Richards, W.G. Molecular similarity in terms of valence electron density. *J. Chem. Soc. Chem. Comm.*, *133*, **1986**, 133-135.
40. Bowen-Jenkins, P.E.; Richards, W.G. Quantitative measures of similarity between pharmacological active compounds. *Int. J. Quantum Chem.*, *30*, **1986**, 763-768.
41. Hodgkin, E.E.; and Richards, W.G. Molecular similarity based on electrostatic potential and electric field. *Int. J. Quant. Chem.*, *14*, **1987**, 105-110.
42. Richards, W.G.; Hodkgin, E.E. Molecular similarity. *Chem. in Britain*, *24*, **1988**, 1141- 1143.
43. Meyer, A.M.; Richards, W.G. Similarity molecular shape. *J. Comput.-Aided Mol. Des.*, *5*, **1991**, 426-439.
44. Good, A.C. The calculation of molecular similarity: alternative formulas, data manipulation and graphical display. *J. Mol. Graph.*, *10*, **1992**, 144-151.
45. Good, A.C.; Hodgkin, E.E.; Richards, W.G. Similarity screening of molecular data sets. *J. Comput.-Aided Mol. Des.*, *6*, **1992**, 513-520.
46. Seri-Levy, A.; West, S.; Richards, W.G. Molecular similarity, quantitative quirality and QSAR for quiral drugs. *J. Med. Chem.*, *37*, **1994**, 1727-1732.
47. Seri-Levy, A.; West, S.; Richards, W.G. Shape similarity as a single independent variable in QSAR. *Eur. J. Med. Chem.*, *29*, **1994**, 687-694.
48. Richards, W.G. Molecular similarity and dissimilarity. In *Modelling of biomolecular structures and mechanisms*. Pullman, A. (Ed.) Kluwer: The Netherlands, **1995**.
49. Richards, W.G. The dominant role of shape similarity and dissimilarity in QSAR. In QSAR and molecular modelling: concepts, computational tools and biological applications. Sanz, F.; Manaut, F. (Eds.) Prous Science: Barcelona, **1995**, 364-373.
50. Luque, F.J.; Sanz, F.; Illas, F.; Pouplana, R.; Smeyers, Y. G. Relationships between the activity of some H2-receptor agonists of histamine and their *ab initio* molecular electrostatic potential (MEP) and electron density comparision coefficients. *Eur. J. Med. Chem.*, *23*, **1998,** 7-10.
51. Burt, C.; Richards, W.G.; Huxley, P. The application of molecular similarity calculations. *J. Comput. Chem.*, *11*, **1990**, 1139-1146.
52. Richard, A. M. Quantitative comparison of molecular electrostatic potentials for structure activity studies. *J. Comput. Chem.*, *12*, **1991**, 959-969.
53. Cioslowski, J.; Fleishmann, E.D. Assessing molecular similarity from results of *ab initio* electronic structure calculations. *J. Am. Chem. Soc.*, *113*, **1991**, 64-67.

54. Cioslowski, J.; Challacombe, M. Maximum similarity orbitals for analysis of the electronic exited states. *Int. J. Quant. Chem., 25,* **1991**, 81-93.
55. Ortiz, J.V.; Cioslowski, J. Molecular similarity indices in electron propagator theory. *Chem. Phys. Lett., 185,* **1991**, 270-275.
56. Cioslowski, J. Differential density matrix overlap: an index for assessment of electron correlation in atoms and molecules. *Theor. Chim. Acta, 81*, **1992**, 319-327.
57. Cooper, D.L.; and Allan, N.L. Bond formation in momentum space. *J. Chem. Soc. Faraday Trans., 83*, **1987**, 449-460.
58. Cooper, D.L.; Allan, N.L. A novel approach to molecular similarity. *J. Comput.-Aided Mol. Des., 3*, **1989**, 253-259.
59. Cooper, D.L.; Allan, N.L. Molecular dissimilarity: a momentum-space criterion. *J. Am. Chem. Soc., 114,* **1992**, 4773-4776.
60. Allan, N.L., and Cooper, D.L. A momentum space approach to molecular similarity. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 587-590.
61. Allan, N.L.; Cooper, D.L. Momentum-space electron densities and quantum molecular similarity. *Top. Curr. Chem., 173,* **1995**, 85-111.
62. Cooper, D.L.; Allan, N.L. Molecular similarity and momentum space. In *Molecular similarity and reactivity: from quantum chemical to phenomenological approaches.* Carbó, R. (Ed.) Kluwer: Amsterdam, **1995**, 31-55.
63. Measures, P.T.; Allan, N.L.; Cooper, D.L. Momentum-space similarity: some recent applications. In *Advances in molecular similarity*; *Vol 1.* Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: Greenwich, **1996**, 61-87.
64. Allan, N.L.; Cooper, D.L. Quantum molecular similarity via momentum-space indices. *J. Math. Chem., 23*, **1998**, 51-60.
65. Fratev, F.; Polansky, O.E.; Mehlhorn, A.; Monev, V. Application of distance and similarity measures. The comparison of molecular electronic structures in arbitrari electronic states. *J. Mol. Struct., 56,* **1979**, 245-253.
66. Fratev, F.; Monev, V.; Mehlhorn, A.; Polansky, O.E. Application of similarity measures. An estimation of the degree of fragmentation of a molecule in ground and exited states. *J. Mol. Struct., 56,* **1979**, 255-266.
67. Amovilli, C.; McWeeny, R. Shape and similarity: two aspects of molecular recognition. *J. Mol. Struct. (Teochem), 227,* **1991**, 1-9.
68. Pepperrell, C.A.; Willett, P. Techniques for the calculation of three-dimensional structural similarity using interatomic distances. *J. Comput.-Aided Mol. Des., 5,* **1991**, 455-474.
69. Petke, J.D. Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem., 14,* 1993, 928-933.
70. Lee, C.; Smithline, S. An approach to molecular similarity using density functional theory. *J. Phys. Chem., 98,* **1994**, 1135-1138.
71. Turner, D.B.; Willett, P.; Ferguson, A.M.; Heritage, T.W. Similarity searching in files of threedimensional structures: evaluation of similarity coefficients and standarization methods for field-based similarity searching. *SAR QSAR Environ. Res., 3,* **1995**, 101-130.
72. Burgess, E.M.; Ruell, J.A.; Zalkow, L.H.; Haugwitz, R.D. Molecular similarity from atomic electrostatic multipole comparisons. Application to anti-HIV drugs. *J. Med. Chem., 38,* **1995**, 1635-1640.
73. Sheridan, R.P.; Miller, M.D.; Underwood, D.J.; Kearsley, S.K. Chemical similarity using geometric atom pair descriptions. *J. Chem. Inf. Comput. Sci., 36,* **1996**, 128-136.
74. Boon, G.; De Proft, F.; Langenäker, W.; Geerlings, P. The use of density functional theorybased reactivity descriptors in molecular similarity calculations. *Chem. Phys. Lett., 295,* **1998**, 122-128.
75. Herndon, W.C.; Leonard, J.E. *Inorg. Chem., 22,* **1983**, 554.
76. Herndon, W.C.; Bertz, S.H. Linear notations and molecular graph similarity. *J. Comput. Chem., 8,* **1987**, 367-374.
77. Herndon, W.C. Graph codes and a definition of structural similarity. *Comput. Math. Applic., 15,* **1988**, 303-309.
78. Herndon, W.C.; Bruce, A.J. *J. Math. Chem., 2,* **1988**, 155.
79. A similarity measure for graphs: reflections on a theme of Ulam. *Los Alamos Sci., 15,* **1987,** 114-121.
80. Herndon, W. C.; Rum, G. Three-Dimensional Topological Descriptors and Similarity of Molecular Structures: Binding Affinities of Corticosteroids. In *QSAR and Molecular Modeling: Concepts, Computational Tools and Biological Aplications. Proceedings of the 10th European*

Symposium on Structure-Activity Relationships, QSAR and Molecular Modelling. Sanz, F.; Giraldo, J.; Manaut, F. (Eds.) Prous Science Publishers: Barcelona, **1995**, 380-384.

81. Mezey, P.G. Shape group studies of molecular similarity: shape groups and shape graphs of molecular contour surfaces. *J. Math. Chem., 2,* **1988**, 299-323.
82. Arteca, E.A.; Jammal, V.B.; Mezey, P.G. Shape group studies of molecular similarity and regioselectivity in chemical reactions. *J. Comput. Chem., 9,* **1988**, 608-619.
83. Arteca, G.A.; Mezey, P.G. Molecular similarity and molecular shape changes along reactions paths: a topological analysis and concequences on the Hammond postulate. *J. Phys. Chem., 93,* **1989**, 4746-4751.
84. Arteca, G.A.; Mezey, P.G. A quantitative aroach to structural similarity from molecular topology of reaction paths. *Int. J. Quant. Chem. Symp., 24,* **1990**, 1-13.
85. Arteca, G.A.; Mezey, P.G. Similarity between the effects of configurational changes and alied electric fields on the shape of electron densities. *J. Mol. Struct. (Teochem), 256,* **1991**, 125-134.
86. Harary, F.; Mezey, P.G. Similarity and complexity of the shapes of square-cell configurations. *Theor. Chim. Acta, 79,* 1991, 379-387.
87. Mezey, P.G. The degree of similarity of three-dimensional bodies: alications to molecular shapes. In *Mathematical modeling in chemistry*. Mezey, P.G. (Ed.) VCH: New York, **1991**, 39-49.
88. Luo, X.; Mezey, P.G. A global characterization and similarity analysis of two-dimensional potential energy surfaces. *Int. J. Quant. Chem., 41,* **1992**, 557-579.
89. Mezey, P.G. Shape-similarity measures of molecular bodies: a 3D topological aroach to QShAR. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 650-656.
90. Mezey, P.G. Similarity analysis in two and three dimensions using lattice animals and polycubes. *J. Math. Chem*., *11*, **1992**, 27-45.
91. Mezey, P.G. Shape in chemistry: an introduction to molecular shape and topology. VCH: New York, **1993**.
92. Ponec, R. Topological Aspects of Chemical-Reactivity - on the Similarity of Molecular-Structures. *Collect. Czech. Chem. Commun., 52,* **1987**, 555-562.
93. Ponec, R. Similarity Measures, the Least Motion Principle and Selection-Rules in Chemical-Reactivity. *Z. Phys. Chem-Leipzig, 268,* **1987**, 1180-1188.
94. Ponec, R.; Strnad, M. Similarity Approach to Chemical-Reactivity - Specificity of Multibond Processes. *Collect. Czech. Chem. Commun., 55,* **1990**, 2583-2589.
95. Ponec, R.; Strnad, M. Topological Aspects of Chemical-Reactivity - Evans-Dewar Principle in Terms of Molecular Similarity Approach. *J. Phys. Org. Chem., 4,* **1991**, 701-705.
96. Ponec, R.; Strnad, M. Electron Correlation in Pericyclic Reactivity - a Similarity Approach. *Int. J. Quantum Chem., 42,* **1992**, 501-508.
97. Ponec, R.; Strnad, M. Similarity Ideas in the Theory of Pericyclic Reactivity. *J. Chem. Inf. Comp. Sci., 32,* **1992**, 693-699.
98. Ponec, R.; Strnad, M. Position Invariant Index for Assessment of Molecular Similarity. *Croat. Chem. Acta, 66,* **1993**, 123-127.
99. Ponec, R. Similarity Approach to Chemical-Reactivity - a Simple Criterion for Discriminating between One-Step and Stepwise Reaction-Mechanisms in Pericyclic Reactivity. *J. Chem. Inf. Comp. Sci., 33,* **1993**, 805-811.
100. Willett, P. *Similarity and Clustering in Chemical Information Systems.* Wiley: New York, **1987**.
101. Dean, P.M. (Ed.) *Molecular Similarity in Drug Design*. Blackie Academic: London, **1995**.
102. Carbó, R. (Ed.) *Molecular Similarity and Reactivity: From Quantum Chemical to Phenomenological Aproaches*. Kluwer: Amsterdam, **1995**.
103. Carbó-Dorca, R.; Mezey, P.G. (Eds.) *Advances in Molecular Similarity*; *Vol. 1.* JAI Press: Greenwich, **1996**.
104. Carbó-Dorca, R.; Mezey, P.G. (Eds.) *Advances in Molecular Similarity*; *Vol. 2.* JAI Press: Greenwich, **1998**.
105. Kubinyi, H.; Martin, Y.C.; Folkers, G. (Eds.) *3D QSAR in Drug Design: Ligand-Protein Interactions and Molecular Similarity (Quantitative Structure-Activity Relationships); Vol.2.* Kluwer: Dordrecht, **1998**.
106. Carbó-Dorca, R.; Gironés, X.; Mezey, P.G. (Eds.) *Fundamentals of molecular similarity*. Kluwer Academic/Plenum Press: New York, **2001.**
107. Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. Quantum Mechanical Origin of QSAR: Theory and Alications. *J. Mol. Struct. (Theochem), 504,* **2000**, 181-228.
108. Webpage of the IV Girona Seminar on Molecular Similarity: http: //iqc.udg.es/gsms99/ [last accessed 28 May 2004].

109. Webpage of the V Girona Seminar on Molecular Similarity: http: //iqc.udg.es/gsms2001/ [last accessed 28 May 2004].
110. Webpage of the IV Girona Seminar on Molecular Similarity: http: //stark.udg.es/gsms2003/ [last accessed 28 May 2004].
111. Messiah, A. *Mécanique Quantique*. Dunod: Paris, **1959**.
112. McQuarrie, D. A. *Quantum Chemistry*. University Science Books: Mill Valley, California, **1983**.
113. Bohm, D. *Quantum theory*. Dover publications: New York, **1989**.
114. Born, M.; Oppenheimer, J. R. *Ann. Phys. Liepzig, 84*, **1927**, 457.
115. Sutcliffe, B.T. Fundamentals of computational quantum chemistry. In *Computational Techniques in Quantum Chemistry*. Diercksen, G.H.F.; Sutcliffe, B.T.; Veillard, A. (Eds.) Reidel: Boston, **1975**, 1.
116. Born, M. *Atomic Physics*. Blackie & Son Limited: London, **1945**.
117. Von Neumann, J. *Mathematische Grundlagen der Quantenmechanik.* Springer-verlag, **1932.** *(*Tr. *Mathematical Foundations of Quantum Mechanics.* Princeton University Press: Princeton, N. J., **1955**).
118. Dirac, P.A.M. *The Principles of Quantum Mechanics*. Clarendon Press: Oxford, **1983**.
119. Schrödinger, E. *Mémoires sur la Mécanique Ondulatoire.* Éditions Jacques Gabay : Paris, **1933.**
120. Löwdin, P. 0. Quantum Theory of Many-Particle Systems. I. Physical Interpretation by Means of Density Matrices, Natural Spin-Orbitals; Convergence Problems in the Method of Configurational Interaction. *Phys. Rev., 97*, **1955**, 1474-1489.
121. McWeeny, R. The Density Matrix in Many-Electron Quantum Mechanics. I. Generalized Product Functions. Factorization and Physical Interpretation of the Density Matrices. *Proc. R. Soc. Lond Ser A, 253*, **1959**, 242-259.
122. Davidson, E. R. Reduced Density Matrices in Quantum Chemistry. Academic Press: New York, **1976**.
123. Roothaan, C.C.J. New developments in molecular orbital theory. *Rev. Mod. Phys., 23*, **1951**, 69-89.
124. Szabo, A.; Otslund, N.S. *Modern Quantum Chemistry*. McGraw-Hill, Inc.: New York, **1991**.
125. Mezey, P.G. The Holographic Electron Density Theorem and Quantum Similarity Measures, *Mol. Phys., 96*, 1999, 169-178.
126. Mezey, P.G. Functional groups in quantum chemistry. *Adv. Quant. Chem., 27,* **1996**, 163-222.
127. Baerends, E. J. ; Ellis, D. E. ; Ros, P. Self-consistent molecular Hartree-Fock-Slater calculations. I. The computational procedure. *Chem. Phys., 2*, **1973**, 41-51.
128. Sambe, H.; Felton, R. H. A new computational approach to Slater's SCF-Xa equation. *J. Chem. Phys., 62*, **1975**, 1122-1126.
129. Delley, B.; Ellis, D. E. Efficient and accurate expansion methods for molecules in local density models. *J. Chem. Phys., 76,* **1982**, 1949-1960.
130. Dunlap, B.I.; Connolly, J.W.D.; Sabin, J.R. On some approximations in applications of Xa theory. *J. Chem. Phys., 71,* **1979**, 3396-3402.
131. Andzelm, J.; Wimmer, E. Density functional Gaussian-type-orbital approach to molecular geometries, vibrations; reaction energies. *J. Chem. Phys, 96,* **1992**, 1280-1303.
132. Gallant, R.T.; St-Amant, A. Linear scaling for the charge density fitting procedure of the linear combination of Gaussian-type orbitals density functional method. *Chem. Phys. Letters, 256,* **1996**, 569-574.
133. Goh, S.K.; St.-Amant, A. Using a fitted electronic density to improve the efficiency of a linear combination of Gaussian-type orbitals calculation. *Chem. Phys. Letters, 264*, **1997**, 9-16.
134. Unsöld, A. Beitrage zur Quantenmechanik de Atome. *Ann. Physik., 82,* **1927**, 355-393.
135. Cioslowski, J.; Piskorz, P.; Rez, P. Accurate analytical representations of the core electron densities of the elements 3 through 118. *J. Chem. Phys., 106,* **1997**, 3607-3612.
136. Carbó, R.; Domingo, L. LCAO-MO similarity measures and taxonomy. *Int. J. Quantum Chem., 32,* **1987**, 517-545.
137. Carbó, R.; Calabuig, B. Molsimil 88 - Molecular Similarity Calculations Using a CNDO-Like Approximation. *Comput. Phys. Commun., 55,* **1989**, 117-126.
138. Carbó, R.; Calabuig, B. Molecular similarity and quantum chemistry. In *Concepts and applications of molecular similarity*. Johnson, A.; Maggiora, G.M. (Eds.) John Wiley & Sons, Inc: New York, **1990**.
139. Carbó, R.; Calabuig, B. Quantum Molecular Similarity Measures and the Normal-Dimensional Representation of a Molecular Set - Phenyldimethylthiazines. *J. Mol. Struc. (Theochem), 86,* **1992**, 517-531.

140. Carbó, R.; Calabuig, B. Molecular Quantum Similarity Measures and N-Dimensional Representation of Quantum Objects .1. Theoretical Foundations. *Int. J. Quantum Chem., 42,* **1992**, 1681-1693.

141. Carbó, R.; Calabuig, B. Molecular Quantum Similarity Measures and N-Dimensional Representation of Quantum Objects .2. Practical Applications. *Int. J. Quantum Chem., 42,* **1992**, 1695-1709.

142. Carbó, R.; Calabuig, B.; Besalú, E.; Martínez, A. Triple Density Molecular Quantum Similarity Measures: a General Connection between Theoretical Calculations and Experimental Results. *Molec. Engineer., 2,* **1992**, 43-64.

143. Carbó, R.; Calabuig, B. Quantum Similarity Measures, Molecular Cloud Description; Structure Properties Relationships. *J. Chem. Inf. Comp. Sci.*, *32*, **1992**, 600-606.

144. Constans, P.; Carbó, R. Atomic Shell Approximation: Electron Density Fitting Algorithm Restricting Coefficients to Positive Values *J. Chem. Inf. Comput. Sci*., *35,* **1995**, 1046-1053.

145. Avilable at http: //iqc.udg.es/cat/similarity/ASA [last accessed 28 May 2004].

146. Ruedenberg, K.; Schwarz, W.H.E. Nonspherical atomic ground-state densities and chemical deformation densities from x-ray scattering. *J. Chem. Phys., 42*, **1990**, 4956-4969.

147. Coppens, P. In *International Tables for Crystallography*; *Vol. B.* Kluwer: Amsterdam, **1992**, 10.

148. Coppens, P.; Becker. In *International Tables for Crystallography*; *Vol. C.* Kluwer: Amsterdam, **1992**, 628.

149. Carbó-Dorca, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum Molecular Similarity Measures: Concepts, Definitions and Applications to Quantitative Structure-Property Relationships. In *Advances in molecular similarity*. *Vol. 1.* Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: Greenwich, CT, **1996**, 1–41.

150. Huzinaga, S.; Klobukowski, M. Well-tempered Gaussian basis set expansions of Roothaan-Hartree-Fock atomic wavefunctions for lithium through mercury. *J. Mol. Struct. (Theochem), 44,* **1988**, 1-87.

151. Ruedenberg, K.; Raffenetti, R.C.; Bardon, D. Energy, structure and reactivity. In *Proceedings of the 1972 Boulder Conference on Theoretical Chemistry*. Smith, D.W. (Ed.) Wiley: New York, **1973**, 164.

152. Schmidt, M.W.; Ruedenberg, K. Effective convergence to complete orbital bases and to the atomic Hartree-Fock limit through systematic sequences of Gaussian primitives. *J. Chem. Phys., 71*, **1979**, 3951-3962.

153. Feller, D.F.; Ruedenberg, K. Systematic approach to extended even-tempered orbital bases for atomic and molecular calculations. *Theor. Chim. Acta, 52,* **1979**, 231-251.

154. Jacobi, C.G.J. Über ein leichtes Verfahren, die in der Theorie der Säkularstörungen vorkommenden Gleichungen numerisch aufzulösen. *J. Reine. Angew. Math., 30,* **1846**, 51-94.

155. Wilkinson, J.H.; Reinsch, C. *Linear Algebra*. Springer-Verlag: Berlin, **1971**, 202-211

156. Pierre, D.A. *Optimization Theory with Applications*. Wiley: New York, **1969**.

157. Available at http: //iqc.udg.es/cat/similarity/ASA/basisset.html [last accessed 28 May 2004].

158. Binkey, J. S.; Pople, J. A.; Hehre, W. J. Self-consistent molecular orbital methods. 21. Small split-valence basis sets for first-row elements. *J. Am. Chem. Soc., 102*, **1980**, 939-947.

159. Gordon, M.S.; Binkley, J.S.; Pople J.A.; Pietro, W.J.; Hehre, W.J. Self-consistent molecular orbitalmethods. 22. Small split-valence basis sets for second-row elements. *J. Am. Chem. Soc., 104,* **1982**, 2797-2803.

160. Dobbs, K.D.; Hehre, W.J. Molecular orbital theory of the properties of inorganic and organometallic compounds. 4. Extended basis sets for third- and fourth-row, main-group elements. *J. Comput. Chem.,* 7, **1986**, 359-378.

161. Available at http: //iqc.udg.es/cat/similarity/ASA/table432.html [last accessed 28 May 2004].

162. Spiegel, M.R. *Mathematical Handbook.* McGraw-Hill: New York, **1968**.

163. Huzinaga, S.; Andzelm, J.; Klobukowski, M.; Radzio-Andzelm, E.; Sakai, Y.; Tatewaki, H. Gaussian basis set for Molecular Calculations. Elsevier: Amsterdam, **1984.**

164. Huzinaga, S. Gaussian-type functions for polyatomic systems. I. *J. Chem. Phys*., *42,* **1965**, 1293-1302.

165. Krishnan, R.; Binkley, J. S.; Seeger, R.; Pople, J. A. Self-consistent molecular orbital methods. XX. Abasis set for correlated wave functions. *J. Chem. Phys., 72,* **1980,** 650-654.

166. McLean, A.D.; Chandler, G.S. Contracted Gaussian basis sets for molecular calculations. I. Secondrow atoms, Z=11-18. *J. Chem. Phys.*, *72*, **1980**, 5639-5648.

167. Richards, W.G. Molecular Similarity and Dissimilarity. In *Modeling of Biomolecular Structures and Mechanisms*. Pullman, A.; Jortner, J.; Pullman, B. (Eds.) Kluwer Academic Publishers: Dordrecht, **1995**, 365-369.

168. Carbó–Dorca, R.; Besalú, E.; Gironés, X. Extended density functions. *Adv. Quantum Chem., 38,* **2000**, 3-63.
169. Girones, X.; Gallegos, A.; Carbó-Dorca, R. Modeling antimalarial activity: Application of kinetic energy density quantum similarity measures as descriptors in QSAR. *J. Chem. Inf. Comp. Sci., 40,* **2000**, 1400-1407.
170. Solà, M.; Mestres, J.; Oliva, J. M.; Duran, M.; Carbó, R. The use of ab initio quantum molecular self-similarity measures to analyze electronic charge density distributions. *Int. J. Quantum Chem., 58,* **1996**, 361-372.
171. Amat, L.; Carbó-Dorca, R.; Ponec, R. Molecular quantum similarity measures as an alternative to log P values in QSAR studies. *J. Comput. Chem., 19,* **1998**, 1575-1583.
172. Ponec, R.; Amat, L.; Carbó-Dorca, R. Molecular basis of quantitative structure-properties relationships (QSPR): A quantum similarity approach. *J. Comput. Aid. Mol. Des., 13,* **1999**, 259-270.
173. Besalú, E.; Carbó-Dorca, R.; Karwowski, J. Generalized one-electron spin functions and self-similarity measures. *J. Math. Chem., 29,* **2001**, 41-45.
174. Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to LFER: substituent and solvent effects on the acidities of Carbóxylic acids. *J. Phys. Org. Chem., 12,* **1999**, 447-454.
175. Amat, L.; Besalú, E.; Carbó-Dorca, R.; Ponec, R. Identification of active molecular sites using quantum-self-similarity measures. *J. Chem. Inf. Comp. Sci., 41,* **2001**, 978-991.
176. Gironés, X.; Carbó-Dorca, R.; Ponec, R. Molecular basis of LFER. Modeling of the electronic substituent effect using fragment quantum self-similarity measures. *J. Chem. Inf. Comp. Sci., 43,* **2003**, 2033-2038.
177. Robert, D.; Carbó-Dorca, R. Analyzing the triple density molecular quantum similarity measures with the INDSCAL model. *J. Chem. Inf. Comp. Sci., 38,* **1998**, 620-623.
178. Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular quantum similarity measures tuned 3D QSAR: An antitumoral family validation study. *J. Chem. Inf. Comp. Sci., 38,* **1998**, 624-631.
179. Robert, D.; Amat, L.; Carbó-Dorca, R. Three-dimensional quantitative structure-activity relationships from tuned molecular quantum similarity measures: Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comp. Sci., 39,* **1999**, 333-344.
180. Carbó, R.; Besalú, E. Extending Molecular Similarity to Energy Surfaces: Boltzmann Similarity Measures and Indices. *J. Math. Chem., 20,* **1996**, 247-261.
181. Sanz, F.; Martín, M.; Lapeña, F. and Manaut, F. Quantitative structure-activity relationships on MAO substrates by means of quantum chemical properties. *Quant. Struct.-Act. Relat., 5,* **1986**, 54-57.
182. Sanz, F.; Manaut, F.; José, J.; Segura, J.; Carbó, M. and De la Torre, R. Automatic determination of MEP patterns of molecules and its application to cafein metabolism inhibitors. *J. Mol. Struct. (Theochem), 170,* **1988**, 171-180.
183. Gironés, X.; Amat, L.; Carbó-Dorca, R. Using molecular quantum similarity measures as descriptors in quantitative structure-toxicity relationships. *SAR QSAR Environ. Res., 10,* **1999**, 545-556.
184. Gironés, X.; Amat, L.; Carbó-Dorca, R. Descripció de propietats moleculars i activitats biològiques emprant l'energia de repulsió electró-electró. *Sci. Gerun., 24*, **1999**, 197-208.
185. Girones, X.; Amat, L.; Robert, D.; Carbó-Dorca, R. Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comput. Aid. Mol. Des., 14,* **2000**, 477-485.
186. Carbó, R.; Besalú, E. *Adv. Quantum Chem., 24,* **1992,** 115.
187. Robert, D.; Carbó-Dorca, R. A formal comparison between molecular quantum similarity measures and indices. *J. Chem. Inf. Comp. Sci., 38,* **1998**, 469-475.
188. Gironés, X.; Ponec, R. Chemical structures from the analysis from of domain averaged Fermi Holes. Hypervalence and the nature of bonding in isocoordinated molecules SF6 and CLi6. *J. Phys. Chem. A, 106,* **2002**, 9506-9511.
189. Measures, P.T.; Mort, K.A.; Allan, N.L.; Cooper, D.L. Applications of momentum-space similarity. *J. Comput.-Aided Mol. Design, 9,* **1995**, 331-340.
190. Amat, L.; Carbó-Dorca, R.; Cooper, D.L.; Allan, N.L.; Ponec, R. Structure-property relationships and momentum-space quantities: Hammett σ Constants. *Mol. Phys., 101,* **2003**, 3159-3162.
191. Amat, L.; Carbó-Dorca, R.; Cooper, D. L.; Allan, N. L. Classification of reaction pathways via momentum-space and quantum molecular similarity measures. *Chem. Phys. Lett., 367,* **2003**, 207-213.

192.Cioslowski, J.; Stefanov, B.B.; Constans, P. Efficient Algorithm for Quantitative Assessment of Similarities among Atoms in Molecules. *J. Comput. Chem.*, *17*, **1996**, 1352-1358.

193.Robert, D.; Carbó-Dorca, R. On the extension of QS to atomic nuclei: Nuclear QS. *J. Math. Chem.*, *23*, **1998**, 327-351.

194.Robert, D.; Carbó-Dorca, R. Structure-property relationships in nuclei. Prediction of the binding energy per nucleon using a quantum similarity approach. *Nuovo Cimento A*, *111*, **1998**, 1311-1320.

195.Bach, A. Xarxes de funcions de base distribuïdes aplicades a càlculs SCF i CI de sistemes bielectrònics. DEA. Institute of Computational Chemistry. Girona, **1999.**

196.Fradera, X.; Duran, M.; Mestres, J. Second order QSM from intracule and extracule densities. *Theoret. Chem. Accounts*, *99,* **1998**, 44-52.

197.Fradera, X.; Duran, M.; Mestres, J. Comparison of quantum similarity measures derived from one-electron, intracule and extracule densities. In *Advances in Molecular Similarity; Vol. 2.* JAI Press, **1998**, 215-243.

198.Carbó-Dorca, R. Stochastic transformation of quantum similarity matrices and their use in quantum QSAR (QQSAR) models. *Int. J. Quantum Chem.*, *79,* **2000**, 163-177.

199.Carbó-Dorca, R. Quantum QSAR and the eigensystems of stochastic quantum similarity matrices. *J. Math. Chem.*, *27*, **2000**, 357-376.

200.Tou, J.T.; González, R.C. *Pattern recognition principles.* Addison-Wesley: Reading, M. A, **1974**.

201.Petke, J.D. Cumulative and discrete similarity analysis of electrostatic potentials and fields. *J. Comput. Chem. 14,* **1993**, 928-933.

202.Lobato, M.; Amat, L.: Besalú, E.: Carbó-Dorca, R. Estudi QSAR d'una familia de quinolones utilitzant índexs de semblança i índexs topològics de semblança. *Sci. Gerun.*, *23*, **1998**, 17-27.

203.Nyburg, S.C. Some uses of a best molecular fit routine. *Acta cryst., B30,* **1974**, 251

204.Haneef, B.I.; Moss, D.S.; Standford, M.J.; Borkakoti, N. Restrained Structure-Factor Least-Squares Refinement of Protein Structures Using a Vector Processing Computer. *Acta cryst., A41*, **1985**, 426-433.

205.Sippl, M.J.; Stegbuchner, H. Superposition of Three Dimensional Objects: A Fast and Numerically Stable Algorithm for the Calculation of the Matrix of Optimal Rotation. *Comput. Chem.*, *15,* **1991**, 73-78.

206.Allen, F.H.; Bellard, S.; Brice, M.D.; Cartwright, B.A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Humelink–Peters, B.G.; Ken-nard, O.; Motherwell, W.D.S.; Rodgers, J.R.; Watson, D.G. The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta cryst., B35*, **1979**, 2331-2339.

207.Artymiuk, P.J.; Bath, P.A.; Grindley, H.M.; Pepperrell, C.A.; Poirrette, A.R.; Thorner, D.A.; Wild, D.J.; Willet, P.; Allen, F.H.; Taylor, R. Similarity searching in databases of three-dimensional molecules and macromolecules. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 617-630.

208.Martin, Y.C. 3D Database Searching in Drug Design. *J. Med. Chem., 35*, **1992**, 35, 2145-2154.

209.Atai, A.; Tomioka, N.; Yamada, M.; Inoue, A.; Kato, Y. Molecular Superposition for Rational Drug Design. In *3D QSAR in Drug Design*. Kubinyi, H. (Ed.) ESCOM: Leiden, **1993**, 200-225.

210.Cramer III, R.D.; Paterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.*, *110*, **1988**, 5959-5967.

211.Kellogg, G.E.; Kier, L.B.; Gaillard, P.; Hall, L.H. The E-State Fields. Applications to 3D QSAR. *J. Comput.-Aided Mol. Design, 10,* **1996**, 513-520.

212.Bravi, G.; Gancia, E.; Mascagni, P.; Pegna, M.; Todeschini, R.; Zalianni, A. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. *J. Comput.-Aided Mol. Des., 11,* **1997**, 79-89.

213.Ren, J. Esnouf, R.; Garman, E.; Somers, S.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D.; Stammers, D. High resolution structures of HIV-1 RT from four RT-inhibitor complexes. *Nature Struct. Biol., 2,* **1995**, 293-302.

214.Bladon, P. A rapid method for comparing and matching the spherical parameter surfaces of molecules and other irregular objects. *J. Mol. Graphics*, *7,* **1989**, 130-137.

215.Masek, B.B.; Merchant, A.; Matthew, J.B. Molecular skins: a new concept for quantitative shape matching of a protein with its small molecule mimics. *Proteins*, *17,* **1993**, 193-202.

216.Kearsley, S.K.; Smith, G.M. An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Computer Methodology, 3*, **1992**, 615-633.

217.Gavuzzo, E.; Pagliuca, S.; Pavel, V.; Quagliata, C. Generation and best fitting of molecular models. *Acta Cryst., B28*, **1972**, 1968-1968.

218.McLachan, A.D. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Cryst., A28*, **1972**, 656-657.

219.Gerber, P.R.; Müller, K. Superimposing several sets of atomic coordinates. *Acta Cryst., A41*, **1987**, 426-428.

220.Redington, P.K. Molfit: a computer program for molecular superposition. *Comput. Chem., 16,* **1992**, 217-222.

221.Barakat, M. T.; Dean, P. M. *J. Comput. Aided Mol. Design, 4,* **1990**, 295.

222.Parretti, M.F.; Kroemer, R.T.; Rothman, J.H.; Richards, W.G. Alignment of Molecules by the Monte Carlo Optimization of Molecular Similarity Indices. *J. Comput. Chem., 18,* **1997**, 1344-1353.

223.McMahon, A.J.; King, P.M. optimization of Carbó Molecular Similarity Index Using Gradient Methods. *J. Comput. Chem., 18,* **1997**, 151-158.

224.Lemmen, C.; Lengauer, T.; Klebe, G. A method for fast flexible ligand superposition. *J. Med. Chem., 23*, **1998**, 4502-4520.

225.Dean, P.M.; Callow, P.; Chau, P.L. Molecular recognition: Blind-searching for regions of strong structural match on the surfaces of two dissimilar molecules. *J. Mol. Graphics, 6,* **1988**, 28-34.

226.Bayada, D. M.; Simpson, R.W.; Johnson, A.P.; Laurenco, C. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 680.

227.Brown, R. D.; Jones, G.; Willett, P. *J. Chem. Inf. Comput. Sci., 34*, **1994**, 63.

228.Artymiuk, P.J.; Grindley, H.M.; Poirrette, A.R.; Rice, D.W.; Ujah, E.C.; Willett, P. *J. Chem. Inf. Comput. Sci., 34*, **1994**, 54.

229.Rarey, M.; Dixon, J. *J. Comput. Aided Mol. Design, 12*, **1998**, 471.

230.Hermann, R.B.; Herron, D.K. OVID and SUPER: Two overlap programs for drug design. *J. Comput.-Aided Molec. Des., 5*, **1991**, 511-524.

231.Perkins, T.D.J.; Mills, J.E.J.; Dean, P.M. Molecular surface-volume and property matching to superpose flexible dissimilar molecules. *J. Comput.-Aided Molec. Des., 9*, **1995**, 479- 490.

232.Grant, J.A.; Gallardo, M.A.; Pickup, B.T. A Fast Method of Molecular Shape Comparison A simple application of a gaussian description of molecular shape. *J. Comput. Chem., 17,* **1996**, 1653-1666.

233.Clark, M.; Cramer III, R.D.; Jones, D.M.; Patterson, D.E.; Simeroth, P.E. Comparative Molecular Field Analysis (CoMFA). 2. Towards its use with 3D-structural databases. *Tetrahedron Comput. Method., 3,* **1990**, 47-59.

234.Manaut, M.; Sanz, F.; Jose, J.; Milesi, M. Automatic Search for maximum similarity between molecular electrostatic potential distributions. *J. Comput.-Aided Mol. Design, 5,* **1991**, 371-380.

235.Good, A.C.; Hodgkin, E.E.; Richards, W.G. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 188-191.

236.Mestres, J.; Rohrer, D.C.; Maggiora, G.M. MIMIC: A Molecular-Field Matching Program. Exploiting Applicability of Molecular Similarity Approaches. *J. Comput. Chem., 18,* **1997**, 934-954.

237.Amat, L.; Constans, P.; Besalú, E.; Carbó-Dorca, R. MOLSIMIL 97. Institute of Computational Chemistry, University of Girona: Spain, **1997**.

238.Nissink, J.W.M.; Verdonk, M.L.; Kroon, J.; Mietzner, T.; Klebe, G. Superposition of molecules: electron density fitting by application of Fourier transforms. *J. Comput. Chem., 18,* **1997**, 638-645.

239.Gironés, X.; Carbó-Dorca, R. TGSA99. Institute of Computational Chemistry, University of Girona: Spain, **1999**.

240.Gironés, X.; Carbó-Dorca, R. Extending the capabilities of the Topo-Geometrical Superposition Algorithm to handle rotary bonds. *J. Comp. Chem., 25*, **2004**, 153-159.

241.Carbó, R.; Besalú, E.; Amat, L and Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards theoretical foundation of quantitative structureproperties relationship. *J. Math. Chem., 18*, **1995**, 237-246.

242.Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat., 16,* **1997**, 25-32.

243.Mestres, J.; Solà, M.; Carbó, R.; Luque, F.J.; Orozco, M. Effect of Solvation on the Charge Distribution of a Series of Anionic, Neutral; Cationic Species. A Quantum Molecular Similarity Study. *J. Phys. Chem., 100*, **1996**, 606-610.

244.Amat, L.; Carbó-Dorca, R.; Ponec, R. Simple linear QSAR models based on Quantum Similarity Measures. *J. Med. Chem., 42,* **1999**, 5169-5180.

245. Carbó-Dorca, R.; Besalú, E. Quantum Theory of QSAR. *Contribution to Science, 1,* **2000**, 399-422.
246. Besalú, E.; Carbó, R. Quantum similarity topological indices: Definition, analysis and comparison with classical molecular topological indices. *Sci. Gerun., 21*, **1995**, 145-152.
247. Lobato, M.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Structure-activity relationships of a steroid family using quantum similarity measures and topological quantum similarity indices. *Quant. Struct.-Act. Relat.* **1997**, 16, 465-472.
248. Besalú, E.; Gallegos, A.; Carbó-Dorca, R. Topological quantum similarity indices and their use in QSAR: application of several families of antimalarial compounds. *MATCH-Commun. Math. Comput. Chem., 44*, 2001, 41-64.
249. Besalú, E.; Gironés, X.; Amat, L.; Carbó-Dorca, R. Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res., 35,* **2002**, 289-295.
250. Solà, M.; Mestres, J.; Carbó, R.; Duran, M. A comparative study of charge density distributions in H2CO derived from HF, MP2, QCISD and DFT methods. In *QSAR and Molecular Modelling: Concepts, Computational tools and Biological Applications*. Sanz, F.; Giraldo, J.; Manaut, F. (Ed.) Prous Science Pub: Barcelona, 403-406.
251. Solà, M.; Mestres, R. Carbó, R.; Duran, M. A comparative analysis by means of quantum molecular similarity measures of density distributions derived from conventional *ab initio* and density functional methods. *J. Chem. Phys., 104,* **1996**, 636–647.
252. Torrent, M.; Duran, M.; Solà, M. How Similar are HF, MP2; DFT Charge Distributions in the Cr(CO)6 Complex? In *Advances in Molecular Similarity; Vol. 1.* Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: Greenwich, **1996**, 167-186.
253. Forés, M.; Duran, M.; Solà, M. A procedure for assessing the quality of a given basis set based on quantum molecular similarity measures. *Theor. Mol. Mod. Electr. Conf., 1,* **1997**, 50-56.
254. Simon, S.; Duran, M. A QMS analysis of changes in molecular electron density caused by basis set flotation and electric field application. *J. Chem. Phys., 107,* **1997**, 1529-1535
255. Solà, M.; Forés, M.; Duran, M. Optimizing hybrid density functionals by means of quantum molecular similarity techniques. In *Advances in molecular similarity; Vol 2*. Carbó-Dorca, R.; Mezey, P.G. (Eds.) JAI Press: London, **1998**, 187–203.
256. Poater, J.; Duran, M.; Solà, M. Parametrization of the Becke3-LYP Hybrid Functional for a Series of Small Molecules Using Quantum Molecular Similarity Techniques. *J. Comput. Chem., 22,* **2001**, 1666-1678.
257. Martí, J. A new method for assessing similarities among atoms in molecules. *Chem. Phys., 265* **2001,** 263-271.
258. Vautherin, D.; Brink, D.M. Hartree-Fock calculations with Skyrme’s interactions. I. Spherical nuclei. *Phys. Rev., C5,* **1972**, 626-647.
259. Vautherin, D. Hartree-Fock calculations with Skyrme’s interactions. II. Axially deformed nuclei. *Phys. Rev., C7*, **1973**, 296-316.
260. Solà, M.; Mestres, J.; Carbó, R.; Duran, M. Use of *ab initio* Quantum Molecular Similarities as an Interpretative Tool for the Study of Chemical Reactions. *J. Am. Chem. Soc., 116*, **1994**, 5909-5915.
261. Solà, M.; Toro-Labbé, A. The Hammond Postulate and the Principle of Maximum Hardness in Some Intramolecular Rearrangement Reactions. *J. Phys. Chem. A, 103,* **1999**, 8847-8852.
262. Hammond, G.S. A Correlation of Reaction Rates. 'Hammond Postulate. *J. Am. Chem. Soc., 77,* **1955**, 334-338.
263. Bultinck, P.; Carbó-Dorca, R. Molecular quantum similarity matrix based clustering of molecules using dendrograms. *J. Chem. Inf. Comp. Sci., 43,* **2003**, 170-177.
264. Fradera, X.; Amat, L.; Torrent, M.; Mestres, J.; Constans, P.; Besalú, E.; Martí, J.; Simon, S.; Lobato, M.; Oliva, J.M.; Luis, J.M.; Andrés, J.L.; Solà, M.; Carbó, R.; Duran, M. Analysis of the changes on the potential energy surface of Menshutkin reactions induced by external perturbations. *J. Mol. Struct. (Theochem), 371,* **1996**, 171–183.
265. Mestres, J.; Solà, M.; Carbó, R.First-order molecular descriptors for molecular steric similarity. *Sci. Gerun., 21,* **1995**, 165-173.
266. Mezey, P.G.; Ponec, R.; Amat, L.; Carbó-Dorca, R. Quantum similarity approach to the characterization of molecular chirality. *Enantiomer, 4,* **1999**, 371–378.
267. Bach, A. Xarxes de funcions de base distribuïdes aplicades a càlculs SCF i CI de sistemes bielectrònics. DEA. Institute of Computational Chemistry. Girona, **1999.**
268. Bach, A.; Carbó-Dorca, R. Aplicació de la Semblança Molecular Quàntica en la reducció de l'espai configuracional per a l`'estat fonamental i primers exitats de l'àtom d'heli. *Sci. Gerun., 24,* **1999**, 183-196.

269. Amat, L.; Carbó-Dorca, R. Use of promolecular ASA density functions as a general algorithm to obtain starting MO in SCF calculations. *Int. J. Quant. Chem*., *87*, **2002**, 59-67.
270. Martín, M.; Sanz, F.; Campillo, M.; Pardo, L.; Pérez, J.; Turmo, J.; Aulló, J.M. Quantum chemical structure-activity relationships on b-carbolines as natural monoamine oxidase inhibitors. *Int. J. Quant. Chem., 23*, **1983**, 1643-1652.
271. Luque, F.J.; Sanz, F.; Illas, F.; Pouplana, R.; Smeyers, Y.G. Relationships between the activity of some H2-receptor agonists of histamine and their *ab initio* molecular electrostatic potential (MEP) and electron density comparison coefficients. Eur. *J. Med. Chem., 23*, **1988**, 7-10.
272. Sanz, F.; Manaut, F.; Dot, T.; López de Briñas, E. Complete or partial comparison of molecular electrostatic potential distributions? Some tests with 5-HT ligands. *J. Mol. Struct. (Theochem), 256,* **1992**, 287-293.
273. Sanz, F.; Manaut, F.; Rodriguez, J.; Lozoya, E.; Lopez de Briñas, E. MEPSIM: A computacional package for analysis and comparison of molecular electrostatic potentials. *J. Comput.-Aided Mol. Design*, *7,* **1993**, 337-347.
274. Richard, A.M. Quantitative comparison of molecular electrostatic potentials for structure activity studies. *J. Comput. Chem*., *12*, **1991**, 959-969.
275. Rum, G.; Herndon, W.C. Molecular similarity concepts. 5. Analysis of steroid-protein binding constants. *J. Am. Chem. Soc*., *113,* **1991**, 9055-9060.
276. Good, A.C.; So, S.S.; Richards, W.G. Structure-activity relationships from molecular similarity matrices. *J. Med. Chem*., *36*, **1993**, 433-438.
277. Good, A.C.; Peterson, S.J.; Richards, W.G. QSAR's from similarity matrices. Technique validation and application in the comparison of diferent similarity evaluation methods. *J. Med. Chem*., *36*, **1993**, 2929-2937.
278. Good, A.C.; Richards, W.G. The extension and application of molecular similarity to drug design. *Drug Information Journal*, *30*, **1996**, 371-388.
279. Besalú, E.; Amat, L.; Fradera, X.; Carbó, R. An application of the molecular quantum similarity: ordering of some properties of the hexanes. In *QSAR and molecular modelling: concepts, computational tools and biological applications*. Sanz, F.; Manaut, M. (Eds.) Prous Science: Barcelona, **1995**, 396-399.
280. Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des*., *13,* **1999**, 597-610.
281. Gironés, X.; Gallegos, A.; Carbó-Dorca, R. Antimalarial Activity of Synthetic 1, 2, 4- Trioxanes and Cyclic Peroxy Ketals, a Quantum Similarity Study. *J. Comput.-Aided Mol. Des.*;, 15, **2001**, 1053-1063.
282. Robert, D.; Carbó-Dorca, R. Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR & QSAR Environ. Res*., *10*, **1999**, 401-422.
283. Gallegos, A.; Robert, D.; Gironés, X; Carbó-Dorca, R. Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity. *J. Comput.-Aided Mol. Design, 15,* **2001**, 67-80.
284. Busacker, R.G.; Saaty, T.C. *Finite Graphs and Networks: An Introduction with Applications.* McGraw Hill Book Co.: New York, N. Y., **1965**.
285. Turner, J. *Proof Techniques in Graph Theory*. Harary, F. (Ed.) Academic Press: New York, N.Y., **1969**, 189.
286. Harary, F. *Graph Theory*. Addison-Wesley Pub. Co: Reading, MA, **1969.**
287. Essam, J. W.; Fisher, M. E. *Revs. Mod. Phys. 42*, **1970**, 272.
288. N. Trinajstic, N. *Rep. Mol. Th., 1,* **1990**, 185.
289. Rouvray, D.H. *Mathematical Chemistry*; *Vol. 1*. Bonchev, D.; Rouvray, D.H. (Eds.) Gordon and Breach: New York, **1991**, 1.
290. Rouvray, D.H. A rationale for the topological approach to chemistry. *J. Mol. Struc. (Theochem), 336,* **1995**, 101-114.
291. Hansen, P.J.; Jurs, P.C. Chemical Applications of Graph Theory. Part I: Fundamentals and Topological Indices. *J. Chem. Educ., 65*, **1988**, 574-580.
292. Hall, H. *A Computational Chemical Graph Theory*. Rouvray, D.H. (Ed.) Nova Science Publishers: New York, **1990**, 202-236.
293. Basart, J.M. *Grafs: Fonaments i Algorismes.* Publicacions UAB: Bellaterra, **1994**.
294. Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Vertex- and Edge-Weighted Molecular Graphs and Derived Structural Descriptors. In *Topological Indices and Related Descriptors in QSAR and QSPR*. Devillers, J.; Balaban, A. T. (Eds.) Gordon & Breach Science Publishers: The Netherlands, **1999**, 169-220.

295. Ivanciuc, O. QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs. *J. Chem. Inf. Comput. Sci., 40,* **2000**, 1412-1422.

296. Trinajstić, N. (1983). Chemical Graph Theory. CRC Press, Boca Raton, FL, 2nd ed.**1992**.

297. Mihalić, Z.; Trinajstić, N. A Graph-Theoretical Approach to Structure-Property Relationships. *J. Chem. Educ., 69,* **1992**, 701-712.

298. Randić, M. Chemical structure – What is she? *J. Chem. Educ., 69,* **1992**, 713-718.

299. Euler, L. Solutio problematis ad geometricam situs pertinentes. *Comm. Acad. Sci. Imp. Petropol. 8,* **1736,** 128-140. (Tr. Euler, L. *Sci. Am., 189,* **1953**, 66-70).

300. Image retrieved from http://www-gap.dcs.st-and.ac.uk/history/Miscellaneous/Konigsberg.html [accessed 25 April 2004].

301. Bošković, R.J. Theoria Philosophae Naturalis. In *Redacta ad unicam legem virium in natura existentium*. Remondini: Venetia, **1763.** (Tr. The Theory of Natural Philosophy. MIT Cambridge, MA, **1966**).

302. Bošković, R.J. *Školska knjiga*. Zagreb, **1987**.

303. Listing. Vorstudien zur Topologie, **1847.**

304. Rouvray, D.H. *J. Mol. Struct. (Theochem), 185,* **1989**.

305. Sylvester, J.J. *Nature, 17,* **1878**, 284.

306. Devillers, J. Balaban, A.T. (Eds.) *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach: The Netherlands, **1999**.

307. Rouvray, D.H. *Sci. Am., 254,* 1986, 40.

308. Rouvray, D.H. The Modeling of Chemical Phenomena Using Topological Indices. *J. Comput. Chem., 8,* **1987**, 470-480

309. Sylvester, J.J. *Am. J. Math, 1,* **1874,** 64

310. Cayley, A. *Philos. Mag., 13,* **1857,** 172-176.

311. A. Cayley, *Philos. Mag. 3,* **1877**, 34.

312. Kirchoff, G. *Ann. Phys. Chem., 72,* **1847**, 497-508

313. Seybold, P.G.; May, M.; Bagal, U.A. Molecular Structure-Property Relationships. *J. Chem-Educ., 64,* **1987**, 575-581.

314. Basak, S.C.; Grundwald, G.D. Molecular Similarity and Estimation of Molecular Properties.*J. Chem. Inf. Comput. Sci., 35,* **1995**, 366-372.

315. Grassy, G.; Calas, B.; Yasri, A.; Lahana, R.; Woo, J.; Iyer, S.; Kaczorek, M.; Floc'h, R.; Buelow, R. Computer-Assisted Rational Design of Immunosuppressive Comppounds. *Nature Biotechnol., 16,* **1998**, 748-752.

316. Ivanciuc, O.; Taraviras, S.L.; Cabrol-Bass, D. Quasi Orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure. *J. Chem. Inf. Comput. Sci, 40,* **2000**, 126-134.

317. Taraviras, S. L.; Ivanciuc, O.; Cabrol-Bass, D. Identification of Groupings of Graph Theoretical Molecular Descriptors Using a Hybrid Cluster Analysis Approach. *J. Chem. Inf. Comput. Sci, 40,* **2000**, 1128-1146.

318. Basak, S.C.; Grunwald, G.D. Molecular Similarity and Risk Assessment: Analog Selection and Property Estimation Using Graph Invariants. *SAR QSAR Environ. Res., 2,* **1994,** 289-307.

319. Lajiness, M. S. In *Computational Chemical Graph Theory.* Rouvray, D.H. (Ed.) Nova: New York, **1990**, 299-316.

320. Auer, C. M.; Nabholz, J. V.; Baetcke, K. P. Mode of Action and the Assessment of Chemical Hazards in the Presence of Limited Data: Use of Structure-Activity Relationships (SAR) Under TSCA, Section 5. *Environ. Health Perspect., 87,* **1990**, 183-197.

321. Johnson, M.A.; Basak, S.C.; Maggiora, G.A Characterization of Molecular Similarity Methods for Property Prediction. *Math. Comput.Modelling, 11,* **1988**, 630-634.

322. Basak, S.C. Grunwald, G.D. Estimation of lipophilicity from molecular structural similarity. *New Journal of Chemistry, 19,* **1995,** 231-237.

323. Austel, V. In *Topics in Current Chemistry; Vol 114*. Charton, M.; Motoc, I. (Eds.) Springer-Verlag: Berlin, **1983**, 7-19.

324. Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. In$ Comput. Sei., 25,* **1985**, 64-73.

325. Willet, P.; Winterman, V. A Comparison of Some Measures for the Determination of Intermolecular Structural Similarity: Measures of Intermolecular Structural Similarity. *Quant. Srruct-Act. Relat., 5,* **1986,** 18-25.

326. Basak, S.C.; Magnuson, V.R.; Niemi, G.J.; Regal, R.R. Determining Structural Similarity of Chemicals Using Graph-Theoretic Indices. *Discrete Appl. Math., 19,* **1988**, 17-44.

327. Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity Searching on CAS Registry Substances 1. Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci., 32,* **1992**, 111.
328. Basak, S.C.; Bertelsen, S.; Grunwald, G.D. Application of Graph Theoretical Parameters in Quantifying Molecular Similarity and Structure-Activity Studies. *Chem. Inf. Comput. Sci., 34,* **1994**, 270- 216.
329. Wilkins, C.L.; Randić, M. A Graph Theoretic Approach to Sttucture- Property and Structure-Activity Correlations. *Theor. Chim. Acta (Berl.), 58,* **1980,** 45-68.
330. Basak, S.C. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res., 15,* **1987,** 605-609.
331. Basak, S.C.; Gute, B.D.; Grundwald, G.D. Characterization of molecular similarity of chemicals using topological invariants. In, *Advances in Molecular Similarity*; *Vol 2.* Carbó-Dorca, R.; P.G. Mezey, P.G. (Eds.) JAI Press: Stanford, Connecticut, **1998,** 171-185.
332. Basak, S.C. Information theoretic indices of neighborhood complexity and their applications. In, *Topological Indices and Related Descriptors in QSAR and QSPR.* Devillers, J.; Balaban, A.T. (Eds.) Gordon & Breach: Amsterdam, **1999,** 563-593.
333. Basak, S.C.; Gute, B.D.; Grundwald, G.D. A hierarchical approach to the development of QSAR models using topological, geometrical and quantum chemical parameters. In, *Topological Indices and Related Descriptors in QSAR and QSPR.* Devillers, J.; Balaban, A.T. (Eds.) Gordon & Breach: Amsterdam, **1999**, . 675-696.
334. Purcell, W.P.; Bass, G. E.; Clayton, J.M. *Strategy of Drug Design*. Wiley: New York, **1973.**
335. Redl, G.; Cramer, R.D.; Berkoff, C.E. *Chem. Soc. Revs., 3,* **1974**, 273.
336. Martin, Y.C. *Quantitative Drug Design.* Dekker: New York, **1978**.
337. Kier, L.B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research.* Academic Press: New York, **1976.**
338. Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press: Letchworth, **1986.**
339. Diudea, M.V.; Ivanciuc, O. *Molecular Topology*. Comprex: Cluj, Romania, **1995.**
340. Balasubramanian, K. Geometry dependent connectivity indices for the characterization of molecular structures. *Chem. Phys. Let., 235,* **1995**, 580-586.
341. Balaban, A.T. From Chemical Graphs to 3D Molecular Modeling. In *From Chemical Topology to Three-Dimensional Geometry.* Balaban, A.T. (Ed.) Plenum: New York, **1997**, 1-24.
342. Balaban, A.T. *J. Chem. Inf. Comput. Sci. 25,* **1985**, 334.
343. Mezey, P.G. Computational Chemical Graph Theory. D.H. Rouvray (Ed.) Nova Science: Commack, NY, **1990**, 175.
344. Roberts, F.S. *Discrete Mathematical Model.* Prentice-Hall: Englewood Cliffs, NJ, **1976**, 58.
345. Randić, M. Molecular topographic descriptors. *Studies Phys. Theor. Chem., 54,* **1988**, 101-108.
346. Randić, M. On the characterization of three-dimensional structures. *Int. J. Quant. Chem.: Quant. Biol. Symp., 15,* **1988**, 201-208.
347. Randić, M.; Jerman-Blazic, B.; Trinajstic, N. Development of 3-dimensional molecular descriptors. *Comput. Chem., 14,* **1990**, 237- 246.
348. Pogliani, L. On a graph theoretical characterization of cisltrans isomers. *J. Chem. Inf. Comput. Sci., 34,* **1994**, 801-804.
349. Bogdanov, B.; NikoliC, S.; TrinajstiC, N. On the three-dimensional Wiener number. *J. Math. Chem., 3,* **1989**, 299-309.
350. Bogdanov, B.; NikoliC, S.; TrinajstiC, N. On the three-dimensional Wiener number. A comment. *J. Math. Chem., 5,* **1989**, 305-306.
351. P. J. Hansen and P. C. Jurs. *J. Chem. Educ., 65,* **1988**, 574-580.
352. Randić, M. *J. Math. Chem.* **1991**, 7, 155.
353. Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc., 69,* **1947**, 17-20.
354. Wiener, H. *J. Am. Chem. Soc., 69,* **1947,** 2636-2638.
355. Wiener, H. *J. Phys. Chem., 15,* **1947,** 766.
356. Wiener, H. *J. Phys. Chem., 52,* **1948,** 425-430.
357. Wiener, H. *J. Phys. Chem., 52,* **1948,** 1082.
358. Platt, J.R. *J. Chem. Phys., 15,* **1947**, 419-420.
359. Platt, J.R. *J. Phys. Chem., 56,* **1952**, 328-336.
360. Gordon, M.; Scantlebury, G.R. *Trans. Faraday Soc., 60,* **1964**, 605.
361. Hosoya, H. Topological index. A Newly proposed quantity Characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. soc. Jpn., 44*, **1971,** 2332-2337.

362.Randić, M.On characterization of molecular branching. *J. Am. Chem. Soc., 97*, **1975,** 6609-6615.

363.Kier, L.B.; Hall, L.H.; Murray, W.J.; Randić, M. Molecular connectivity. I. Relationship tot nonspecific local anaesthesia. *J. Pharm. Sci*., *64*, **1975,** 1971-1974.

364.Balaban, A.T. Highly discrimating distance-based topological index. *Chem. Phys. Lett., 89*, **1982,** 399-404.

365.Balaban, A.T. *J. Mol. Struct. (Theochem), 165*, **1988,** 243.

366.Schultz, H.P. Topological organic chemistry. 1. Graph theory and topological indices of alkanes. *J. Chem. Inf. Comput. Sci., 29*, **1989,** 227-228.

367.Mueller, W.R.; Szymanski, K.; Knop, J.V.; Trinajstić, N. Molecular topological index. *J. Chem. Inf. Comput. Sci., 30,* **1990**, 160-163.

368.Mihalic, Z.; Veljan, D.; Amic, D.; Nicolić, S.; Plavšić, D.; Trinajstić, N. The Distance Matrix in Chemistry. *J. Math. Chem., 11*, **1992,** 223-258.

369.Gutman, I.; Trinajstić, N. Graph theory and molecular orbitals. Total -electron energy of alternant hydrocarbons. *Chem. Phys. Lett., 17,* **1972**, 535-538.

370.Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, Jr. C.F. Graph theory and molecular orbitals. XII. Acyclic polyenes. *J. Chem. Phys., 62*, **1975**, 3399-3405.

371.Ren, B. A New Topological Index for QSPR of alkanes. *J. Chem. Inf. Comput. Sci., 39*, **1999**, 139-143.

372.Sabljić, A.; Trinajstić, N. *Acta Pharm. Jugosl., 31,* **1981**, 189.

373.Balaban, A.T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Topics Curr. Chem., 114,* **1983**, 21.

374.Rouvray, D.H. In *Chemical applications of Topology and Graph Theory*. King, R.B. (Ed.) Elsevier: Amsterdam, **1983**, 159.

375.Stankevich, M.I.; Stankevich, I.V.; Zefirov, N.S. *Russ. Chem. Rev., 57,* **1988**, 337.

376.Randić, M. *J Math Chem., 4,* **1990**, 157.

377.Besalú, E.; Carbó, R. Quantum Similarity Topological indices: definition, analysis and comparison with classical molecular topological indices. *Sci. Gerun., 21*, **1995,** 145-152.

378.Lobato, M.; Besalú, E.; Carbó, R. Relacions estructura-propietat per un conjunt d'hidrocarburs a partir de nous descriptors tridimensionals de la Semblança Molecular. *Sci. Gerun., 22*, **1996,** 79-86.

379.TOPO. Software developed to calculate topological indices. Institute of computational Chemistry: Girona, **2000**.

380.Balaban, A.T. *J. Chem. Inf. Comput. Sci*., 35, **1995**, 339-350.

381.Basak, S.C.; Grunwald, G.D.; Niemi, G.J. Use of graph theoretical and geometrical molecular descriptors in structure-activity relationships. In, From Chemical Topology to Three Dimensional Molecular Geometry. A.T. Balaban, A.T. (Ed.) Plenum Press, **1997**, 73-116.

382.Skvortsova, M.I.; Baskin, I.I.; Slovokhotova, O.L.; Palyulin, V.A.; Zefirov, N.S. Inverse problem in QSAR/QSPR [quantitative structure-property] studies for the case of topological indexes characterizing molecular shape (Kier indexes). *Dokl. Akad. Nauk., 324*, **1992,** 344-348. (Tr. *Doklady Chemistry, 324*, 103-107).

Quantitative Structure-Activity Relationships

When you can measure what you are speaking about,
and express it in numbers, you know something about it;
but when you cannot measure it,
when you cannot express it in numbers,
your knowledge is of a meager and unsatisfactory kind.
It may be the beginning of knowledge,
but you have scarcely, in your thoughts,
advanced to the stage of science.

Popular Lectures & Addresses 1891-1894
William Thomson (Lord Kelvin)

1 INTRODUCTION

Our society is faced with challenges such as new diseases like AIDS, drug resistance, and aggressive agricultural pest control processes, which can have a chemical agreement. Thus, there is an urgent need of predictive models for health hazard purposes, id est.; to design new drugs with improved properties and diminished side-effects, and to assess the safety of some chemicals, i.e. cosmetics. In addition, the assessment of the risk of chemicals released to the environment and the evolvement of environmentally benign synthetic methods is strongly required [1-2]. Furthermore, there is also a demand on scientific methods that replace or at least refine and reduce the use of laboratory animals. In particular, the U. S. Environmental Protection Agency (EPA) [4], and the European Centre for the Validation of Alternative Methods (ECVAM) [4] aims to develop and implement non-animal alternative tests into regulatory and validation procedures. These methods should be used in the design, and evaluation of experimental tests, and in the selection of appropriate test chemicals for validation studies.

Although the Chemical Abstracts Service registers an always-increasing number of pharmaceutical lead compounds every year, there is still a tremendous need to design quickly new drugs for curing human diseases. However, the cost to bring a new drug onto the market has dramatically risen. Therefore, the high cost in money and time of discovering and developing effective medicines has raised the investment of pharmaceutical companies. Notwithstanding, many of the concepts and methodologies applied to the design of pharmaceutically interesting compounds can be also applied to other compounds of scientific or commercial interest.

For a long time, medicinal chemists have systematically modified lead compounds with the driving force of synthetic feasibility, experience, and intuition. Using traditional techniques, it may take months to synthesise a new compound for biological testing. However, over the last decades, important contributions to the design of biologically active new compounds reducing experimental research costs come from **rational molecular design strategies** [5-7], such as biostructural research, computer-assisted data handling, data storage, retrieval, and processing from chemical databases [8-10], molecular modelling, and, specially, structure-based design, **structure-function correlation studies**, and other statistical techniques.

In rational drug design for human health hazard, and environmental risk assessment purposes, several statistical mathematical techniques are employed to unravel information obtained from the available biological and chemical data, and to obtain a sound chemical insight of the problem for several applications. Among the most useful applications, priority setting, risk assessment, classification and labelling for regulatory purposes must be remarked. But the discovery of biologically active compounds and their development as drugs is a highly complex process which involves many scientific disciplines [11], i.e. structural, cell, and molecular biology, microbiology, biochemistry, synthetic inorganic and organic chemistry, medicinal chemistry, biophysical chemistry, toxicology, pharmacology, natural products chemistry, chemical ecology, mathematics, computing, and information technology, among others. Hence, selected and clinically tested drugs developed from bioactive molecules require specific conditions. Thus, not all the biologically active compounds are suitable to be used as drugs due to toxicity, unfavourable side-reactions, or pharmacokinetics. So, after the synthesis and aside from the desirable therapeutic properties, the testing and approval of many rigorous tests are required to ascertain whether the compounds are worthy of becoming drugs. Therefore, there is not just one technique of computational chemistry that plays a leading role in drug discovery, but rather an integrated approach of experimental science with computational techniques.

Both molecular modelling techniques and quantitative statistical methods may be useful in elucidating structural information of active compounds. Since a biological effect seldom depends on just one or two chemical properties, the **multidimensional problem** takes into account a large number of factors, rationalised to cover a broad parameter space. In order to be able to deal with complex data sets, consisting of more than one biological activity and many descriptors, advanced statistical and computational tools have been developed in the field of chemometrics. The term **chemometrics**, coined in the 1970s, is the chemical discipline that uses statistical and mathematical methods for selecting and optimizing procedures for the analysis and interpretation of data. These techniques allow the rapid retrieval and prediction of molecular and biological properties by means of multivariate methods and artificial intelligence techniques [12-16].

The strategy of **structure-based molecular design** has been proven to be very successful in the pharmaceutical industry [17]. However, when structural information about the biological target is lacking, the strategy of lead finding involves the synthesis and testing of widely diverse compounds. The systematic variation of substituents in a molecule has been the subject of various studies in the past. As it is not straightforward to select a representative subset of substituents that adequately covers the multidimensional parameter space, relevant properties derived from large sets of property descriptors and selected by using statistical techniques can be used to make rational choices.

Besides, **combinatorial chemistry** has arrived on the scene; nowadays, instead of manually synthesising all compounds, it is possible to assemble chemical building blocks in all combinations, generating large virtual libraries of structurally related compounds by means of automated procedures [18]. **High Throughput Screening (HTS)** and **data mining** techniques screen the databases with a defined query, usually a pharmacophore, testing hundreds up to millions compounds, and looking for relevant information. In absence of a known pharmacophore these techniques can also detect the most occurring fragments. Combinatorial approaches seek to maximise the structural diversity of the final library, i.e. the degree of heterogeneity, that is, the structural range or dissimilarity, to ensure the coverage of the largest possible expanse of chemical space in the search for bioactive molecules [19]. These computational tools improve molecular diversity and the chance of lead discoveries. The ready availability of chemical structure databases plays an important role in enhancing the drug discovery approach [20]. These databases find increasing use in environmental, inorganic, and organic chemistry. The combinatorial chemistry supporting technologies not only have risen the number of compounds synthesised and tested for every new chemical entity, but also provide a far more cost-effective approach to the discovery of bioactive compounds, in comparison with traditional approaches that require the sequential synthesis and testing of individual molecules.

Hence, as the economical resources for chemical synthesis and biological testing are limited, there is a need for preventing or at least slowing down further increases in the synthesis of compounds too specific. A promising way to achieve this purpose is to investigate the causes of diseases and the possibilities of intervention at the molecular level and to design promising test compounds by means of statistical **experimental design techniques**.

Experimental design methods can be divided into two categories. On one hand, methods for the direct optimization of lead compounds, only suitable in the final stages of an optimization procedure. These methods usually cover only a limited area of the parameter space surrounding a previously identified active compound. Therefore, active compounds located in other areas cannot be detected. On the other hand, methods for the systematic investigation of the parameter or descriptor space, applicable at any stage of the search for new drugs. These techniques provide a strong basis to derive reliable qualitative and quantitative structure-function relationships. These methods, which can be used for the complete coverage of the descriptor space by a minimum number of compounds, analyse if the potential drug presents an appropriate pattern of properties in the correct spatial arrangement.

Computer-Aided Molecular Design (CAMD) or, more specifically, Computer-Aided Drug Design (CADD), is a unifying discipline focused on the prediction of chemical reactivity for non-synthesised, virtual structures. CAMD emphasises the development of predictive tools for molecular properties in order to understand structure-function relationships. Rational molecular design assisted by computer embraces an interdisciplinary combination of methodologies of computational chemistry and information technology that aims to discover and design new and useful compounds. The main techniques of computational chemistry are molecular graphics and data visualization, quantum chemistry, molecular dynamics and mechanics, and structure-based methods, such as molecular and homology modelling, molecular databases and diversity analysis, receptor-based pharmacophore modelling, docking, scoring, and **Quantitative Structure-Activity Relationships (QSAR)**. CAMD techniques have a wide application in several fields of chemistry, such as organic chemistry, medicinal chemistry, environmental chemistry, guest-host chemistry (design of enzyme inhibitors, clinical analytical reagents, and catalysts), and also agricultural, veterinary, human health, and materials science (polymer chemistry, supramolecular chemistry, semi-conductors, and nonlinear phenomena), among others.

However, in relation to the application of computers as tools in the drug design process, it is important to emphasise that computers cannot substitute for a clear understanding of the system being studied. That is, the computer is only an additional tool to gain better insight into the chemistry and biology of the problem. Researchers have attempted for many years to develop drugs based on rational drug design methods [21]. Easy access to computational resources was not available when these efforts began. Hence, attempts consisted primarily of statistical correlations of structural descriptors with biological activities. However, as access to high-speed computers and graphics workstations became common, this field evolved into what is nowadays known as rational drug design or computer-assisted drug design.

Structure-function correlation studies aim at broadening the understanding of relationships between molecular intrinsic chemical features, and biological properties. Particular cases of these studies are the **Quantitative Structure-Activity Relationships (QSAR),** and **Quantitative Structure-Property Relationships (QSPR),** which extend the same notion to general physicochemical property prediction, and **Quantitative Structure-Toxicity Relationships (QSTR),** within the environmental and health sciences field**.** For the sake of simplicity, throughout this work they will be generally named as QSAR [22-24].

2 COMPUTER-AIDED MOLECULAR DESIGN

Starting in the 1950s, improving computer technology made possible the transformation of quantum mechanics from a pencil and paper attempt assisted by a hand calculator to a feasible task, where interesting molecular properties could be computed by solving the Schrödinger equation. Different treatments of quantum chemists were approached. On one hand, the rigorous application of the theory, confined to small molecules with only few atoms, that is, the so-called *ab initio* (from first principles) approach, of interest for theorists. On the other hand, the semiempirical molecular orbital approach, going before theoretical purity to render the methods applicable to large enough molecules of biological and commercial interest. In 1964, the Lilly pharmaceutical company initiated a research program to exploit this approach in the study of drugs. Meantime, several empirical modelling methods were developed by physical organic chemists, allowing the three-dimensional conformational treatment of large biologic systems. These so-called molecular mechanics methods, faster than either *ab intio* or semiempirical quantum methods compute simulations and relative energies of large biomolecules, including enzymes, nucleic acids, polypeptides and proteins. Besides, in the 1960s and 1970s other physical organic chemists developed methods for computing the lipophilicity of compounds, with strong implications in drug design. Specially, the lipophilic o hydrophilic character of a compound determines its ability to cross the membrane’s lipid barrier. If a medicine is administrated, the drug has to dissolve in the aqueous environment to enter the blood and then pass through several membranes to reach the receptor.

Nowadays, one of the main goals of drug research is to discover ligands (potential drugs) that are predicted to interact favourably and bind strongly to its intended host (receptor active site), without interfering with the operation of other bio-macromolecules in the living organism. Alternatively, this procedure can be reversed to search for hosts that interact strongly with a given ligand.

Whereas most drugs are ligands, only few ligands are drugs, because even small variations in chemical structure can influence whether the compound will be curative, physiologically inert, or toxic. Receptor sites have the ability to first attract and then bind to a ligand through hydrophobic interactions, and electrostatic interactions between oppositely charged atoms or groups, such as hydrogen bonding interactions. In addition, solvation energies of the ligand and receptor site are also important because partial to complete desolvation must occur prior to binding. To a rough approximation, compounds possessing similar shape, volume, lipholicity, electronic distribution, and chemical stability, cause similar effects in a biochemical system. The tools to simulate and visualise these molecular properties, which determine the occupation of the “lock” by the “key”, are provided by computational chemistry. As organic and physical chemists search for guest-host systems with specificity in binding and catalysis, the basic concepts of molecular field analysis, receptor mapping, and molecular recognition [25], or how enzymes recognise and bind the proper substrates, are unifying tools in this research area. Basically, CAMD entails a collection of computer-based methods that study molecular structures and properties and allow the determination of activities as well as the access of prior knowledge of databases. Computational tools make possible the discovery of new molecules with useful characteristics or old molecules with new uses.

However, the optimum fit of a ligand in a receptor site does not guarantee that the desired activity of the drug will be enhanced or that undesired side effects will be diminished. Moreover, this approach does not consider the pharmacokinetics of the drug, and therefore this approximation is dependent upon the amount of information that is available about the ligand and the receptor. In order to simulate the behaviour of an unknown chemical system, ideally, three-dimensional structural information for the receptor and the ligand-receptor complex from X-ray diffraction or **Nuclear Magnetic Resonance** (NMR) are required to assist in building models of the ligand and receptor. Or, at least, a simplified description of the system obtained by modelling is required. From here, the need for enhanced uses of molecular mechanics and dynamics, molecular orbital calculations, and chemical information technology becomes apparent. Despite the simplifications, the model should capture the essence of a particular property or process of interest. In CAMD, molecular models are subjected to computational experiments to deduce information about their properties and thereby to ascertain which hypotetical new molecular structures will have the desired properties. The modelling methods in CAMD, although there are not yet good enough to predict exactly how a molecule will behave in a test system or what properties a compound will have, can provide helpful information, especially if they are used in a complementary or even synergistic way.

The fundamental concept behind CAMD relies in the fact that molecular and electronic structure underlies all physical and chemical properties of molecules, including their biological activity. These physical particles determine the reactivity and the physical properties of the compound that, in turn, describe the interaction of the molecule with other molecules. The interactions influence solubility, lipophilicity, association, and stability, which affect transport of a compound to the active site. The drug-receptor interactions also define how well the compound attaches to the receptor by first being recognised as complementary to the receptor structure in shape and electronic structure. The affinity between the compound and the receptor determines how well a biochemical or conformational change in the receptor is induced. The latter change is eventually observable in the patient as a therapeutic response to the drug.

In applying this approach, it must be noted that lead molecules must be aligned so that the active functional groups of molecules are overlapped in the space. Therefore, conformational analysis is desirable, the extent of which is dependent on the flexibility of the compounds under investigation and the presence of rotatable bonds. So, once the active conformer is found, the molecule should be rotated to be aligned with the other molecules in the study. To guess the active conformer, the number of possible conformations can be restricted using aromatic drugs with ring systems. In this way, an inflexible active lead compound is chosen so that conformations of the more flexible leads mimic the inflexible ones. Another strategy is to find the lowest energy conformers of the most rigid compounds and superimpose them. Conformational searching on the more flexible compounds is done while applying distance constraints derived from the structures of the more rigid compounds. Ultimately, all of the structures are superimposed to generate the pharmacophore. This template may then be used to develop new compounds with functional groups in the desired positions. In applying this strategy, the minimum energy conformers are assumed to bind most favorably in the receptor site although, in fact, there is no *a priori* reason to exclude higher energy conformers as the source of activity. Another difficulty in alignment is that the active functional groups in the pharmacophore are usually unknown.

Based on the available information, either ligand-based or receptor-based molecular design methods can be applied. Ligand-based design uses a known set of ligands, but an unknown receptor site, whereas receptor-based design starts with a known receptor, such as a protein binding site or supramolecular host. Both approaches are actually very similar.

2.1 Receptor-based approach

The receptor-based approach applies when a reliable model of the receptor site is available. With the availability of the receptor site, docking techniques design ligands that interact favourably at the site. The first phase is to determine the structure of the binding site using standard structural analysis from X-ray diffraction, NMR, homology modelling, or calculations involving molecular orbital or molecular mechanics and molecular dynamics techniques. In the absence of structural information, homology of the unknown receptor sequence with known structures that have been identified through database searches may be a good starting point. Thus, 2D substructure searches include functional groups, and connectivity; 3D substructure searches are related to spatial relationships, definition of ranges for distances and angles, and the stored conformation usually taken to be the lowest energy conformation. 3D conformationally flexible searches, involve rotation around all freely rotatable bonds, the consideration of many conformations to generate a large number of hits, and the consideration of guest-host chemistry. Several three-dimensional databases can be used as source [26-29].

The next phase is to generate a query for database searching. The query is generated by building a simplified model of the receptor site. This model may be based on a pharmacophore, which identifies few specific interactions that are responsible for the binding. These interactions include hydrogen bond donors and acceptors, charged groups, and hydrophilic regions such as hydrocarbon side chains, and phenyl groups. The pharmacophore can be generated by visual inspection or by computational techniques. In general, it is assumed that the active site properties are complementary to active lead drugs (ligands). The receptor must also minimise steric repulsion and maximise favourable Van der Waals interactions. To guess the shape of the active site, the 3D-volume occupied by the active leads is examined. The receptor model, which must not have groups that extend into the volume occupied by the drug, is taken as the combined Van der Waals surfaces of the active leads. In docking-based searches, the model is based on an analysis of steric interactions over the receptor site. Typically, a solvent accessible surface map is generated and binding pockets are identified on the host surface.

The next phase is to search databases for new ligands that may bind to the chosen receptor. The searches can be from the bidimensional substructure (functional groups and connectivity), steric search (docking), or three-dimensional substructure searches (spatial conformation, and pharmacophores). The three-dimensional pharmacophore is used in conformationally flexible searches for ligands that match the spatial distribution of the receptor. Alternatively, the receptor pocket can be used with auto-docking to find ligands that avoid close-contacts. The 3D-pharmacophore approach and the binding pocket approach are actually very similar, and queries can be fashioned that incorporate aspects of both approaches. Pharmacophores emphasise a few specific and varied types of interactions, while binding pockets emphasise steric interactions over the entire ligand. The results of the database search may be used directly or modified to produce candidates for further study, which constitute the design element of the procedure. The new ligands or hosts are then assessed for the use at hand. This assesment first involves docking the new molecule and evaluation of the full interaction by molecular orbital calculations or molecular mechanics. Next, calculations are done to predict the binding constant or activity of the compound. Prediction of the binding constants is usually performed using Gibb's free energy perturbation studies based on either Monte Carlo or molecular dynamics simulations. Activity predictions are usually based on QSAR extrapolation. Increasingly these QSAR predictions are based on the 3D-QSAR used to generate the pharmacophore in the search stage. Finally, the candidates are synthesised and tested in the laboratory.

2.2 Ligand-based approach

The ligand-based approach is applicable when the structure of the receptor site is unknown, but a series of compounds has been identified to exert the activity of interest. To be used most effectively, structurally similar compounds with high activity, with no activity, and with a range of intermediate activities are required. Recognition site mapping attempts to identify a pharmacophore, which is the template derived from the structures of the compounds. In this case, the pharmacophore is a three-dimensional space representation of the collection of common functional groups within the group of active compounds, complementary to the geometry of the receptor site. It can be pictured as a search query used to search a database for compounds with might have similar biological activity.

Ligand-based design starts with a group of ligands that have known binding constants or biological activities. The first phase is to determine the structure of the ligands using the same elucidation techniques as before. The next phase is to generate a query for database searching, based on a pharmacophore, as in receptor based design. The pharmacophore can be also generated by visual inspection or by statistical techniques, such as 3D-QSAR, which maps the steric, charge, and hydrogen bonding interactions into a 3-D grid for each known ligand. These maps are then compared to find features that the active compounds have in common. The map of common features is then converted into a pharmacophore. The next phase is to search databases for new ligands that may also bind to the chosen receptor. 2D-substructure searches based on the known ligands can be used, but such searches have not been very successful. Instead, the 3D-pharmacophore is used in conformationally flexible searches for ligands that match the spatial distribution of the known ligands. The remainder of the phases are identical for ligand and receptor based design.

3 OVERVIEW OF QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

QSAR attempt to correlate structural molecular properties (descriptors) with functions (i.e. physicochemical properties, biological activities, toxicity, etc.) for a set of similar compounds, by means of statistical methods. As a result, a simple mathematical relationship is established:

$$\text{Function} = f\ (\text{structural molecular or fragment properties}) \qquad (1)$$

QSAR techniques include from chemical measurements and biological assays to the statistical techniques and interpretation of results.

Applications of QSAR can be extended to any molecular design purpose, including environmental sciences, prediction of different kinds of biological activity by correlation of congeneric series of compounds, lead compound optimization, classification, diagnosis and elucidation of mechanisms of drug action, and prediction of novel structural leads in drug discovery.

3.1 Objective

The goal of structure-activity modelling is analyse and detect the determining factors for the measured activity for a particular system, in order to have an insight of the mechanism and behaviour of the studied system. For such purpose, the employed strategy is to generate a mathematical model that connects experimental measures with a set of chemical descriptors determined from the molecular structure for a set of compounds. The model derived should have as good predictive capabilities as possible to predict the studied biological or physicochemical behaviour for new compounds. The factors governing the events in a biological system are represented by a multitude of physicochemical descriptors, which can include parameters to account for hydrophobicity, electronic properties, steric effects, and topology, among others. These descriptors were determined empirically in the past but, more recently, they can be calculated by using computational methods.

In particular, in CADD, the analysis of statistical relationships between molecular structural features and the therapeutical effect of a medicine derived by correlation facilitates the understanding of how chemical structure and biological activity relate. The building of a model with relevant and consistent chemical descriptors provides insight into various underlying biological processes; and the prediction for new compounds increases the number of candidate compounds to be considered.

In summary, the objectives of QSAR models are to allow the prediction of biological activities of untested and sometimes yet unavailable compounds, and to provide insight of which relevant and consistent chemical properties are determinant for the biological activity of compounds.

3.2 Underlying principles

The QSAR approach, based in the numerical representation of chemical structure, aims to understand how structural variation affects the biological activity of a class set of compounds. The main assumption is that the factors governing the events in a biological system are represented by the descriptors characterizing the compounds, whose biological activity is expressed via the same mechanism. So, QSAR attempt to find what features of a molecule affect its activity and what can be modified to enhance their properties. Hence, for a series of biologically active molecules, any systematic variation in chemical structure from one to another is expected to be reflected in a proportional analogous variation in the biological response.

The QSAR paradigm is based on the assumption that there is an underlying relationship between molecular structure and biological activity, which arises from this systematic variation. Also, it is assumed that the multivariate physicochemical description of the set of compounds reveals these analogies. All physical, chemical, and biological properties of a chemical substance can be computed from its molecular structure, encoded in a numerical form with the aid of various descriptors.

However, it is important to remark the difference between correlation and causation. A satisfactory QSAR correlation does not mean that a particular descriptor causes the efficient action of a compound. The lack of evidence on causation might be complemented by additional information on the various mechanisms leading to the biological activity.

Indeed, QSAR is based in the well-known similarity principle, which states that similar compounds have similar activities. In addition, in linear relationships, besides from the afore-mentioned similarity principle also the neighbourhood principle holds. This principle states that, in principle, molecules located in the same region of the descriptor space present similar activity.

3.3 QSAR model

QSAR expresses a multivariate mathematical relationship between a set of physicochemical properties or descriptors, $\{x_{ij}\}$, and a experimental function or biological activity, $\{y_i\}$. The QSAR relationship is expressed as a mathematical model, quantitative in the sense that it is used to account for the observed activity. For a compound *i*, the linear equation that relates molecular properties, x_1, x_2…, to the desired activity, *y*, is:

$$y_i = x_{i1}b_1 + x_{i2}b_2 + \ldots + x_{in}b_n + e_i \tag{2}$$

Expressing the previous equation in a compact form for the general case of *n* selected descriptors, x_j, the QSAR equation results into:

$$y_i = \sum_{j=1}^{n} x_{ij}b_j + e_i \tag{3}$$

where *b* are the linear slopes that express the correlation of the particular molecular property x_{ij} with the activity y_i of the compound *i*; and e_i is a constant. The slopes and the constant are often calculated using regression analysis. In this work, only the models with a single dependent variable, or *y* observation will be considered, although some models can deal with several biological activities. The strength of a QSAR model depends on the quality of this variable.

The independent variables, so-called descriptors, are usually physicochemical properties that describe some aspects of the chemical structure, which may be either experimentally or theoretically determined. The improper choice of independent variables can result in poor QSAR models. In a typical QSAR study, a large number of descriptors can be used; however, attention must be paid to overfitting, because with enough parameters any model can be successfully correlated. The final QSAR equation seeks to find the smallest number of descriptors that can adequately model the activity of the compounds in the study. The maximum recommended ratio is a single independent variable to five compounds.

3.4 Conditions for applicability of QSAR

To develop a valid and reliable QSAR model, there are a number of conditions related to the different aspects that conform the model that must be fulfilled [30-31]. The biological activity of the series of compounds should be related to the physicochemical properties being considered. Also, the activities of the chemicals covered should be elicited by the same elucidated mechanism, which is both common and relevant. A related chemical structure is not strictly required; however, it is often difficult, if not impossible, to determine the mechanism of action, whereas it is usually less difficult to establish chemical similarity. Hence, QSARs are generally developed for congeneric molecular series, that is, for sets of chemically similar compounds, in the hope that they will also have the same mechanism of action. The compounds with a different mechanism of action are likely to fit the correlation only poorly and to appear as outliers. Therefore, the main required conditions to build a valid QSAR model regard the leading areas involved in the development such models.

3.4.1 Selection of Compounds

The selection of appropriate chemical sets to develop QSAR models is of great importance to obtain valid results [41]. A suitable set should consist of those chemicals that exert a given activity effect via a common mechanism that can be modelled by a single QSAR equation.

3.4.1.1 Homogeneity

The requirement of chemical similarity and homogeneity of compounds implies that the investigated system must have the same mechanism of influence and that there are some limits on the variability and diversity of chemical structures and properties. Thus, the absence of influential outliers (compounds that do no fit the model), and strong clustering is desired. Clustering occurs when several classes of compounds can be separated into different subgroups. In this case, the option o treat each class independently must be considered.

3.4.1.2 Representativity

The selection of the set of compounds must span the chemical domain of interest, according to the definition of the chemical space. Therefore, a wide range as possible of relevant chemicals must be selected to assess its utility. For such purpose multivariate design [32] known as **Statistical Molecular Design (SMD)** can be used [33]. A variety of selection procedures has been proposed including **non-statistical methods** [36], **Cluster Analysis (CA)** [37], **Factorial Design (FD), Fractional Factorial Design (FFD)** [38], **Central Composite Design (CCD)**, **Principal Component Analysis (PCA)** [39] and **D-Optimal Design (DOD)** [40, 43].

3.4.2 Selection of Descriptors

Concerning the physicochemical or structural descriptors, especially when a large number of descriptors are calculated, some of them may content redundant information, resulting in the collinearity problem. The parameters used in QSAR should be meaningful, and easily interpretable, in physical terms. Thus, the selected descriptors should provide valuable insight into the mechanism [34].

3.4.3 Biological data

Concerning to the experimental activity, high quality and reliable biological data is required. Biological activities should have been measured in a consistent manner, by using well standardised assays with a clear and unambiguous endpoint [34]. Ideally, the data source should come from the same protocol and, if possible, the same laboratory. In data extracted from the literature, it is convenient to take only a single source into account. Besides, mechanistic insights of the chemicals must be attentively considered, in order to have a basis to detect and reject outliers of the model. Finally, it must be taken into account that biological measurements are subjected to experimental error.

The methods to evaluate biological and, in particular, toxicological endpoints are in order of increasing complexity: *in silico* methods, accounting for electronic and general molecular properties, *in vitro* methods, which provide a satisfactory description at cellular level, and *in vivo* methods, suitable to more detailed studies on specific organs and individuals.

3.4.3.1 Types of data

Biological data can be distributed on a continuous scale, so that a quantitative QSAR equation can be derived by means of correlation techniques, or sometimes it can be classified into discrete categories. For example, a chemical may be classified as either active or inactive, or in several classes according to the potency of the activity.

In such cases, other statistical techniques, such as classification methods must be applied, in which the physicochemical properties of the compounds are used to discriminate between activity and inactivity. If more than two such properties are used, they can be combined into principal components, and a plot of two major principal components may distinguish the different classes.

3.4.3.2 Data scaling

Logarithmic expressions are usually used for different reasons. From one point of view QSARs represent free energy changes, and by analogy the logarithm is used. From another point of view, sometimes an effective compound has a low concentration for the production of the desired effect. In QSAR studies, it is often desirable to have a higher activity value for the more effective compounds. Therefore, it is very common to transform the concentration for a desired effect, [*C*], to an activity by a logarithmic expression, log (*C)* or log (1/*C*), which increases with compound efficacy. Also, biological data is often skewed, so that logarithmic transforms fit the data to a normal distribution.

3.4.4 Some advice

Finally, QSAR models should be simple, transparent, and mechanistically comprehensible [34]. Although statistical techniques will be considered afterwards in detail, it is important to remark that overfitting and non-linearity of data should be avoided, whereas transparency and validation of the model are strongly recommended.

In summary, the so-called SETUBAL principles [42] are required to obtain a valid QSAR model. These principles state some conditions for the successful development of QSAR models that should:

1) Be associated with a defined pharmacotoxicological endpoint which it serves to predict
2) Take the form of an unambiguous and easily applicable algorithm for predicting the endpoint
3) Ideally have a clear mechanistic basis
4) Be accompanied by a definition of the domain of its applicability
5) Be associated with a measure of goodness-of-fit, assessed by internal validation
6) Be assessed in terms of its predictive predictive capacity by external validation

Complementary to the desired principles, there are a number of caveats to be borne in mind for the valid application of QSAR:

- QSAR can be applied only to pure compounds. Some work has been undertaken on their application to mixtures but, up to now, there are no firm guidelines for their use in this respect.
- The set of compounds used to derive the QSAR should be selected from knowledge, or assumptions, of a common mechanism of action. The training set should also be chosen to cover appropriate ranges of parameter values.

- The parameters or descriptors used in the QSAR equation should be selected on the basis of mechanistic considerations, in order to provide mechanistic interpretation.
- For comparative purposes, concentrations or doses must be in molar, not weight, units.
- Each QSAR should be validated by investigating its predictive ability using the test set, a different set of compounds, which should cover the same ranges of parameter space.
- The QSAR should not be applied outside of its domain of validity, i.e. outside of the parameter space covered by the training set.

3.5 QSAR Steps

The **strategy for QSAR development** in drug design consists of several iterative steps, based on **statistical experimental design** and **multivariate data analysis**, which hopefully lead to the design of compounds with the desired activity profile [43-45].

1) Formulation of classes of similar compounds

The first step consists of the selection of the biological activities of interest, the choice of structural domain (structural class) and the choice of structural features to be varied. Since the mechanism of biological action usually differs between different types of compounds, it is not desirable that QSARs are based on compounds too diverse structurally. Thus, the ideal situation corresponds to **classes of chemically and biologically similar compounds** were, within each class, all the compounds are structurally similar and function according to the same **mode of action**. However, the compounds must be dissimilar enough to cause some **systematic change** in the biological activity.

The formation of classes of similar compounds consists on dividing the series of compounds of interest into categories on the basis of their chemical structure. This may be achieved according to their general backbone, their substituents, reactivity, and knowledge of the biological mechanism. If the subsequent data analysis reveals that the compounds do not form a homogeneous class, new classes should be defined.

2) Quantitative description of structural variation and choice of the QSAR model

To appropriately describe the structural variation, in general, several descriptor variables are required to contain sufficient relevant information about the biological phenomenon. For that reason the structural description is **multivariate**. However, is difficult to predict in advance which descriptor variables will be useful.

Thus, it is convenient to have a set of **independent design variables**, which might have an influence on the biological effect. However, in the optimization of molecules, where substitution patterns or the whole molecular structure is changed, usually is not possible to discern design variables that can be changed independently of each other.

3) Selection of the training set of compounds (series design)

For any QSAR model, it is of crucial importance that the training set selected to calibrate the model exhibits a well-balanced distribution and contains representative compounds. This calibration can be attained by a **systematic selection of the training set**, where the major structural features are varied systematically and simultaneously.

4) Synthesis and Biological testing

Provided that the biological testing should be minimised, the basic idea is to subject merely the representative training test to extensive testing, in order to obtain a broad and stable picture of their biological properties. The response matrix contains biological variables that span as many aspects of the biological profiles of the investigated compounds as possible. The more biological tests are performed, the better the stability of the resulting QSAR model is, and this lead to an improved predictive capability. The testing of a few representative compounds saves time and thus money and adheres to the principles of animal welfare. Biological measurements are commonly recorded as dose-response curves, showing the relationships between the administered doses and the responses that they elicit.

5) QSAR development: data analysis

To calculate the best mathematical expression linking together the physicochemical descriptors and biological responses, information regarding the essential features of the chemical and biological data structure is obtained. There may be need o transform some of the descriptors variables, or delete compounds (**outliers**), exhibiting deviating chemical and/or biological properties. The QSAR analysis also provides information on whether a descriptor variable is relevant for a certain application.

6) Validation and predictions for non-tested compounds

Finally, the final purpose of a QSAR is to predict the biological activities of non-tested compounds, which belong to the class under investigation. However, the **predictive ability** of the model first is verified experimentally. This is accomplished by **biological testing** of some additional compounds in the same way as the training se and then comparing the experimental finding with the values predicted by the QSAR. If the QSAR predicts within acceptable limits, it may be used for a more extensive prognostication. The prediction errors should be compared with the precision and range of the biological measurements obtained.

It is desirable that the compounds in the validation set adequately span the physico-chemical domain and the biological activity range of interest. Conveniently, the validation set may be selected according to a statistical experimental design in order to result in a series of representative compounds.

7) Data analysis and interpretation of results for the proposal of new compounds

In fact, any QSAR development is an iterative cycle, in which the steps are repeated a number of times, until sufficient knowledge about a class of compounds is obtained in order to either design compounds with the desired activity profile or to conclude that such a profile cannot be attained.

4 ORIGINS AND EVOLUTION

4.1 The birth of QSAR

Strictly, the first attempt to correlate properties accounting for chemical and physical behaviour with structures begun about 2500 years ago with **Plato**. Plato described a gasification reaction as a transformation of fire into air in an aqueous solution, assuming the knowledge of the structure of the reactants.

$$2\ \text{Fire} \xrightarrow[\text{solution}]{\text{aqueous}} 1\ \text{Air} \begin{cases} \text{Fire = pyramid} \\ \text{Water = icosahedron} \\ \text{Air = octahedron} \end{cases} \qquad (4)$$

Mendeleev could be also considered as a predecessor of QSAR, as a result of the predictions of new elements and their properties that lead him to the construction of the periodic table of elements in 1869 [46].

But reasonably tracing back the history of QSAR, the first trials to relate a biological response or a physicochemical property of a series of compounds with a structural feature were reported in the nineteenth century. The earlier studies were, of course, qualitative. Quantitative analyses where the bioactivity is mathematically related to a set of parameters using statistical considerations came later on.

In 1863, **Cros** observed that the toxicity of alcohols in mammals increased as the solubility of the alcohols in water decreased [47]. Afterwards, in 1968, **Crum-Brown and Fraser** postulated that there ought to be a relationship between physiological activities and chemical structures [48-49]. They proposed an equation linking changes in biological activity to changes in chemical structure but they did not show a way to characterise chemical structure in quantitative terms. From this basis, **Richardson** expressed the chemical structure as a function of solubility [50].

Opening the field of QSPR, considered to be comprised within QSAR, **Mills** developed in 1884 a QSPR for the prediction of melting and boiling points in homologous series, accurate to better than one degree [51]. Later, **Richet** correlated toxicities of a set of alcohols, ethers and ketones with aqueous solubility [52] establishing the empirical principle “Plus ils sont solubles, moins ils sont toxiques”, that is, « more they are soluble, less they are toxic ».

In the 1900's, **Meyer** from University of Marburg and **Overton** from University of Zurich, working independently, noted that the narcotic potencies of organic compounds depended on their lipophilicity. Basing on biological experiments, they correlated partition coefficients with anaesthetic potencies [53-55]. Besides, Overton also determined the effect of functional groups in the increase or decrease of partition coefficients [56]. Afterwards, **Lazarev** in St. Petersburg continued where Overton and Meyer left off, applying partition coefficients to the development of industrial hygiene standards. Lazarev reported correlations on a log scale, and developed a system for estimating partition coefficients from chemical structure.

But the earliest mathematical formulation is attributed to **Ferguson**, who announced a principle for toxicity. He observed the increase in anaesthetic potency when ascending in a homologous series of either n-alkanes or alkanols to a point where a loss of potency, or at least no further increase occurred, using physical properties such as solubility in water, distribution between phases, capillarity and steam pressure [57].

Little additional development of QSAR occurred until the work of Louis P. Hammett within the organic chemistry field, who is considered to be the father of Linear Free Energy Relationships (LFER). In fact, QSAR methodology as applied nowadays is attributed to the independent contemporary publications of the Free-Wilson model [58] and the Hansch model [59].

4.2 Linear Free-Energy Relationships (LFER)

In the mid-1930s, **Hammett** observed that the addition of substituents to the aromatic ring of benzoic acid had an orderly and quantitative effect on the dissociation constant. He also observed a similar effect on the dissociation of other organic acids and bases [60-61]. From empirical observation, he consequently derived the following linear relationship, the so-called **Hammett equation (1953)**:

$$\log \frac{K}{K_o} = \rho\sigma \tag{5}$$

or

$$\log \frac{k}{k_o} = \rho\sigma \tag{6}$$

where the slope ρ is a proportionality reaction constant pertaining to a given equilibrium that relates the effect of substituents on that equilibrium to the effect on the benzoic acid equilibrium. σ is a parameter that describes the electronic properties of aromatic substituents, i.e. electron-withdrawing or donating power.

These relationships are termed **linear free energy relationships** because they recall the equation relating the free energy, ΔG, to an equilibrium constant, *K*, or rate constant, *k.* In other words, the energetics of the reaction is related to concentration measurements by logarithmic relationships.

$$\Delta G = -RT \ln K \tag{7}$$

The Hammett's correlation to describe the reactivity of aromatic systems was instrumental to the ulterior derivation of QSAR as a discipline by Hansch and his colleagues in the late 1960s. Besides, it was the first study that partitioned the molecule and explained its activity from its fragments, instead of referring to its totality.

Closely related to Hammett’s equation, Taft worked on the steric effects (E_S) and derived the so-called **Taft equation** [62-63]:

$$\log \frac{k}{k_o} = \rho^* \sigma^* + \delta E_S \tag{8}$$

Working in the same direction, Swain studied the effects of field and resonance. He investigated the variation of reactivity of a given electrophilic substrate towards a series of nucleophilic reagents, deriving the linear free-energy relation, called **Swain-Scott equation** [64]:

$$\log \frac{k}{k_o} = sn \tag{9}$$

where *n* is a measure of the nucleophilicity characteristic of the reagent and *s* is a mesure of the sensitivity to the nucleophilicity of the reagent characteristic of the substrate. He also derived the **Swain-Lupton equation,** a dual parameter approach to the correlation analysis of substituent effects, which involves a field constant (*F*) and a resonance constant (*R*). [65]

Free and Wilson partitioned the molecule in a different manner as Hammett. They postulated that the biological activity of a molecular set can be related with the addition of substituents, taking into account the number, type and position in the parent skeleton. Thus, they formulated an additive model, where the activity is discretized as a simple sum of contributions [58]:

$$\log\frac{1}{C}=\sum a_i+\mu \qquad (10)$$

where C is the molar dose, a_i the group contribution of the substituent X_i and μ the biological activity of the parent structure. In this additive model, the descriptors, also called indicator variables, codify the presence or absence of particular structural characteristics. They are assigned the binary values of 1 and 0, accordingly.

But the first application of the Free-Wilson type analysis (properly defined 8 years later) had already been reported in 1956 by **Bruice** et al. [66]. Also, **Bocek and Kopecký** [67-68] developed Free-Wilson models including crossed terms in order to consider the probable interactions among close substituents.

Afterwards, **Fujita and Ban** simplified the Free-Wilson equation estimating the activity for the non-substituted compound of the series [69].

Besides, in the middle 50's, centering on the year 1954, the work from several laboratories came together to offer a quantitative explanation and relationship to a biological activity. The work focused on the carcinogenicity of polycyclic aromatic hydrocarbons. **Daudels, Pullmans, and Coulson** studies used valence bond theory and molecular orbital theory to quantify the involvement of certain bonds in an event initiating the onset of a carcinogenic outcome. From the basis of theoretical structure descriptors, the theory of the *K* and *L* regions illustrating a possible mechanism of the hydrocarbons was derived [70-71].

QSAR based on Hammett's relationships use electronic properties as descriptors of the structure. Difficulties were encountered when investigators attempted to apply Hammett-type relationships to biological systems, indicating that **other structural descriptors** were necessary. To deal with the complexity of biological systems, independently of Free-Wilson analysis, Hansch and Fujita based their model on empirical searches among multiple descriptors and data analysis methods to predict regularities. To such an extent, for the first time, pencil and paper were substituted by computers.

4.3 Hansch Analysis

The origins of QSAR as practiced today began with the research of Robert Muir, a botanist at Pomona College, who was studying the biological activity of plant growth regulators. In attempting to correlate the structures of the compounds with their activities, he consulted his colleague in chemistry, Corwin Hansch [72].

Later on, **Hansch and Fujita** published a LFER related model considered to be the formal beginning for QSAR [73-74]. Their fragment and additive group contribution theory added two things to what had been done before: the use of calculated properties to correlate with biological activities, and the recognition that multiple properties may influence the biological activity. With such purpose, they implemented the use of the computer to fit QSAR equations.

Starting from the hypothesis that substituents on a parent molecule have a quantitive relationship with biological activity, they used Hammett sigma parameters to account for the electronic effect of substituents did not lead to meaningful QSAR. However, Hansch recognised the importance of the lipophilicity, expressed as the octanol-water partition coefficient, on biological activity. This parameter provides a measure of the bioavailability of compounds, which determines, in part, the amount of the compound that gets to the target site.
The so-called Hansch equation [75] was developed to correlate physicochemical properties with biological activities can be expressed in a general form by:

$$\log \frac{1}{C} = a\left(\log P\right)^2 + b\log P + c\sigma + ... + k \qquad (11)$$

where C is the molar concentration that produces the biological effect; P is the octanol/water partition coefficient and σ is the electronic Hammett constant.

The definition of a parabolic model and the combination of different physicochemical properties in one model allowed for the first time the description of SARs which could not be correlated with a single term. As an alternative to logP values, a lipophilicity parameter π can be used, defined in an analogous way as the Hammett's electronic parameter σ. In addition, Rudolf **Zahradnik** formulated Hansch-type relationships relatively early [76-78].

The combination of Hansch and Free-Wilson analysis in a mixed approach widens the applicability of both QSAR methods. SAR are now developed using a variety of parameters as descriptors of the structural properties of molecules.

4.4 Topological Studies

Paralelly to these techniques, but avoiding the use of phsysicohemical parameters, the molecular topology was developed. It considers that the biological activity is related with the molecular topological characteristics, numerically represented by means of the distance and connectivity indices.

The modern formulation of topological studies is due to Wiener [79], Kier and Hall [80-81], Randić [82], Hosoya [83], Balaban and Platt [84], among others. More recently, Rum and Herndon [85-86], have also used similarity matrices derived from bidimensional topological descriptors. Also, the aforementioned TQSI, widely exposed in the second half of **Chapter 3** provide a structure-based insight on the molecular structure characterization.

4.5 Spatial Methods: 3D-QSAR

More recently, the need to include the influence of the conformations and stereochemistry in QSAR studies has opened the three-dimensional QSAR field.
These new techniques, which introduce three-dimensional parameters in the description of compounds, allow calculations extensive to the space surrounding the molecules and require the alignement of the molecules to a common pharmacophore. An application example of these QSAR techniques is the interaction study of a ligand with a receptor, where the molecules are approximated in three dimensions. The interactions are governed by electrostatic and steric factors. This method takes into account the different conformations, stereoisomers, enantiomers, and diastereomers of the studied compounds.

The first approach dealing with electrostatic and steric interactions of molecules with their environment taking into account 3D shape was the Comparative Molecular Field Analysis, CoMFA [87-88]. Nowadays is still a common technique in the modelling of receptor and ligand. Later, Good and Richards [89-91] compared the electronic similarity between molecules, using the CoMFA superposition and then they correlated the topological indices using Neural Networks and Partial Least Squares.

Also, 4D-QSAR [92-93] and 5D-QSAR [94-95] have been developed as representation of ensembles of conformations, allowing for the representation of multiple conformations, orientation and protonation states.

5 MOLECULAR DESCRIPTORS

A common issue in QSAR is how to describe molecules and their properties. The nature of the descriptors used and the extent to which they encode the structural features related to the biological activity is a crucial part of a QSAR study [96]. It has been estimated that more than 3,000 molecular descriptors are now available [97-99]. Most of them can be theoretically calculated by using commercial software packages such as ADAPT [100-101], OASIS [102], CODESSA [103], and DRAGON [99], among others.

From the extensive available bibliography, some of the most widely used in order of increasing complexity are the topostructural, topochemical, geometrical, relativistic, and biodescriptors. The main descriptors used to characterise chemical compounds can be arbitrarily classified in different fashions. The more general overview classes them into three groups:

- **Empirical parameters derived from organic chemistry**, employed in the classical QSAR models, using for example the **Hansch analysis**. Initially, these models were based on several varieties of physicochemical descriptors, classified into electronic, hydrophobic, and steric. But afterwards descriptors of diverse type were also included, i.e. experimental properties like solubility, melting point, boiling point, spectroscopic descriptors, etc.

- **Properties theoretically determined**: this group includes topological descriptors, parameters derived from computational chemistry. The great advantage of theoretical descriptors is that they are calculable for not yet synthesised chemicals.

- More recently, from the eighties, the so-called **tridimensional descriptors** have apperared into scene. These parameters, used in the so-called 3D-QSAR techniques, take into account the three-dimensional structure of molecules and they may require a molecular superposition procedure. This group includes molecular similarity indices and topological quantum similarity indices.

Besides, the influence of a structural characteristics in the activity can be localised to a part of a molecule or it may also be global. This is another usually employed classification pattern of descriptors.

5.1 Substituent constants or parameters based on fragment constants or physicochemical parameters

An important amount of these descriptors belongs to the empirical parameters category derived from physical organic chemistry. These parameters focus on how chemical reaction rates depend on differences in molecular structure. The characterization of these differences in structure, due to different substitutions of functional groups on to a fixed core pattern, leads to the development of substituent constants. These constants relate the effect of substituents on a reaction centre from one type of process to another.

5.1.1 Electronic Substituent Constants

They describe electronic interactions and they are a direct result of the empirical observation of certain chemical systems, where substituents have the same relative effects on the rates of reaction equilibria, regardless of the reaction.

Hammett's constant σ [60]. It is the pionering electronic substituents descriptor, and the subsequently determined constants for substituents in the positions orto, meta and para [104], which define the electronic properties of the aromatic substituents. They can also be of conformational type, for example, the conformational energy of a particular compound.

5.1.2 Hydrophobic Substituent Constants

The hydrophobicity can be experimentally determined for a substituent working within a set of derivatives from the lipophilicity coefficient π, or from the partition coefficients [105-106].

Partition Coefficient (log P). The natural logarithm of the octanol/water partition coefficient is the relative affinity of a compound for an aqueous or lipid medium, closely related with the absorption transport, and partitioning phenomena of a drug in a biological system. The octanol/water partition coefficient is measured by placing the compound in a separatory funnel with octanol and water. Octanol and water are immiscible, and the compound under study partitions between the two phases. The concentration of the compound in the two phases and hence the partition coefficient are a measure of the hydrophobic-hydrophilic character of the compound. The more hydrophobic, the larger are *P* and *log P*. *Log P* is a common descriptor in QSAR studies because drugs must often cross membranes. Cell membranes are composed of phospholipids, which have hydrophobic tails that produce a very hydrophobic environment in the middle of the membrane bilayer. In the absence of active membrane transport, more hydrophobic drugs have an easier time getting through a membrane [104-107].

5.1.3 Steric Substituent Constants

They parametrize the possible interferences with the molecular reactions in homologous series of compounds, by the description of molecular shape and size. These descriptors parameterise, for instance, how well a ligand fits the receptor site, and the nature and relative positions of appropriate functional groups on the ligand, which affect the type and strength of the interaction with complementary groups on the receptor.

Taft steric parameter. The Taft steric parameter, E_S, is the first steric parameter developed by Taft [108], which describes the intramolecular steric effects on the rate of reaction**.**

Charton's steric constant. From the previous wrok, Charton solved some of the limitations of Taft method [109], introducing the so-called Charton's steric constant**.**

STERIMOL parameters. Developed by Verloop et al. [110], they characterize the steric features of substituents in more complex biological systems**.**

5.2 Whole molecule representations

Some of the descriptors derived from entire molecular structures are extensions of the substituent constant approach but many of them are completely new. In addition, several of them are based on the **spatial** conformation of compounds; for that reason, they require a molecular superposition process.

5.2.1 Electronic whole molecule descriptors

These descriptors are derived from a **three-dimensional conformation** of the molecule, and, thus, they are dependent of the modelling program employed. They distinguish from the electronic substituent constants in that a single value is assigned for a given compound. The values range from **experimental** to **semi-empirical** and to **quantum mechanical values** derived from molecular orbital calculations; also some of them are derived from **thermodynamics**. They may encode either general features of the entire molecule or local features of a specific site [111-112]. Electronic descriptors include **polar and energetic descriptors.**

5.2.2 Polar descriptors

They describe the force fields acting on the molecule, and thus they encode the effects or strength of different intermolecular interactions.

Intermolecular forces. They encode the strength of polar-type interactions [115], They can be either experimentally determined or theoretically calculated by using quantum mechanical techniques. The more commonly recognized intermolecular forces arise from the following interactions: ion-ion, ion-dipole, dipole-dipole, dipole-induced dipole, dispersion forces, hydrogen bonding [113]**.**

Molecular polarizability and molar refractivity. They are a measure of a molecule of being polarized. These descriptors are calculated from the refractive index [114] and the molar volume**.**

Ionization constants. They encode ionic interactions and they provide information about the absorption and distribution of a drug [116]**.**

5.2.2.1 Energetic Descriptors

They are obtained by molecular orbital calculations and they mainly describe electronic interaction. Some examples of these descriptors are: **electrostatic potentials, bond order, atomic charges, number of hydrogen bond donors and acceptors, measures of the *π*-*π* donor-acceptor ability of molecules**, and, specially, **reactivity indices**.

Reactivity indices. The energy of the highest occupied molecular orbital, E_{HOMO}, is a quantitative measure for the chemical reactivity of the compound-ionization potential of a molecule. The energy of the lowest unoccupied molecular orbital, E_{LUMO}, accounts for the electron affinity [117]. The HOMO-LUMO band gap energy can also be employed. These magnitudes provide measures of overall susceptibility of a molecule to loose a pair of electrons to an electrophile or to accept an electron pair from a nucleophile**.**

5.2.3 Geometric descriptors

They provide information about the shape and size of active compounds, as well as the degree of complementarity of a ligand and the receptor. They are developed from three-dimensional models of molecules, and derived from molecular surface area calculations.

Molecular volume. It is an overall measure of molecular size, calculated by placing a sphere on each atom with the radius given by the Van der Waals radius of the atom. The most widely used volume estimation technique was developed by Pearlman [118]**.**

Molecular surface area. Molecular surface area can be calculated by using some approximations proposed by Lee and Richards [119], Herman [120], and Pearlman [121]**.**

Charged partial surface area. It provides information on surface area and charge information to understand the properties influenced by interactions in polar molecules [122]**.**

5.2.4 Topological descriptors

These descriptors are based on the connection table for a molecule, a compact representation of connectivity within a molecule. Their values may be or not independent of the three-dimentional conformation.

5.2.4.1 Structure-based descriptors or information-content indices

They count the frequency of occurrence of a substituent or substructures present in molecules as indicator variables: such as **number of bonds** and **number of atoms**.

5.2.4.2 Topological indices

They have been treated in more extension on the second half of **Chapter 3.** Derived from graph representation of chemical structures, they attempt to encode the size, shape, or branching in the compound by manipulation of graph-theoretical aspects of the structures [123]. Among the most important, there are the molecular connectivity indices [81], Wiener index (sum of the chemical bonds existing between all pairs of heavy atoms in the molecule), Zagreb index (sum of the squares of vertex valences), Hosoya index, Kier and Hall molecular connectivity index (a series of numbers designated by order and subgraph type, that emphasize different aspects of atom connectivity within a molecule), Kier & Hall valence-modified connectivity index, Kier & Hall subgraph count index, Kier's shape indices, Kier's alpha-modified shape indices, Molecular flexibility index, and Balaban indices.

5.2.4.3 Electrotopological descriptors

The electrotopological state indices are numerical values computed for each atom in a molecule, which encode information about both the topological environment of the atom and the electronic interactions due to all other atoms in the molecule. The topological relationship is based on the graph distance to each other atom.

5.2.4.4 Kappa indics

The constitute a series of graph theoretical indices developed by Kier, which relate to the molecular shape of the molecule [124-126].

5.2.5 Quantum Similarity Indices

Derived from quantum mechanical calculations, they also take into account the three-dimensional conformation. They can be calculated either as molecular QSI or fragment QSI they have been extensively exposed in the first half of **Chapter 3**.

5.3 Other descriptors

Receptor Surface Analysis (RSA) Descriptors. They calculate the energy of interaction between each point on the receptor surface and each model to the study table.

Molecular Field Analysis (MFA) Descriptors. They evaluate the energy between a probe and a molecular model at a series of points defined by a rectangular or spherical grid.

Molecular Shape Analysis (MSA) Descriptors. Also called pharmacophoric descriptors or 3D Keys. They constitute a collection of combinations of three features (triplets) and four features (quadruplets) in the 3D space for all conformers. The features can be negative and positive charges, negative and positive ionisable groups, hydrogen bond donors and acceptors, hydrophobic groups, aromatic rings, etc.

Absorption-Distribution-Metabolism-Excretion (ADME) Descriptors. They help to the understanding and prediction of drug responses, based on a balance of potency, selectivity, pharmacokinetics, and toxicity profiles required for an ideal drug as well as the minimization of undesired potential side effects. ADME descriptors can be used to predict problematic new chemical entities at an early stage of development, thus reducing drug discovery expenses, minimising the development, and evaluation time for successful candidates, and used to populate or bias libraries with molecules that are likely to yield developable leads.

6 STATISTICAL ADJUSTMENT AND MULTIVARIATE ANALYSIS TECHNIQUES

6.1 Introduction

Statistical methods are the mathematical foundation for the development of QSAR models. Multivariate chemometric methods [31] are applied when it is not easy to extract enough relevant information for the problem from single original variables. The applications of multivariate analysis, data description, classification, and regression modelling, are combined with the ultimate goal of interpretation and prediction of non-evaluated or non-synthesised compounds [127].

Modelling techniques can be roughly classified into two different categories. **Quantitative regression techniques** aim to develop correlation models by using statistical adjustment techniques. Complementary, **qualitative pattern recognition techniques** are devoted to the descriptive data analysis and classification. Besides, methods for the quantitative analysis of continuous properties or methods for the semiquantitative analysis of categorical properties or continuous properties partitioned into discrete classes can be applied, depending on the type of variables being studied.

Among an increasing pool of different modelling methods, the selection of the appropriate method for the statistical analysis is crucial [128]. There is an important number of regression analysis methods available in the literature and **Multiple Linear Regression (MLR),** also termed as **Ordinary Least Squares (OLS)** [129], can be considered as an easy interpretable regression-based method, indicated for QSAR analysis. Some of its variants are simple linear regression, multiple linear regression, and stepwise multiple linear regression.

Recently, such methods have been substituted by multivariate **projection methods**, namely projection to latent variables, such as **Principal Component Regression (PCR) and Partial Least Squares (PLS)** [130], which in turn reduce the information content of data matrices. Thus, these techniques project multivariate data into a space of lower dimensions, reducing obviously the number of dimensions, and indeed providing insight to visualise, classify, and model large sets of data. The position of the observations on the new space is given by the scores and the orientation of the plane in relation to the original variables is indicated by the loadings.

Also, more sophisticated methods, like **Adaptive Least Squares (ALS), Canonical Correlation Analysis (CCA), Continuum Regression (CR),** and the **non-linear Genetic Function Approximation (GFA), and Genetic Partial Least Squares (G/PLS)** have appeared into scene.

Besides, in QSAR field other methods to perform classification studies have been extensively developed. The so-called **Pattern Recognition (PARC)** analysis methods [131-132], i.e. discriminant analysis and decision trees, can be applied to classify compounds in a model. PARC methods are based on a set of classes that contain a number of observations mapped by variables, guidelines, and rules, so that new compounds can be classified as similar or dissimilar to the members of the existing classes. The main assumption to compare the similarity of observations within each class is the application of the principle of analogy. Thus, **pattern recognition techniques** [133-134] such as **Linear Discriminant Analysis (LDA), k-Nearest Neighbours (kNN), Principal Component Analysis (PCA), Correspondence Factor Analysis (CFA) and Factor Analysis (FA), Non-Linear Mapping (NLM), Canonical Correlation Analysis (CCA), Cluster Analysis (CA),** and **Artificial Neural Networks (ANN)** provide qualitative information on the property-structure relationships, by means of representation techniques.

6.2 Overview for the choice of the method

There is no single method that works best for all problems and that has the perfect balance of predictivity, interpretability, and computational efficiency.

Simple linear regression

The simple linear regression method performs a standard linear regression calculation to generate a set of QSAR equations that include one equation for each independent variable. Each equation contains one variable from the descriptor set. This method is suitable for exploring simple relationships between structure and activity.

Multiple linear regression

The Multiple Linear Regression (MLR) [135] is an extension of the classical regression method to more than one dimension. MLR calculates QSAR equations by performing standard multivariable regression calculations using multiple variables in a single equation.

Sepwise multiple linear regression

The stepwise multiple linear regression is a commonly used variant of MLR. In this case, also a multiple-term linear equation is produced, but not all independent variables are used. Each variable is added to the equation at a time and a new regression is performed. The new term is retained only if the equation passes a test for significance. This regression method is especially useful when the number of variables is large and when the key descriptors are not known.

However, if the number of variables exceeds the number of structures, alternative methods such as **projection methods** should be considered.

Principal Components Analysis

The Principal Components Analysis (PCA) method [136] does not create a model but searches for relationships among the independent variables. PCA creates new variables (the principal components) which represent most of the information contained in the independent variables.

Principal Components Regression

The Principal Components Regression (PCR) method uses linear regression to create a model using the principal components as independent variables.

Partial Least Squares

The Partial Least Squares (PLS) regression method [137-138] carries out regression using latent variables from the independent and dependent data that are along their axes of greatest variation and are most highly correlated. PLS can be used with more than one dependent variable. It is typically applied when the independent variables are correlated or the number of independent variables exceeds the number of observations. Under these conditions, it gives a more robust QSAR equation than multiple linear regression.

Genetic function approximation

Genetic Function Approximation (GFA) applies the natural principles of evolution of species that assume that conditions that lead to better results prevail over the poorer ones. So, improvement is obtained by recombination of independent variables, that is, reproduction, mutation, and crossover. This method provides multiple models that are created by evolving the random initial models using the genetic algorithm. The method is good for generating QSAR equations dealing with a large number of descriptors. GFA can build linear and higher-order polynomials, splines, and other nonlinear equations.

Genetic partial least squares

Genetic Partial Least Squares (G/PLS) is a variation of GFA that is derived from two methods: GFA and PLS. Both GFA and PLS are valuable analytical tools for datasets that have more descriptors than samples.

Simple and multiple linear regression are very quick and easy to interpret, but do not work when the number of independent variables is larger than or even comparable to the number of molecules.

Stepwise multiple linear regression and genetic function approximation work with any number of variables but do not perform well if important information is spread over more of them than can be included in the model.

Partial least squares can handle any number of independent variables, but creates only linear relationships. Genetic partial least squares offers automatic creation of nonlinear terms, combining the best features of GFA and PLS.

The advantage of PLS over PCR is that when the dependent variable information is used directly when regressing, the model very often turns out to be better for prediction purposes, and handles noisy data better than PCR.

In summary, MLR is considered a reverse classical regression method placing all weight on the dependent variable when regressing; this means that the prediction error is minimised. On the other hand, PCR is considered a forward inverse regression method placing all weight on the independent variable, hence minimising the calibration error. PLS uses both variables equally. Depending on the nature of the data and the intended use of the model each of the methods has qualities that must be considered in each case.

6.3 Pre-treatment of data

Before applying some multivariate analysis methods, and for the sake of quality of results, a previous treatment of the data is required. Depending on the method to be used and the amount of data available, the data set needs to be transformed by means of pre-processing methods in order to enhance the information.

6.3.1 Scaling and centring

The results of projection methods depend on the normalisation of the data. Descriptors with small absolute values have a small contribution to overall variance; this biases towards other descriptors with higher values, and leads to biased principal components. With appropriate scaling, equal weights are assigned to each descriptor, so that the more important variables in the model can be focussed.

In order to give all variables the same importance, they are standardised by **autoscaling.** The standard procedure consists of normalising each variable to mean centring and variance scaling. Centring sets the mean value to zero, that is, moves the centre of the entire the data set to origin by subtracting the mean value from each variable. This gives all variables the same distance from origin. Scaling sets the variance to one for all variables. This means that all variables get equal importance or weight in the model.

$$\mathbf{x}_i' = \frac{\mathbf{x}_i - \bar{x}_i \mathbf{1}}{s_{\mathbf{x}_i}} \tag{12}$$

$\mathbf{x}_i$ is the original column vector of descriptors, $\mathbf{x}_i = \left(x_{1i}, x_{2i}, \ldots x_{ni}\right)^T$, $\bar{x}_i$ is the arithmetic mean, $s_{\mathbf{x}_i}$ is the standard deviation, and $\mathbf{x}_i'$ the transformed vector of autoscaled variables.

These transformations are recommended for ease of interpretation and numerical stability, but do not lead to changes in the coefficients or weights of variables and does not alter the interpretation of the results.

6.3.2 Data correction and compression

If the data have skew distributions or systematic error, specially when handling with spectroscopic data, they may need to be **linearized**.

6.3.3 Other transformations

Furthermore, a variable may contain one or a few extreme measurements that may influence the model. If the variables differ in range, variation, and size, they are often transformed, so that their distribution is consistent with chemical and biological theory. Activity variables with a range covering more than one order of magnitude of ten can be logarithmically transformed, and the same applies to structure descriptor variables. By transforming them to a frequency, or a scale domain, important information can be more easily extracted. Another example of such transforms is the Fourier transform.

6.4 Construction of the predictive model: Linear Regression Analysis

Regression models help to build models, estimate their predictive ability, and find underlying relationships between descriptors. The first part of data analysis consists of using the data to determine values of parameters in the model so that the model fits the data well. The most general purpose of regression techniques is to construct a model to represent the dependence of several independent predictor or observable variables on dependent criterion or explicative variables. This relationship is expressed in the form of an equation that expresses dependent variables, **y**, in terms of independent variables, **X**. Therefore, MLR predicts **y** from the knowledge of **X** variables, which can be either quantitative or qualitative. Also, **X** parameters may be the selected principal axis resulting from a MDS configuration from a reduction dimensionality process.

In particular, in this thesis, for the analysis applied to a QSAR study, the main goal is to correlate a biological property, forming a column vector (**y**), with molecular descriptors, arranged in the columns of the so-called data matrix (**X**). The columns are associated with variables or descriptors, whereas the objects, in this case, the molecules, are associated with the rows. The model can then be used to predict activities for new molecules, or screening a large group of molecules with unknown activities among other uses. Usually, the prediction model is elaborated using the parameters calculated for a well-determined data of a training set on the unknown test set. If the training set is a sufficiently representative pattern of the system, then, it can be assumed that the introduction of new elements with an unknown property will not affect their stability and that confident predictions can be attempted.

A model's ability to provide insight into the system is as important as its predictive ability. Possibly more valuable than being able to predict an activity or property is to get insight into underlying relationships between descriptors. Finally, validation methods are needed to establish the predictive capacity of a model on test data and to help determine the complexity of an equation that the amount of data justifies.

Regression methods can be classified according to the parameters being optimized. Linear regression methods fit parameters to data so as to **minimise the sum-of-squared residual errors**. Some of them have the side effect of minimising or maximizing other quantities at the same time.

6.4.1 Simple Linear Regression Analysis

In the simplest case, i.e. the simple linear regression, there is a single independent **x** variable, and a dependent **y** variable. Thus, a one-term linear equation is produced separately for each independent variable. This is useful for discovering some of the most important descriptors; however, it ignores the interaction of multiple descriptors. The goal of a linear regression procedure is to fit a regression x/y line through the points (observations) so that the squared deviations of the observed points from that line are minimised.

The regression equation in a two-variable or two dimensional space is defined by the equation:

$$\mathbf{y} = a + b\mathbf{x} \tag{13}$$

where the response variable **y** can be expressed in terms of the variable **x** by means of two parameters: the constant *a* and the slope *b*. The constant is also referred to as the intercept, and the slope as the estimator, the **regression coefficient** or *b* coefficient.

The coefficients {*a*,*b*} are usually determined by a **least squares estimation**. Once determined the $\{a,b\}$ coefficients the experimental property can be estimated:

$$y'_i = a + bx_i \tag{14}$$

where *y'* is the adjusted value of *y*.

The regression line expresses the best prediction of the dependent variable, $\hat{\mathbf{y}}$, given the independent variable **x**. However, there is usually a substantial variation of the observed points around the fitted regression line. The deviation of a particular point from the regression line, that is, the distance of the experimental value ***y*** from the adjusted value ***y'***, is called quantified by the **residual error** ($\mathbf{e} = \mathbf{y} - \mathbf{y}'$). The smaller the variability of the residual values around the regression line relative to the overall variability, the better is the prediction. Customarily, the degree to which the predictor (**x**) is related to the dependent (**y**) variable is expressed by the **correlation coefficient r**, whose values are comprised between 0 and 1.

To estimate *a* and *b* regression coefficients, the least squares method finds the values that minimize the sum of squared residuals: $\min\|\mathbf{e}\|^2$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_i')^2 = \sum_{i=1}^{n} \left[y_i - (a + x_i b) \right]^2 \tag{15}$$

From successive demonstrations, it follows that the expressions for the slope and the intercept are, respectively:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \tag{16}$$

$$a = \frac{\left(\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i\right)}{n} \tag{17}$$

or

$$a = \overline{y} - b\overline{x} \tag{18}$$

being $\overline{x} = \frac{1}{n}\sum_{i}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i}^{n} y_i$ the arithmetic means (average values) of **x** and **y**.

The sign of *b* coefficients is an indicator of the relationship between variables. If *b* is positive, then the relationship of this variable with the dependent variable is positive; if *b* is negative so the relationship is, and if *b* is equal to zero there is no relationship between the variables.

A fundamental principle of least squares methods is that the variance of the dependent variable can be partitioned according to the source. Thus, it can be demonstrated that the total sum of squared values of the dependent variable equals the sum of squared predicted values plus the sum of squared residual values. Stated more generally,

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n}\left(y_i^{'} - \overline{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2 \tag{19}$$

Taking the following definitions,

$$SS_T = \sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 \tag{20}$$

$$SS_R = \sum_{i=1}^{n} \left(y_i^{'} - \overline{y} \right)^2 \tag{21}$$

$$SS_E = \sum_{i=1}^{n} \left(y_i - y_i^{'} \right)^2 = \sum_{i=1}^{n} e_i^2 \tag{22}$$

the balance of variances can be stated in another way:

$$SS_T = SS_R + SS_E \tag{23}$$

The term *SS*$_T$ on the left is the total sum of squared deviations of the observed values of the dependent variable from the dependent variable mean. The terms on the right are *SS*$_R$ and *SS*$_E$. *SS*$_R$ is the sum of squared deviations explained by the regression model, of the predicted values from the dependent variable mean, and *SS*$_E$ is the residual sum of the squared deviations of error, of the observed values from the predicted ones, that is, the sum of the squared residuals. The standard error of the estimate is a measure of the accuracy of predictions. As the difference between observed and predicted approaches zero ($SS_E \simeq 0$), the sum of the squares due to regression approaches the sum of the squares about the mean ($SS_R \simeq SS_T$). Thus, the sum of the squares of the residuals (*SS*$_E$) can be considered a measure of goodness of fit. *SS*$_T$ is always the same for any particular data set, but *SS*$_R$ and the *SS*$_E$ vary with the regression equation.

6.4.2 Multiple Linear Regression Analysis

In the multivariate case, i.e. MLR analysis, when there is more than one independent variable, the regression cannot be visualized as a line in the two-dimensional space, but it can be computed just as easily.

The regression model also assumes a linear relationship between *m* molecular descriptors and the response variable. This relationship can be expressed with the single multiple-term linear equation:

$$\mathbf{y} = b_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + ... + b_m\mathbf{x}_m + \mathbf{e} \tag{24}$$

The MLR analysis estimates the **regression coefficients,** $\{b_i\}$, by minimising the residuals, **e**, which quantify the deviations between the data (**y**) and the model (**y'**), as in the case of simple linear regression.

The multiple regression model can be also expressed compactly using matrix notation [139-140].

$$\mathbf{y = Xb + e} \tag{25}$$

In this case, the vector containing the activity value, $\mathbf{y}(n)$, experimentally calculated for n objects, i.e. chemicals, is $\mathbf{y} = (y_1, y_2, ..., y_n)^T$. The matrix of descriptors calculated for the n objects and m associated descriptors, $\mathbf{X}(n \times m)$, is $\mathbf{X} = \{\mathbf{x_1, x_2, ..., x_m}\}$. **b** is the column vector of regression coefficients $\mathbf{b} = (b_0, b_1, b_2, ..., b_m)^{\mathrm{T}}$, and **e** is the error vector. The first column of **b** is composed by constants (b_0); this column becomes zero after the centring. Regression coefficients represent the independent contributions of each independent variable to the prediction of the dependent variable.

y_i, x_{ij}, and e_i respectively represent the activity, the j-th descriptor, and the residual value between the experimental value and the adjusted one, for a compound i.

Analogously as before, to estimate **a** and **b** regression coefficients, the least squares method also minimises the sum of the squares of the residuals: $\min \|\mathbf{e}\|^2$, taking into account that, in turn, the residuals are defined from the adjusted values ($\mathbf{e = y - y' = y - Xb}$):

$$\begin{aligned} \sum_{i=1}^{n} \mathbf{e_i^2} &= \mathbf{e^T e} = \mathbf{(y - Xb)^T (y - Xb)} = \\ &= \mathbf{y^T y - b^T X^T y - y^T Xb + b^T X^T Xb} = \\ &= \ \mathbf{y^T y} - 2\mathbf{b^T X^T y + b^T X^T Xb} \end{aligned} \tag{26}$$

By applying the minimisation condition, the first partial derivative must be equal to zero:

$$\begin{aligned} \left.\frac{\partial S}{\partial \mathbf{b}}\right|_{\mathbf{b}} &= -\mathbf{2X^T y + 2X^T Xb = 0} \\ \mathbf{X^T Xb} &= \mathbf{X^T y} \end{aligned} \tag{27}$$

One of the ways for solving the system above is to premultiply both sides of the matrix formula by the inverse matrix $\mathbf{X^T X}$ to give

$$\mathbf{b} = \left[\mathbf{X^T X}\right]^{-1} \mathbf{X^T y} \tag{28}$$

The condition for existence of the least squares solution is that the inverse product of matrices, $\mathbf{X^T X}$**,** exists, so that it is non-singular. This, in turn, requires that **X** is non-singular, i.e. the columns of the **X** matrix are linearly independent, in order that no column of **X** may be written as a linear combination of other columns of **X**. In this case, the system of equations is of full rank, ($n > m$), and there is a unique solution.

When the number of descriptors m is close to or greater then the number of molecules n, ($n \approx m$ or $n < m$), **b** cannot be estimated by multilinear regression because **X** is not full-rank, the $\left|\mathbf{X^T X}\right|$ determinant is null, and $\left(\mathbf{X^T X}\right)^{\mathbf{-1}}$ is not defined. In this case, there is a linear function relating predictor variables and so the descriptors are linearly dependent. The problem of collinearity [135] can be solved by suppressing descriptors or by using other Multivariate methods.

Using the calculated values of the correlation coefficient, the values of the experimental property are estimated. The vector with experimental properties can be related with the adjusted properties fitted by the linear model, $\mathbf{y'} = \left(y_1^{'}, y_2^{'}, ..., y_n^{'}\right)^T$, by means of the expression:

$$\begin{aligned} \mathbf{y'} &= \mathbf{Xb} = \mathbf{X}\left[\mathbf{X^T X}\right]^{\mathbf{-1}} \mathbf{X^T y} \\ &= \mathbf{Hy} \end{aligned} \tag{29}$$

The prediction matrix or **hat matrix** of $(n \times n)$ dimension that relates the observed values with the adjusted ones has been defined as: $\mathbf{H} = \left\{h_{ij}\right\} = \mathbf{X}\left[\mathbf{X^T X}\right]^{\mathbf{-1}} \mathbf{X^T}$, that is: $h_{ij} = \mathbf{x}_i^T \left[\mathbf{X^T X}\right]^{\mathbf{-1}} \mathbf{x}_j$, where $\mathbf{x}_i^T = \left(x_{i1}, x_{i2}, ...x_{ik}\right)$ is the *i*-th row of **X** matrix.

6.5 Conditions

In the usual models, the calculations are based on three simplifying assumptions: independence of observations, normality of the sampling distribution, and uniformity of residuals.

6.5.1 Independent variables

Independence of observations refers to the notion that the value of one datum is unrelated to any other datum. Observations that are not independent are also said to be correlated or interdependent. MLR assumes the predictor variables **X** to be linearly independent or orthogonal. So the method fails when the descriptors are correlated or collinear. The method also requires at least as many molecules as independent variables. However, to produce reliable results minimising collinearities and the possibility of chance correlations, typically the ratio of compounds to variables should be at least five to one [141]. When the number of independent variables is greater than the number of molecules, multiple linear regression cannot be applied.

6.5.2 Linear relationship

The dependence of **y** on **X** to construct the QSAR model is assumed to be linear, while the real world is always non-linear. Thus, it is not easy to obtain the exact functional relationship $\mathbf{y} = f(\mathbf{X})$, because there is the influence of the random residual **e**, acting as a perturbating term.

6.5.3 Normal distribution

If the study was repeated many times, the expected sampling distribution for the **y** values of the statistic is assumed to follow normal distribution. This means that the typical error has the same average magnitude for every subject and that the condition of homodedasticity holds.

The residuals are uniform if their mean is zero and their scatter or standard deviation is the same for any subgroup of observations. If there are non-uniform residuals, the heteroscedasticity condition prevails. In this case, there are two possible solutions: to fit a non-linear model instead of a straight line, or transform the data to obtain a straight line. The most common transformations are log transformation and rank transformation.

6.5.4 Other

The available data of a given system consists of uninteresting information, termed as noise, and the data of interest. Noise is partly random, but it can also be systematic, due to inadequacies in structure descriptors, as well to deficiencies in the model. If the data set contains a lot of noise, MLR model can fail.

If there is a low degree of coherence between **X** and **y**, MLR can also fail. The lack of selectivity, i.e. all **X** information is used in the regression, irrespective of its correlation with **y**, is also another drawback.

6.6 Reduction of Dimensions: Multivariate Data Analysis Methods

Usually, molecular descriptor matrices can not be directly used as independent variables in the MLR analysis due to their lack of homogeneity, the high correlation between descriptors, or the excessive number of involved descriptors. If the number of descriptors is larger than the number of compounds, some of them may be redundant. Thus, previous to the MLR analysis, normally a reduction of variables is necessary in order to obtain a concentrated set of significant underlying variables, not correlated between them, loosing the minimum amount of information.

The goal of many multivariate analysis methods is to find a mathematical function of the multivariate data to define a relatively small number of new **latent variables** possessing the maximum amount of information relevant for the problem. The latent variables, obtained as statistical scores, are also called **factors**, **components**, **coordinates**, or **principal properties**. They are orthogonal and can, thus, be used in multiple linear regression.

Graphically, multidimensional data are transformed, and projected into a more intuitive **space of lower dimensionality,** with a **minimal loss of information**. These transformations suppress the dimensions contributing with an insignificant percentage of information. Thus, the basic features behind the data are extracted and visualised into a pictorical form, with the aim of interpretation. In this way, the dimensionality of complex multivariate data is effectively reduced.

Several multivariate analysis methods try to explain an extended set of variables by means of a reduced number of new latent variables. These methods can be grouped into **linear**, i.e. Principal Component Analysis (PCA), Principal Component Regression (PCR), Linear Discriminant Analysis (LDA), Partial Least Squares (PLS), and Factor Analysis (FA), and **non-linear** methods, i.e. Quadratic PLS (QPLS), Non-Linear Mapping (NLM), Artificial Neural Networks (ANN). Besides, the techniques can be also described by the criterion optimized in each particular case (i.e. variance, covariance, correlation, discriminating power…). The objective is to represent the multivariate data by a minimum number of latent variables. The extracted latent variables are orthogonal and can, thus, be used in Multiple Linear Regression (MLR) [135].

In this thesis, only some widely used statistical techniques for multivariate analysis used in QSAR studies such as **Principal Component Analysis (PCA)** [142], **Principal Coordinates Regression (PCR)** [143-144], and **Partial Least Squares Regression (PLS)** [130,137] will be presented in detail.

Given a set of original variables, $\mathbf{X}_{(n\times m)}$, the main goal is to find *k* new latent variables or **scores**, $\mathbf{T}_{(n\times k)}$, where $k \le m$, and to determine the contribution of these new variables to the original ones by means of the coefficient matrix or matrix of **loadings,** $\mathbf{P}_{(m\times k)}$. The decomposition of the variables matrix corresponds to the matricial equation:

$$\mathbf{X} = \mathbf{TP}^{\mathbf{T}} + \mathbf{E} \tag{30}$$

where **E** is the **residual error** matrix.

The main objective is to determine the coefficient matrix **P**. Once decomposed the original variables matrix, the MLR equations can be built between the experimental properties **y** and the new transformed variables **T**.

In particular, **Multidimensional Scaling Techniques (MDS)** [145] seem to be a suitable method to treat the similarity or dissimilarity data, as proximities. The reduction dimensionality methods for similarity matrices try to map the original proximities $\left\{p_{ij}\right\}$ into distances d_{ij} of a multidimensional space. These distances are expressed as a function of a configuration matrix **X**. [146]. The mapping between proximities and distances is specified by means of a function of representation:

$$f : p_{ij} \to d_{ij}\left(\mathbf{X}\right) \tag{31}$$

or more generally:

$$f : g\left(p_{ij}\right) \to d_{ij}\left(\mathbf{X}\right) \tag{32}$$

where p_{ij} is the quantitative measure of the proximity between two objects and $g\left(p_{ij}\right)$ represents a transformation of the original proximities. In this particular case, p_{ij} would correspond to the Z_{ij} elements of the similarity matrix **Z,** and $g\left(p_{ij}\right)$ could be the elements of a similarity indices matrix.

In practice, it is not possible to find an exact function, but a configuration **X** that approximates the distances between points to the original proximities. The choice of the parameter defining the error measure yields different types of reduction techniques.

Each point in the multidimensional space of configuration **X** is described by:

$$\mathbf{x}_i = \left(x_{i1}, x_{i2}, \ldots, x_{im}\right) \tag{33}$$

where x_{ia} is the projection of the object i in the axis a. This vector is the coordinates vector of the object i. The origin of coordinates is assumed to be $\mathbf{0} = \left(0, 0, \ldots, 0\right)$.

Mathematically, the distance between two points i and j that belong to a **Euclidian space** corresponds to the length of the segment of their connecting line

$$d_{ij}\left(\mathbf{X}\right) = \left[\sum_{a=1}^{m}\left(x_{ia} - x_{ja}\right)^2\right]^{1/2} \tag{34}$$

This distance depends on the **X** configuration. This equation can be rewritten in terms of vectors as:

$$d_{ij}^2\left(\mathbf{X}\right) = \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \left(\mathbf{x}_i - \mathbf{x}_j\right) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2 \tag{35}$$

The matrix of Euclidian distances satisfies the following constraints:

$$\begin{aligned} \delta_{ij} &\geq 0 \quad \forall i, j \\ \delta_{ii} &= 0 \quad \forall i \\ \delta_{ij} + \delta_{ik} &\geq \delta_{jk} \quad \forall i, j, k \end{aligned} \tag{36}$$

6.6.1 Classical Scaling

Classical Scaling [143-144] was one of the first multivariate analysis tools used in the treatment of similarity matrices. This technique considers the objects as points in a multidimensional Euclidean space and finds the coordinates for these points so that the distances between them fit the original similarities.

To develop the classical scaling formalism, it is usually supposed that the **T** coordinates matrix is known for a set of n points in a Euclidian space. The squared distance matrix can be constructed from the **T** coordinates matrix:

$$\mathbf{D}^{(2)} = \mathbf{c1}^{\mathbf{T}} + \mathbf{1c}^{\mathbf{T}} - 2\mathbf{TT}^{\mathbf{T}} = \mathbf{c1}^{\mathbf{T}} + \mathbf{1c}^{\mathbf{T}} - 2\mathbf{B} \tag{37}$$

where $\mathbf{B} = \mathbf{TT}^{\mathbf{T}}$ is the so-called scalar product matrix, **1** is a *n*-dimensional unitary vector and **c** is a vector with the diagonal elements of **B** as components.

The solution provided by classical scaling tries to identify the primitive proximities with distances between points in a multidimensional space. The solutions of the method are invariant in relation to translations of the data set. The coordinate origin usually is selected to coincide with the centre of gravity of the points.

There are two ways of treating the proximity matrices of quantum objects. The first one is to use a dissimilarity quantum similarity index (D class) and, afterwards, find the **T** matrix by means of spectral decomposition. The second consists of using the Gower transformation from the Carbó indices and transform the elements of the distance matrix. Both transformations, as well as the direct use of original similarities, do not influence the final result of the projection; that is, the quality of the model remains unchanged.

Taking the transformed, squared and centred, quantum similarity matrix **Z** as **D**, the **B** matrix can be easily obtained. Afterwards, the **T** matrix is obtained from **B** by means of the **spectral decomposition**:

$$\mathbf{B} = \mathbf{TT}^{\mathbf{T}} = \mathbf{V\Delta V}^{\mathbf{T}} \tag{38}$$

$$\mathbf{T} = \mathbf{V\Delta}^{1/2} \tag{39}$$

where **V** is the matrix of eigenvectors, and **Δ** is a diagonal matrix of eigenvalues of **B**. **B** is symmetric and, hence, its eigenvalues are always real and the associated eigenvectors can be taken as orthogonal [147]. The eigenvalues are ordered in descendent order, according to the magnitude of the associated eigenvalue.

In classical scaling, the dimensions are nested, that is, the first and second principal coordinates of the whole solution coincide with the bidimensional solution. Thus, once defined the similarity matrix, classical scaling only needs to be calculated once, independently of the dimensionality of the prediction model. Hence, out of the total number of the generated variables set, it is possible to use only a selected subset as a source of QSAR parameters.

The classical scaling technique considers similarity matrices as distance matrices and, thus, it neglects the diagonal elements, considered as nulls. But, in fact, the self-similarities that occupy the diagonal of the quantum similarity matrices are non-null elements, different from each other. To avoid the loss of information derived from this assumption, original matrices can be transformed into similarity indices that lead to matrices with uniform diagonal elements, without relevant information.

The transformation of a normalized similarity matrix element z_{ij}, where $0 \leq z_{ij} \leq 1$, and $z_{ii} = 0$, into a Euclidean distance matrix element can be done using the **Gower and Legendre transformation** [148]:

$$\delta_{ij} = \sqrt{1 - z_{ij}} \tag{40}$$

It has been demonstrated that when transforming a Euclidean distance matrix by classical scaling, there always exists an exact solution where the distances between the points in the multidimensional space exactly fit the original distances [143,149].

The goodness of fit of the model can be quantified by means of **scree plots** [150]. The eigenvalues associated to the eigenvectors of **B** matrix define a proportion of the variation between the points, $V^{(m)}$:

$$V^{(m)} = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{41}$$

The numerator calculates an extended sum for all the selected subspace, while the denominator takes the complete solution, all the eigenvectors.

The scree plot plots the values of $V^{(m)}$ function versus the dimensionality (number of components) of the space. As the number of dimensions increases, the configuration approaches to the ideal one, and the $V^{(m)}$ function approaches the unit. These curve representations present often a bow: they increase monotonally, until a given dimension; from that point, the increase rate is lower. This limit indicates that the addition of new dimensions to the configuration does not lead to a better description. Thus, this point is an adequate **selection criterion** to deduce the number of dimensions for a classical scaling configuration.

Another way to define the loss of information function is by means of the **stress function** [151-152], defined as:

$$Stress = \sum_{i,j} \left[p_{ij} - d_{ij}\left(\mathbf{X}\right) \right]^2 \qquad (42)$$

The main application of Quantum Similarity in this field consists of describing a set of external data, i.e. molecular properties, with the principal coordinates of the system [153-154].

6.6.2 Principal Components Analysis

Principal Component Analysis (PCA) is one of the oldest and most widely used data reduction techniques to reduce the dimensionality of a multivariate data set of descriptors **X** [136, 155-157]. It seeks to determine a new set of variables -referred to as **Principal Components (PC)-** describing the data in order of decreasing variance with the purpose to express the main information in the variables by a lower number of variables, called the principal components of **X.** Equivalently, PCA can be described as a method to determine the natural dimensionality of the dataset allowing subsequent embedding of the data into a space of lower dimensionality within a margin of prescribed original variance percentage.

The first axis, the so-called **first Principal Component (PC1**), describes the maximum variation in the whole data set; alternatively, it can be also pictured as the direction of greatest variance. The second PC describes the maximum remaining variance, and so forth, with each axis orthogonal, that is, linearly independent, to the preceding axis. Some of the last components may be discarded to reduce the size of the model and avoid over-fitting. In effect, this method reduces the size of the model to the amount of data available.

The PC technique is sensitive to scale changes; thus, **X** columns must be previously centred and scaled.

PCA method aims to extract the maximum amount of variance of the initial variables. To such an extent, the original descriptors are described by means of linear combinations of principal components or scores, **T**:

$$\mathbf{X} = \mathbf{TP^T} \rightarrow \mathbf{X} = \mathbf{t_1 p_1^T} + \mathbf{t_2 p_2^T} + \mathbf{...} + \mathbf{t_k p_k^T} + \mathbf{e} \qquad (43)$$

Mathematically, the number of PCs which can be extracted from a data matrix is usually equal to *m*, the number of original variables. With this number of components, the data matrix can be exactly reproduced. However, this is not a desired result, since it would not lead to a reduction of the dimensionality of the data space. The goal is to find the minimum number of components, *k*, such that in the space they span, the original variables can be represented without loss of relevant information. These components reflect the basic effects behind the data. The non-selected components are assumed to represent irrelevant or residual information comprising errors of measurement and errors in the model.

Thus, the objective of PCA is the decomposition of $\mathbf{X}_{(n\times m)}$ in *k* **principal components** or factor score vectors $\mathbf{T}_{(n\times k)}$ -characteristic of the features of the objects - and *k* **loading** vectors $\mathbf{P}_{(m\times k)}$ -characteristic of the measurements-, where $k<m$.

The PCA method studies the decomposition of the covariance matrix between the predictor variables. The starting point is to evaluate the principal components of the covariance matrix R:

$$\mathbf{R} = \mathbf{X}^{\mathbf{T}}\mathbf{X} \tag{44}$$

This equation can be solved by diagonalising **R** using the standard **Single Value Decomposition (SVD)** procedure. Scores and loadings are obtained from the resulting eigenvectors and eigenvalues, where eigenvalues represent the variance contribution of the components in decreasing order, and the eigenvectors are the PC.

To carry out the SVD factorization of **R**, it must be assumed that **X** is a centred, quadratic and orthonormal, i.e. orthogonal and normalized, data matrix. Then, it follows that $\mathbf{X}^{\mathbf{T}}$ is also orthonormal. For such a matrix it can be demonstrated that: $\mathbf{X}^{\mathbf{T}}\mathbf{X} = \mathbf{X}\mathbf{X}^{\mathbf{T}} = \mathbf{I}$, so that $\mathbf{X}^{\mathbf{T}} = \mathbf{X}^{\mathbf{-1}}$.

The SVD of **X** may be written as:

$$\mathbf{X} = \mathbf{TDP}^{\mathbf{T}} \tag{45}$$

where **T** is an $(n\times n)$ orthonormal matrix, **P** is a $(m\times m)$ orthonormal matrix, and **D** is a $(n\times m)$ diagonal matrix. The non-negative diagonal elements $d_1,\ldots,d_k$ of **D** matrix, where $k=\min\{n,m\}$, are called the singular values.

Appropriately applying the rules for transposing a matrix product,

$$\begin{aligned}\mathbf{R} &= \mathbf{X}^{\mathbf{T}}\mathbf{X}\\ &= \mathbf{P}\mathbf{D}^{\mathbf{T}}\mathbf{T}^{\mathbf{T}}\mathbf{T}\mathbf{D}\mathbf{P}^{\mathbf{T}}\end{aligned} \tag{46}$$

Since **T** is orthogonal, $\mathbf{T}^{\mathbf{T}}\mathbf{T} = \mathbf{I}$, and so $\mathbf{D}^{\mathbf{T}}\mathbf{T}^{\mathbf{T}}\mathbf{T}\mathbf{D} = \mathbf{D}^{\mathbf{T}}\mathbf{D} = \boldsymbol{\Delta}$, where Δ is a diagonal matrix with non-negative values.

Then, the decomposition of **R** from the SVD of **X** can be expressed as

$$\mathbf{R} = \mathbf{P}\boldsymbol{\Delta}\mathbf{P}^{\mathbf{T}} \tag{47}$$

where $\Delta = diag\left\{\lambda_1,...,\lambda_k\right\}$ contains the eigenvalues of $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ in its diagonal. It can be assumed that the singular values are in decreasing order so that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \geq 0$.

By using that **P** is orthonormal, it is also possible to write

$$\mathbf{T} = \mathbf{X}\mathbf{P} \tag{48}$$

There are two cases to be considered in the formation of Δ:

1) When $m \leq n$, $\mathbf{D}^{\mathbf{T}}\mathbf{D}$ is a $\left(n \times m\right)$ diagonal matrix with diagonal elements

$$\lambda_m = d_m^2 \quad \forall m = 1,...,n \tag{49}$$

If the singular values of **X** are all strictly positive, then so are the eigenvalues of **R** and **R** is **positive-definite**.

2) When $m > n$, $\mathbf{D}^{\mathbf{T}}\mathbf{D}$ is a $\left(m \times m\right)$ diagonal matrix with diagonal elements

$$\begin{aligned}\lambda_m &= d_m^2 \quad \forall m = 1,...,n\\ \lambda_m &= 0 \qquad \forall m = n+1,...,m\end{aligned} \tag{50}$$

Even if the singular values of **X** are all non-negative, there are still $m-n$ null eigenvalues of **R** and so **R** is **positive semi-definite**. This is a result of **R** being singular, that is, the set of *m* vectors can not be linearly independent.

3) If the eigenvalues are both positive and negative, then **R** is called **indefinite**.

In the particular case where **X** is a $(n \times n)$ square symmetric matrix, a number λ such that there exists a non-zero vector **p** so that $\mathbf{Rp} = \lambda\mathbf{p}$ is called an eigenvalue of **R,** and **p** is called the corresponding **standardised eigenvector**. Any scalar multiple of **p** is then also an eigenvector for the same eigenvalues. Then the set of equations $\mathbf{Rp}_i = \lambda_i\mathbf{p}_i \; i = 1,....,n$ may be written in matrix form as follows: $\mathbf{RP} = \mathbf{P\Delta}$ where $\mathbf{\Delta} = diag\{\lambda_1,...,\lambda_n\}$

Since **P** is orthonormal, $\mathbf{P^T} = \mathbf{P^{-1}}$ so by right-multiplying the equation by $\mathbf{P^T}$: $\mathbf{R} = \mathbf{P\Delta P^T}$ or

$$R = \sum_{i=1}^{n} \lambda_i \mathbf{p}_i \mathbf{p}_i^{\mathbf{T}} \tag{51}$$

If R is positive semi-definite, the so-called eigenvalue or spectral decomposition of R, $\mathbf{R} = \mathbf{P\Delta P^T}$ leads to the same result as the SVD. In that case, the matrices of the SVD for **R** are:

$$\begin{aligned} \mathbf{T} &= \mathbf{P} \\ \mathbf{D} &= \mathbf{\Delta} \end{aligned} \tag{52}$$

Mathematically, each extracted PC (score) is orthogonal to the previously generated PCs ($\mathbf{t}_i^{\mathbf{T}}\mathbf{t}_j = 0$ m for $i \neq j$) and describes a decreasing percentage of the variance of the original **X** matrix.

Similarly, the loading vectors are also orthonormal, that is, orthogonal ($\mathbf{p}_i^{\mathbf{T}}\mathbf{p}_j = 0$) and scaled to the unity ($\mathbf{P^TP} = \mathbf{I}_k$), i.e.; normalized. The components of the eigenvectors are the weights of the original variables in the PCs, and form the columns of the orthonormal **P** matrix.

The eigenvalues represent the percentage of variance explained for each associated PC. But there is not a unique solution of **P** to verify the previous equation. Thus, an additional condition is introduced, whereby the components are determined in sequence, in such a way that the first component accounts for the largest amount of correlation in **R,** the second component for the next largest amount of correlation**,** and so on. Thus, the PCA analysis chooses the first principal component that explains the maximum amount of variance o the variables.

The factor scores and loadings can be obtained in many different ways. NIPALS algorithm was developed in 1923 [158], later modified in 1966 [159-160], and SIMPLS algorithm [161] resulted from work by de Jong in 1993. Singular value decomposition is another commonly used method for calculating scores and loading [162].

In practice, in the PCA analysis the **NIPALS (Nonlinear Iterative Partial Least Squares)** algorithm is usually employed to find the principal components of **R**. In order to find the minimum number of components necessary for data reproduction within residual error, the components are added step by step to the model. After each step, the data matrix is reproduced and the procedure is continued until only non-systematic noise remains. The **first component** is substracted from the other variables and subsequently the second PC is selected with the same criterion:

$$\mathbf{X}_2 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T \tag{53}$$

and so on:

$$\mathbf{X}_k = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T - \mathbf{t}_2\mathbf{p}_2^T - \ldots - \mathbf{t}_k\mathbf{p}_k^T \tag{54}$$

The essence of the PCA method is to decompose the **X** matrix as

$$\begin{aligned} \mathbf{X} &= \mathbf{t_1 p_1^T} + \mathbf{t_2 p_2^T} + \mathbf{...} + \mathbf{t_k p_k^T} + \mathbf{X_k} \\ &= \mathbf{T_k P_k^T} + \mathbf{X_k} \end{aligned} \tag{55}$$

where **T**$_k$ and **P**$_k$ contain the first *k* columns of **T** and **P**, respectively. *k* is chosen in such a way that **X**$_k$ is small and represents only noise, while the term $\mathbf{T_k P_k^T}$ represents the salient features of **X**. In order to accomplish this, *k* must be chosen in such a way that the terms ignored correspond to zero or negligible eigenvalues. In order to help rationalize the choice of *k*, taking into account that the percentage of variance attached to each component is proportional to its eigenvalue, the relative size of the eigenvalues is expressed as a percentage of the sum of all eigenvalues:

$$\frac{\lambda_1}{\lambda_1 + ... + \lambda_k} \times 100 \tag{56}$$

and this percentage is interpreted as the percent variation explained by the corresponding principal component. Also, the **cumulated percentages** are used, so that the percent variation explained by the first *k* components is:

$$\frac{\lambda_1 + \cdots + \lambda_\mathrm{k}}{\lambda_1 + \cdots + \lambda_\mathrm{j}} \times 100 \tag{57}$$

Normally, the majority of the variance contained in the original variables matrix can be reproduced by means of the first components $k < m$. It can be assumed that the first PCs are a valid representation of **X,** because the residual variance can be assumed as negligible background noise. The obtained loading vectors are important to discriminate between relevant and redundant **X** descriptors, and to understand the score vectors, which inform on the similarities and dissimilarities between the studied objects in the model.

The graphical representation of the PCs is obtained by plotting the first PC vs. the second, and eventually vs. the third. In addition, the scores or loadings for each principal component can be plotted against each other in order to examine the data structure, and to allow the visualisation of relationships in the descriptor space.

6.6.3 Principal Components Regression

Principal Components Regression (PCR) method, applies the scores from a PCA decomposition as regressors in the QSAR model. Hence, a multiple-term linear equation is built, based on a principal components analysis transformation of the independent variables.

The PCA method, after choosing a suitable value for *k*, assumes that the important features of **X** have been retained by $\mathbf{T}_k$. Provided that score components are orthogonal and contain the majority of the variance of **X,** they are adequate as regression variables for **y** using MLR:

$$\mathbf{y} = \mathbf{Tq} + \mathbf{e} \rightarrow \mathbf{y} = q_1\mathbf{t}_1 + q_2\mathbf{t}_2 + ... + q_k\mathbf{t}_k + \mathbf{e} \qquad (58)$$

where q_i are the regression coefficients describing the relationship between the response variable (**y)** and the *k* score components (**T**) , and **e** is the error term. Analogously to MLR, the minim squared solution in the estimation of **q** is:

$$\mathbf{q} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)^{-1}\mathbf{T}^{\mathrm{T}}\mathbf{y} \qquad (59)$$

where q is the so-called vector matrix of regression coefficients for T.

For prediction with PCR, it is necessary to turn to **X** again. The substitution of **T** by **XP** yields:

$$\begin{aligned}\mathbf{y} &= \mathbf{Tq} + \mathbf{e} \\ &= \mathbf{XPq} + \mathbf{e}\end{aligned} \qquad (60)$$

The matrix **Pq** is called the regression matrix, and may be compared with the vector **b** of MLR:

$$\mathbf{b} = \mathbf{P}\mathbf{q} \tag{61}$$

However, this method does not work well if some of the variables contain a lot of variance but do not correlate with activity. Such variables may be given a high loading in the components, pushing out other variables more relevant to activity.

6.6.4 Partial Least Squares

PCR and PLS regression differ in the methods used for extracting factor scores. PCR produces the weight matrix reflecting only the covariance of the predictor variables, while PLS regression includes the response variables **y** in the process of reduction of the variables, and so the covariance is between the independent and dependent variables.

The most important aspect of the algorithm is that the score vectors for **X** and **Y** are calculated interdependently, with the score vector used as starting point for each iteration. In this way the **Y** data affects the decomposition of **X**, and the other way around.

PLS is a generalization of regression of particular interest in QSAR because, unlike MLR, data with strongly correlated (collinear), noisy or numerous **X** variables can be analyzed. In addition, several activity variables, **Y**, can be modelled simultaneously. Therefore, PLS is able to investigate complex structure-activity problems, to analyze data in a more realistic way, and to interpret how molecular structure influences biological activity. PLS gives a reduced statistically robust solution and, in fact, it contains MLR as a special case when a MLR exists. Besides, scores and loadings provide useful information about the correlation structures of the variables, and the structural similarities or dissimilarities between compounds.

In this thesis, only the particular case considering a single response variable, **y**, that is, the process is called **PLS1**, will be presented in detail.

It has to be noted that, first, it is convenient to autoescale the data matrix **X,** so that each column has a mean of zero and a unitary standard deviation.

The goal of PLS, is to seek the direction in the space of **X**, which yields the biggest covariance between **X** and **y**. This direction is given by a loading vector, such as the unitary weight vector $\mathbf{w}$:

$$\mathbf{w} = \frac{\mathbf{X}^T\mathbf{y}}{\left\|\mathbf{X}^T\mathbf{y}\right\|} \tag{62}$$

The scores, **t**, to be used in the regression are linear combinations of the original variables, **X**, calculated as:

$$\mathbf{t} = \frac{\mathbf{Xw}}{\mathbf{w^T w}} \tag{63}$$

By definition $\mathbf{w}^T\mathbf{w} = 1$, so that the previous expression can be simplified to $\mathbf{t} = \mathbf{Xw}$, where the columns of **W** are weight vectors for the **X** columns producing the corresponding factor score matrix **T**. The loading vector **p** needed for the calculation of the residuals of the new model can be obtained by regression of **X** with **t**.

$$\mathbf{p} = \frac{\mathbf{X^T t}}{\mathbf{t^T t}} \tag{64}$$

To make possible the estimations of **y** on **t**, it is necessary to calculate *q*, by means of ordinary least squares procedures:

$$q = \frac{\mathbf{y^T t}}{\mathbf{t^T t}} \tag{65}$$

Finally, the new **X** residuals are calculated from the previous component:

$$\mathbf{X}_{k+1} = \mathbf{X} - \mathbf{t}_k\mathbf{p}_k^T \tag{66}$$

$$\mathbf{y}_{k+1} = \mathbf{y} - \mathbf{t}_k b_k \tag{67}$$

The iterative process is repeated until a certain tolerance limit between two consecutive iterations of **y** is smaller than a given threshold.

Analogously o the PCA, the regression coefficients $\mathbf{b}_{\text{PLS}}$ are useful for the interpretation of the model, and o predict external compounds as $\mathbf{y} = \mathbf{Xb_{PLS}}$. The $\mathbf{b}_{\text{PLS}}$ coefficients are then calculated as:

$$\mathbf{b_{PLS}} = \mathbf{W}\left(\mathbf{P^T W}\right)^{-1}\mathbf{q} \tag{68}$$

where **W**, **P**, and $\mathbf{q}^T$ are calculated from their *k* components.

This algorithm is also called the orthogonalized PLS algorithm, because the estimated scores and the weight vectors are orthogonal ($\mathbf{t}_i^T \mathbf{t}_j = 1$ and $\mathbf{w}_i^T \mathbf{w}_j = 1$, where $i \neq j$). If the number of extracted components *k* is equal to the number of descriptors contained in **X**, then the PLS solution is equal to the MLR solution.

The iterative PLS method was originally designed by Wold [163]. Afterwards, several variants of the PLS algorithm have been proposed. Between them, the algorithm by Martens, which considers non-orthogonal scores and weights [164], and the PLS2 algorithm, for the case of more than a single column in **Y**, must be remarked.

6.6.5 Other methodologies

Neural Networks. Artificial Neural Networks (ANN) method [165-167] is a system inspired in the human brain, composed of many simple processing units operating in parallel, the so-called neurones. This discipline is also known with the names of parallel distribution, neurocomputation, natural intelligence systems or learning algorithms. In any case, the objective is to simulate the multiple shells of the simple neurones, where each neuron is attached to a number of neighbouring neurones with variable coefficients of connectivity that represent the force of these connections. The learning process consists of adjusting the coefficient so that the network provided as an output the appropriate results.

Other methods are **non-linear regressions** [168-171], which can be linear, non-linear and polynomial, **PARAFAC decomposition** [172-173], **Multilinear PLS regressions** [174],an extension of the PLS2 model, **multimode regression of principal covariances** [175], and **Hybrid Intelligent Systems (HIS)**, among others.

6.7 Selection of variables

In any QSAR study, the driving purpose is to build a QSAR model relating the independent variables matrix, **X,** with the property vector, **y**. The $(n \times m)$ data matrix may contain the *m* original molecular variables or transformed components. Independently of the provenance of **X** elements, before deducing the MLR model, the selection of variables is required in order to choose the descriptors that will act as descriptors in the QSAR model. The main goal in this step is the obtaining of the best QSAR model, using the minimum number of parameters, *k*, so that $k < m$.

There are different criteria to establish the best model and different methodologies to choose the variables.

There are statistical adjustment methods that imply an intrinsic variables selection, such as MLR [185], discriminant analysis [186], PLS [187-189], and more recently evolutive and genetic algorithms, such as k-nearest neighbours regression [190] or neural networks [191-192].

In MLR analysis, the determination of the variables to be selected cannot be made by means of the *r* correlation coefficient, because it progressively augments with the addition of new parameters to the MLR equations. Instead, other parameters that will be presented in the following section may indicate if the model is over-parameterised for the inclusion of an excessive number of descriptors. To conclude, there are also methods that contribute to select the optimal descriptors, by eliminating redundant descriptors or descriptors contributing with negligible information. These methods are based on the elimination of linear combinations of descriptors, descriptors with a small variation coefficient or, like in the PCA, with small associated variance.

6.7.1 Selection independent of the external variables

The most common case considers a data matrix transformed by means of any multivariate analysis technique, where the matrix has been reduced in order to alleviate computational costs and optimise the results. In this situation, the obvious selection of variables comes from the choice of the *k* leader factors, components or coordinates. For example, in PCA studies the common choice is made according the maximal variance, in order to define a subspace of dimension *k* that fits the original data. In the PCR analysis, also the leader *k* vectors are selected in order of decreasing eigenvalues.

Alternatively, in classical scaling techniques, the resulting configuration keeps the cardinality of the original matrix, that is, the configuration has the same number of components. The most obvious selection is to choose the *k* axis with the maximum eigenvalue, which provide the best *k* dimensional space that fits to the primitive data:

$$\mathbf{x}_1 \succ \mathbf{x}_2 \succ \ldots \succ \mathbf{x}_k \quad \Leftrightarrow \quad \lambda_1 \succ \lambda_2 \succ \ldots \succ \lambda_k \tag{69}$$

The main drawback of this selection independent of the external variables is that it assumes that the property correlates better with the axis that describe better the differences between the members in the set, but this is not necessarily true [176]. In QSM, the first PCs contain information on the more external structural differences of molecules. However, if the property is not associated to these structural characteristics, the regression parameter may not contribute significantly to the predictivity of the model.

6.7.2 Most Predictive Variable Method

Conversely, selection techniques dependent on the external variables allow the selection of other principal axis subspaces than the subspaces of maximal variance, which have into account the studied property. The most predictive variable method, MPVM [177-178] reorders the principal axis following an expression that measures the individual correlation of each axis with the external data (**y**):

$$\chi^2(\mathbf{x}_i,\mathbf{y})=\frac{\left(\mathbf{y}^{\mathrm{T}}\mathbf{x}_i\right)^2}{\sum_j\left(y_j-\bar{y}\right)^2\lambda_j} \tag{70}$$

where **y** is the properties vector, $\mathbf{x}_i$ is the *i*-th principal axis, and λ_j the eigenvalue corresponding to each principal coordinate. Hence, principal axes are arranged according to the decreasing value of the χ^2 coefficient:

$$\chi^2\left(\mathbf{x}_{i_1},\mathbf{y}\right)>\ldots>\chi^2\left(\mathbf{x}_{i_p},\mathbf{y}\right) \tag{71}$$

This method [179] chooses the *k* predictive variables that maximize the $\mathrm{X}^2_{(k)}$ determination-like coefficient:

$$\mathrm{X}^2_{(k)}=\sum_{\alpha=1}^{k}\chi^2\left(\mathbf{x}_{i\alpha},\mathbf{y}\right) \tag{72}$$

Thus, MPVM selects the principal axes that project better the data to be correlated. The limit case is when the axis is identical to the property vector; then, the projection is maximal and this vector is the selected as the more predictive one.

In order to avoid the participation of variables with low variance and to avoid the parameterisation of background noise, a filter can be imposed [180]. In this way, only the principal axes with a variance higher than a certain threshold are selected.

6.7.3 Nested Summation Symbols method

However, in the MPVM, the correlation measure between the axes and the experimental data is individual and does not take into account the inclusion of new parameters that can affect the total model. Stated differently, the maximal individual correlation does not ensure the maximal collective correlation. An alternative is the **Nested Summation Symbols (NSS)** method [181-184], which scans all the principal axes. This technique, applicable to any matrix, systematically searches the best descriptors among all the possible combinations of variables of the data set. Hence, NSS algorithm performs all the possible combinations of the total set with *m* descriptors over *k* parameters, which is the number of descriptors to be selected $\begin{pmatrix} m \\ k \end{pmatrix}$. This mathematic procedure allows selecting the most statistically significant MLR model generated with *k* parameters by generating loops of different profundity. The main inconvenient of this method is the high computational cost due to the high number of combinatorial possibilities when *k* is relatively small.

6.8 Evaluation of the quality of the model: Statistical Adjustment Parameters

The basic descriptive statistics [193] summarises specific features of the independent variable data set, such as count of elements, minimum, and maximum value that define the range of data; mean, standard deviation, variance, covariance, PRESS, and median; kurtosis (thickness of the tails of a distribution curve), and skewness (symmetry of the distribution of values), and other parameters like *F* statistic and Student *t* statistic. Besides, confidence intervals assess the significance, that is, the precision, of statistical parameters, by indicating reference limits within the value of the statistics is reasonably declared.

The two most important aspects of **precision** are reliability and validity. **Reliability** refers to the reproducibility of a measurement; poor reliability reduces the ability to track changes in experimental measurements. The reliability can be quantified simply by taking several measurements on the same subjects, and studying changes in the mean, standard deviation, and other parameters such as the Retest correlation, Kappa coefficient, and Alpha reliability, which refer to the reproducibility of values of a variable.

Validity refers to the agreement between the value of a measurement and its true value. Poor validity reduces the ability to characterize relationships between variables in descriptive studies. Validity is quantified by comparing the measurements with values that are as close to the true values as possible. The main measures of validity are the estimation equation, the typical error of the estimate, and the validity correlation, among others.

Once built the model, the **goodness of fit** quantitatively measures the **precision** of the fit, that is, the extent to which an estimated equation fits the data. It can be evaluated by means of the **standard and typical deviation,** but the most commonly quoted statistic used to such an extent is the **correlation coefficient**, and the related **coefficient of multiple determination**. Atlhough goodness of fit for models in which the dependent variable is discrete or cathegoric is not straightforward, various analogs of the correlation coefficient can be used.

The measures of **sensitivity** and **specificity** can also be regarded as measures of goodness of fit. Sensitivity is obtained when positive experimental results coincide with predicted ones (true positives); in contrast, if negative experimental values agree with negative predictions (true negatives), then specificity occurs. Parallely, an active compound which is predicted to be inactive is a false negative, whereas an inactive compound predicted to be active is named false positive.

	Pred (+)	Pred (–)	
Exp (+)	a	b	a+b
Exp (–)	c	d	c+d
	a+c	b+d	N

$$sensitivity = \frac{a}{a+c} \tag{73}$$

$$specificity = \frac{d}{b+d} \tag{74}$$

$$false\ positive = c \tag{75}$$

$$false\ negative = b \tag{76}$$

Finally, the evaluation of the **stability of the model** can be inspected by means of the leave-one-out cross-validation method. Also the biological confidence in reproducibility and the confidence in the source of data should be carefully inspected.

6.8.1 Standard Deviation

Standard Deviation (SD) is the squared root of the variance, or the Root Mean Square (RMS) error of deviations.

$$s_y = \sqrt{\frac{\sum_i^n \left(y_i - \overline{y}\right)^2}{n-1}} \tag{77}$$

where $n-1$ is he number of degrees of freedom, i.e. the number of parameters to be determined is subtracted from the total number of parameters.

This is the standard deviation usually employed in statistics, which measures the dispersion of a data set in relation to the arithmetic mean, that is, it is a measure of the magnitude of the residuals, accounting for **accuracy**.

Conversely, the standard deviation of the dependent variable, before trying to fit any model represents the amount of variation in the dependent variable, and the error represents the variation left over after fitting the model.

$$s_x = \sqrt{\frac{\sum_i^n \left(x_i - \overline{x}\right)^2}{n-1}} \tag{78}$$

6.8.2 Variance

The variance is the average of the squared standard deviation from the mean. Sums of squares are directly related to variances.

$$s_x^2 = \frac{\sum_i^n \left(x_i - \overline{x}\right)^2}{n-1} \tag{79}$$

$$s_y^2 = \frac{\sum_i^n \left(y_i - \overline{y}\right)^2}{n-1} \tag{80}$$

1.1.1 Typical Deviation

The typical deviation of the adjusted data, also known with the name of **Standard Deviation of Errors of Calculation** (**SDEC),** is expressed as:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - y_i^{'}\right)^2}{n}} \tag{81}$$

In contrast to the aforementioned standard deviation, the typical deviation ponders the residual sum of the squared deviations.

1.1.2 Covariance

Covariance is the average of product of deviations from means, for the {x_i,y_i} pairs, namely, the :

$$c_{XY} = \frac{\sum_{i}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{n-1} \tag{82}$$

This measure, that depends on the scaling of the features, describes the relationship between the *x* and *y* features. The covariance is the mean value of all the pairs of differences from the mean for independent variables multiplied by the differences from the mean for dependent variables. If *x* and *y* are not closely related to each other, they do not co-vary, so the covariance is small, so the correlation is small. If *x* and *y* are closely related, C_{XY} turns out to be almost the same as the product of standard deviations of *x* and *y*, so the correlation is almost 1.

All variances and covariances of a multivariate data set can be arranged in a symmetric matrix, with the variances located in the main diagonal. The covariance matrix describes the dispersion of multivariate data. Highly correlated features make **C** singular and then the inversion of impossible.

The regression coefficient **b** can be redefined in terms of variance and covariance: $b = \frac{c_{XY}}{s_x^2}$

1.1.3 Correlation coefficient

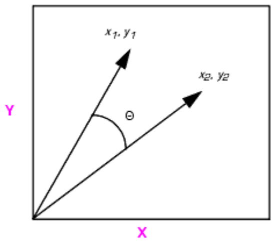
The correlation coefficient is a normalized covariance independent from scaling: $r = \frac{c_{XY}}{s_x s_y}$ that measures the quality if adjustment, that is, the degree of correlation between *x* and *y*, and detects if the variables contain redundant information, and thus they are highly correlated.

Substituting the corresponding identities in the previous equation:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\left[\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2\right]^{1/2}\left[\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2\right]^{1/2}} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]^{1/2}\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]^{1/2}} \quad (83)$$

The correlation coefficient is a measure of the degree of linearity of the relationship, i.e. it indicates the extent to which the pairs of numbers for these two variables lie on a straight line.

The correlation coefficient can be also pictured using vector notation. If two vectors of the same length are correlated, the angle between them approaches zero and the cosine approaches one.



Then the normalized correlation coefficient between the two variables *x* and *y* is computed as:
$\cos\theta = r$.

The correlation coefficient is comprised between -1 and 1,

$$-1 \le r \le 1 \quad (84)$$

Variables positively correlated have $0 > r > 1$, and those negatively correlated have $-1 < r < 0$. For perfect linearity, $r = \pm 1$. If there is no linear trend at all, but there is a random scatter of points, the value of *r* is close to zero.

Correlation coefficients for the variables in a dataset are compiled in a correlation matrix, which shows the correlation of one descriptor with another, and thus the relationships among descriptors. This matrix is a symmetric matrix in which the diagonal elements are one and the off-diagonal elements are the correlation coefficients for the appropriate variable pairs. The correlation coefficients for independent variables that are not correlated, i.e. orthogonal variables, are zero.

The addition of new parameters to the model always increases the *r* value, unless the new parameter is a constant of a linear combination of other parameters, which would not produce any effect. The increase in *r* when adding new parameters can result in overfitting, that is, a spurious correlation which parameterises the background noise.

6.8.3 Multiple determination coefficient

The multiple determination coefficient is the squared correlation coefficient used to describe the goodness-of-fit of the data. It informs about how well the model reproduces the experimental data. However, when a large number of free parameters intervene in the model, r^2 can arbitrarily be close to the value of one.

$$r^2 = \frac{c_{XY}^2}{s_x^2 s_y^2} \tag{85}$$

An alternative definition of the squared correlation coefficient can be deduced [193]:

$$r^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i - y_i' \right)^2}{\sum_{i=1}^{n} \left(y_i - \overline{y} \right)^2} \tag{86}$$

Using the abbreviated notation, the goodness-of-fit of the model can be expressed by:

$$r^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_R}{SS_T} \tag{87}$$

The multiple determination coefficient is a quantitative measure of the **precision** of adjustment for the fitted values to the observed ones, which measures the fraction of the variance explained by the model. The coefficient mainly informs if the variation of *y* explained by the regression equation permits to assume that there is a linear relationship between *y* and x..

The squared coefficient multiplied by 100 is the percent of total variance explained by the model. This percentage expresses the strength of the relationship between *x* and *y*.

r^2 is defined in the [0,1] interval, that is, it ranges from 0 to 1. The closer to the unity, the more similar are the adjusted values to the experimental ones. The limit case, when r^2=1, is obtained when all the residuals are null, that is, the residual sum of the squares approaches to zero, and, thus, the model fits exactly the data.

It must be noted, however, that a coefficient close to the unit does not mean that the model is good; the simple addition of parameters to the regression induces an ever-increasing of r^2, even if the newly added descriptor does not contribute to the model. To determine the predictive capacity of the model, other measures are required.

6.8.4 Statistical significance

A method to check if there are too many parameters in the model is to calculate the probability of obtaining a statistical result only by chance. The statistical significance or importance is the probability of obtaining statistical results only by chance. The statistical significance of two MLR models with the same number of points and the same number of parameters *m* can be easily evaluated, by means of the direct comparison of the corresponding correlation coefficients. However, when the number of points or parameters is different, it is not trivial to decide which model is the most statistically significant. Models can be compared using a recently proposed analytical criterion [194], based on the calculation of the probability that a model with a given correlation coefficient *r* is obtained accidentally. The probability is computationally calculated by simulating thousands of random correlations and comparing the obtained percent correlation explained with the original one. This probability is given by *P*, where *n* and *m* denote the number of data points and the number of parameters, respectively.

$$P = \frac{\int_0^{\arccos(r)} \cos^{n-1}\theta \sin^{n-m-2}\theta d\theta}{\int_0^{\pi/2} \cos^{n-1}\theta \sin^{n-m-2}\theta d\theta} \tag{88}$$

Given a $(n \times m)$matrix of independent variables $\mathbf{X} = \{x_{ij}\}$, and a vector of dependent parameters $\mathbf{y} = (y_1, y_2, ..., y_n)^T$, first the *r* coefficient is calculated. Then the same correlation is analyzed using a set of variables randomly generated, instead of the original **y** and **X**. The correlation coefficient *R* of the MLR generated with the stochastic variables is probably inferior than the initial value of *r*. Repeating the same experiment a number of times, there is a non null probability *P*, that the correlation coefficient for one of the variables set randomly generated is equal o superior than *r*. This probability depends on the number of points *n*, on the value of the correlation coefficient *R* and the number of parameters *m*; the lesser the value of *P*, the most difficult is to obtain a correlation with $R > r$. A simple geometric model that calculates analytically the probability *P* given the *n, m, r* values has been proposed. If the probability to obtain randomly the same result is high, the model must be rejected.

These probabilities are directly related to the so-called confidence level of a correlation, CL:

$$\%\ \mathrm{CL} = 100 * (1\text{-P}) \tag{89}$$

6.8.5 Fischer statistic

The **Fischer statistic** parameter is one of several variance-related parameters that can be used to compare two models differing by one or more variables. This statistic is used to determine whether a more complex model is significantly better than a less complex one.

$$F(k, n-k-1) = \frac{SS_R}{ks^2} \tag{90}$$

The *F* statistic is computed and compared with standard tabulated values. If *F* is larger than the tabulated value, the more complex model can be accepted as significant. The *F* statistic is related to the *t* Student statistic by F = t2.

where s^2 is an unbiased estimate of the residual or error variance, and $(k, n-k-1)$ are the degrees of freedom.

6.9 Statistical plots

Other tools that provide visual information about the model can be easily obtained by graphical representation. Graphical examples for statistical plots can be found in the chapter of application examples.

The plot of **predicted versus experimental** data displays the activity predicted by a QSAR equation against the experimentally measured or observed activity. The data are plotted as a scatter plot, where each point represents one structure of the molecular set. The QSAR equation is plotted as a regression line, which should ideally form a straight line drawn through origin with slope 1. This standard plot may be useful to identify outliers.

The **plot of residuals** displays the residuals, that is, the differences between predicted and observed activities, for a QSAR equation and set of structures. This plot is usually presented as a histogram, plotting residual values against observations, each observation representing the data for a single structure.

The depiction of the cross-validated or the raw **correlation coefficient versus the number of descriptors** aids to select the number of descriptors that presents a satisfactory compromise with the qulity of adjustment.

For **randomisation tests**, usually q^2 (or r_{CV}^2) is represented against the adjustment correlation coefficient (r^2) for all the generated models, marking distinctively the real obtained model with those calculated with the permuted responses.

7 EVALUATION OF THE PREDICTIVE CAPACITY OF THE MODEL: VALIDATION TECHNIQUES

Once the regression equation is obtained, in addition to the **goodness of fit** and the **stability** of the model, it is also important to evaluate the **robustness** and the **predictive capacity or validity** of the model before using the model on the interpretation and prediction of the biological activity.

To validate a method is to establish the reliability and relevance of the method for a particular purpose. The reliability refers to the reproducibility of results, the relevance is related to the scientific use and practical usefulness, and the purpose refers to the intended application. The validation of a QSAR model is the process by which the predictive ability of a QSAR and the mechanistic basis are assessed for practical purposes. Validation assesses if the model accurately represents the reality, from the perspective of the intended model application.

It must be paid special attention to **outliers**, structures with a residual greater than two times the standard deviation of the residuals that do not fit the model. Once identified, diagnostic data that help making decisions about them should be examined. Outliers should be iteratively removed from the observations used to calculate the QSAR equation, and then the equation recalculated until the satisfactory results were obtained.

It may happen, for example, if the structure of one or more elements of the training set differs significantly form the rest, that these elements determine the quality and shape of the model. Several procedures can be used to check the **reliability and significance** of the model, i.e. that the size of the model is appropriate for the quantity of data available of non synthesised compounds, as well as provide some estimate of how well the model can predict activity for new molecules. There are two techniques to determine the confidence and robustness of the model, namely internal and external validation techniques.

7.1 Internal Validation

Internal validation uses the dataset from which the model is derived without adding new elements to the model and checks for internal consistency. The quality of the model can be internally calculated using different criteria.

7.1.1 Cross-validation technique

The most used process to determine the **stability** of a predictive model is by means of the analysis of the influence of each of its elements upon the final model. To such an extent, the **Cross-Validation** (**CV**) **technique** is extensively employed as an internal validation method of statistical models [195-197]. The procedure derives a new model using a reduced set of structural data. The new model is used to predict the activities of the molecules that were not included in the new-model set. This is repeated until all compounds have been deleted and predicted once. The cross-validation process consists on extracting a certain number of objects, *k*, of the initial set, construct a new model with the remaining $n-k$ data, and use the reduced model to predict the dependent variable of the objects initially excluded. The process is repeated as many times as necessary until the vector with all the predicted values of the properties is obtained

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, ..., \hat{y}_n)^{\mathrm{T}} \qquad (91)$$

That is, the process is performed until all molecules have $\begin{pmatrix} n-1 \\ m-1 \end{pmatrix}$ predictions, where *n* is the number of molecules on the set and *m* the number of extracted molecules. The depth of the CV study depends on the number of extracted elements, *m*, of the data set.

Usually, one element of the set is extracted each time, and then the model is recalculated using as a training set the $n-1$ remaining objects, so that the property value for the extracted element is predicted once for all compounds. This process is repeated *n* times for all the elements of the initial set, thus obtaining a prediction for each object. This is the so-called **leave-one-out (LOO**) method.

Analogously, other stability measures of the prediction models can be defined when leaving out more than a molecule of the system at each time. These procedures are generally termed **leave-n-out** or **Leave-many-Out (LmO) CV method**. In this case, if *k* molecules are removed at once from a total set of *n* molecules, then $k \times n$ regressions are performed.

The capacity of prediction of the model can be obtained by two coefficients: the coefficient of prediction (q^2), and the CV coefficient of correlation (r_{CV}).

From the predicted values for each object, the **Predictive Residual Error Sum of Squares (PRESS)** is calculated [198]:

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{\mathrm{cv}} \right)^2 \tag{92}$$

where $\hat{y}_i^{\mathrm{cv}}$ is the predicted y value by cross-validation: $\hat{\mathbf{y}}^{\mathrm{cv}} = \left(\hat{y}_1^{\mathrm{cv}}, \hat{y}_2^{\mathrm{cv}}, ..., \hat{y}_n^{\mathrm{cv}} \right)^{\mathrm{T}}$.

It must be noted the difference between PRESS and the Total Sum of Squared deviations $\mathrm{SS_T}$ used in the definition of r^2 that was between the experimental property and the fitted value by the model.

By analogy with the expression of the coefficient of multiple determination, an estimation of the correlation coefficient of the cross-validation procedure, that is, the **coefficient of prediction** q^2 is defined from PRESS:

$$q^2 = 1 - \frac{\sum_{i=1}^{n} \left(y_i -, \hat{y}_i^{\mathrm{cv}} \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} = 1 - \frac{PRESS}{SS_T} = \frac{SS_T - PRESS}{SS_T} \tag{93}$$

where SS_T is the quadratic deviation of the observed values in relation to their arithmetic mean, and *PRESS* has been defined before.

q^2 is by definition smaller or equal (in the limit case where $r^2 = q^2 = 1$) than the overall r^2 for a QSAR equation. It is used as a diagnostic tool to evaluate the predictive power or the goodness of prediction of an equation generated using a regression method.

The actual computational advances allow the achievable realisation of CV with a deeper profundity. Besides, improved algorithms, which do not require the recalculation of MLR models, have been effectively conceived.

Another possibility is to calculate directly the **cross-validated correlation coefficient** between the original **y** and the predicted $\hat{\mathbf{y}}$ response variables, which can be represented by r_{cv} [199]. This definition is analogous to the coefficient of multiple determination, but replacing fitted MLR values by CV predicted ones. r_{cv} is defined within the interval [-1, 1], and it is calculated as the correlation coefficient but replacing the fitted by predicted activity. It must be noted that the squared coefficient r_{cv}^2 may lead to inverted predictions due to the loss of the sign of r_{cv}. To such an extent, usually the cross-validated values, $\hat{\mathbf{y}}$ are represented versus the experimental ones, **y**, in a bidimensional graph, ad the sign of the slope of the fitted line is examined.

However, the calculation of r_{cv} is computationally expensive. The CV predicted value of properties can be also calculated by means of the diagonal elements of the Hat matrix [129], thus avoiding the time-consuming expensive calculation of all the MLR models involved in the cross-validation. If the data of the *p* element is eliminated, the predicted property for *p* can be calculated as:

$$\hat{y}_p = \frac{1}{1-h_{pp}} \sum_{i \neq p}^{n} h_{pi} y_i \qquad (94)$$

Once calculated $\hat{y}_p$, the q^2 and r_{cv} coefficients can be evaluated. In this case, it can be demonstrated that the predictions for each element are not required in the expression of the PRESS; instead, the remaining objects are considered.

$$PRESS = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1-h_{ii}} \right)^2 = \sum_{i=1}^{n} \left(SS_{\mathrm{E}}^{\mathrm{cv}} \right)^2 \qquad (95)$$

PRESS is simply calculated by means of the CV $SS_{\mathrm{E}}^{\mathrm{cv}}$. However, this demonstration is only valid in absence of selection of variables.

However, in contrast to the r^2 parameter, the coefficient of prediction q^2 can be negative, ranging in the interval $(-\infty,1]$; so, the notation between parenthesis is recommended. The negative q^2 values, $q^2 < 0$, are easy o interpret: when the predictions of the model are worst than if the arithmetic mean of **y** vector was assigned as the predicted value for the data, the numerator of q^2 is negative. Consequently, the q^2 coefficient is smaller than zero. This indicates an awful predictive capacity. If $q^2 = 0$, the model considers the mean of activity or, in general, any arbitrary constant as the predicted value. This case is indicative of a null predictive capacity. Finally, when $q^2 > 0$, the model has the ability to correctly predict the model in a variable extent, varying according to the absolute value of q^2. The closer to the unity, the better predictiviness is achieved. As reference values, the commonly accepted values for a satisfactory QSAR model are $r^2 > 0.8$, and $q^2 > 0.5$.

r_{cv} and q^2 parameters can be used to determine the number of descriptors of the optimal model. Conversely to the classical adjustment coefficient, *r,* which augments with the progressive addition of parameters into the regression, the q^2 coefficient presents a curve with a maximum that corresponds to the optimum number of parameters and after this maximum, the curve decreases monotonally. This means that the increase of the number of parameters of the model always improves the adjustment of data but it is not related to the predictivity of the model. The limit case is when the number of parameters is equal to the number of elements of the system; in this case, the adjustment should be perfect, whereas the predictive capacity would be insignificant. Indeed, the descriptors of the model could be random values perfectly fitted to the data.Thus, the difference between both parameters can be indicative of the stability of the model. As a reference value, if $\left|r^2 - q^2\right| > 0.3$, this may indicate the presence of outliers, the selection of irrelevant descriptors, an insufficient number of data points, or the obtaining of an overfitted model, among others.

Another indicator of the quality of the model is the standard deviation between the original property values and the adjusted or predicted values. The value of σ is a measure of the error in adjustment or predictions. The smaller the σ, the better the model is.

It is important to note that when dealing with a reduced dimension data matrix, the prediction of each activity should be carried out by means of reduces models generated in the same conditions as the global model.

7.1.2 Randomization test

Another procedure to test the validity of the model is the **randomization test.** Even with a large number of observations and a small number of terms, an equation can still have very poor predictive power. This can come about if the observations are not sufficiently independent of each other. One way to test for this is by **randomization** of the independent variable. The set of activity values is re-assigned randomly to different molecules, and a new regression is performed. This process is repeated many times. If the random models' activity prediction is comparable to the original equation within a given estimated confidence level, the set of observations is not sufficient to support the model.

A typical **random test** [200] consists on arbitrarily permuting the response activity vector a certain considerable number of times. The new random vector is used as the real one to build a QSAR model in the same conditions as the original one, and analyse the capability of the prediction of the new ordered vector by means of the r_{cv} and q^2 values. The random test analyses the ability of the model to derive real structure-activity relationships.

If the model is correct, there must be a clear separation between the original fitted and predicted values and the values obtained from the random test. Instead, if relevant models are obtained using randomly ordered activity vectors, the model is suspicious of correlating whatever external data set. This may be an indication of overfitting, i.e.; an excess of degrees of freedom, or in other words, the number of descriptors is too large in comparison with the number of compounds.

There are different types of random tests, depending on the freedom of the model to select the data o the restriction to the regression descriptors that provide the optimal model. To avoid overparameterisation embedded in sophisticated statistical methods, the random test must be carried out allowing a totally algorithm that recalculate the regression coefficient and reselects the most predictive variables for the model without aprioristic conditions.

The common representation of the results of the random test is by means of a bidimensional graph representing q^2 versus r^2_{cv}, distinctively marking the points corresponding to the original activity data from the eventually generated ones. When the random test is satisfactorily achieved, there exists a clear separation between the original points and the random ones. If the test is not achieved, the points appear mixed, and the randomly generated models may even achieve more significant coefficients than the original one. In this case, the model is considered as spurious and it should be rejected. If there is not a clear separation, that is, only a number of points corresponding to random models approach the real one, the correlation between the permuted and the original activity vectors must be re-examined. If there is a significant correlation, this could indicate that the random vector does not significantly differ from the original one, or that permutations would have interchanged molecules with similar property and descriptors values. This could lead to results similar to the originals. In this case, the random test should be repeated.

7.2 External validation

Any model, even with excellent goodness of fit and satisfactory predictions, may lack of a real relationship between structural descriptors and activity. To evidence the existence of chance correlations [201], a reliable validation procedure must be carried out. The definitive validity of the model is examined by mean of **external validation**, which evaluates how well the equation generalizes. Two possible methods to carry out this method, canbe envisaged.

7.2.1 External test set

If a sufficiently large series of molecules with known activity is available, the original data set can be split into two subgroups, the **training set** and the **test set.** The training or calibration set is used to derive an adjustment model that is after used to predict the activities of the test or validation set members. Alternatively, also an external object set that has not been included in any phase of the construction of the model can be used as **test set.**

The obtained predictions of the new generated model for the test set determine the validity of the model. If the activity data is known, predictions are carried out on the same family with known activity that does not intervene in the exploration series.

The parameters quantifying the quality of prediction of the external test set may be the same used for the internal validation (q^2, r_{cv}, σ of prediction), substituting the CV predicted values in the formulas.

The Standard Deviation of Errors of Prediction (SDEP) or the Sum of Squares Prediction Errors (SSPE) are extensively used to account for the variability:

$$SSPE = \sum_{i=1}^{n_{\mathrm{T}}} \left(\hat{y}_i^{\mathrm{T}} - y_i^{\mathrm{T}} \right)^2 \tag{96}$$

where n_{T} is the number of molecules of the test set, $\hat{y}_i^{\mathrm{T}}$ is the prediction, and y_i^{T} the real activity value.

7.2.2 Internal test set

An inconvenient of this validation method is the availability of enough data to split the original set into two significant sets. If the set is not large enough, this method can be statistically in viable. Alternatively, an **internal test set** (ITS) can be simulated [202] using a procedure methodologically similar to LOO but with the total absence of intervention o the molecule extracted form the calculation thus achieving real predictions.

However, in some dimension reduction techniques, the PC for the test set can not be obtained directly from the descriptors between new objects and objects from the training set. Besides, it is erroneous to accept that the training set reduced space corresponds to the test set one. To such an extent, the dimension of the total set is reduced ad then the test set coordinates are extracted. Thus the elements of the test set contribute in the reduction of dimensions process. Although the model might not seem totally transparent, the effect of the test set on the calibration model is minimal and does not include the effect of the property.

8 DATA CLASSIFICATION: QUALITATIVE ANALYSIS OF DISCRETE PROPERTIES

Sometimes it is more desirable to obtain a qualitative association between structural descriptors and a biological property. Obviously, the required information to fit numerical values is higher than the information required to discriminate among discrete values range.

The motivation for this type of study can be either in compounds with a naturally grouped distribution of property values into discrete classes, in the case of continuous properties tat can be partitioned into categorical groups according to a predetermined threshold, and even in the case of originally discrete value properties, i.e. where experimental studies classify the compounds as active or inactive.

In the particular case of Quantum Similarity, reduction of dimension and selection of variables techniques allow the construction of **subspaces of similarity** where the objects can be graphically represented as points. Thus, structural similarity can be associated to the geometric proximity in the graph. In this way, qualitative information can be extracted on structure-property relationships from the distribution of the points in the space.

The main objective of **analysis discriminant techniques** [128-134] is to find a linear combination of factor that best discriminate between different classes. In **Linear Discriminant Analysis** (**LDA**), the mathematic functions defined to separate the classes are linear. Afterwards, the prediction for new object is based on the localization of the point-molecule in the subspace. The proximity of this new object to any class determines the prediction.

Cluster Analysis (CA) and **Cluster Significance Analysis (CSA)** [203-204] divide objects into isolated groupings, namely clusters, **k-nearest neighbours (k-NN)** [205] detects false connections between objects, and **facets theory** [206] provides a systematic scheme to relate regions with features of the data by partiotioning the multidimensional space in the regions, so-called facets diagrams. Also, **SIMCA (Soft Independent Modeling Class Analogy method)** has been developed fro pattern recognition and classification**.**

The **partition of a continuous property into discrete classes** can be done by different ways. In principle, the number of groups that can be considered can be any number; with a maximum corresponding to the total number of compounds. If the property presents values grouped in *k* ranges, the classes are simply constituted by the ranges. Conversely, if the distribution of the values is continues, the criterion to partition the property is only based in a homogenous distribution of the range, that allows classes with different size.

To determine the limits between classes the following expression is used:

$$\frac{n}{k}\left(y_{\max}-y_{\min}\right) \qquad n=1,\dots,k\text{-}1$$

where y_{max} and y_{min} are the maximal and minimal value of the property, k is the number of classes.

However, in QSAR studies, usually a rough separation into two groups, active versus inactive, is often considered.

Given a data matrix X for a set of binary classified compounds, two groups of compounds must be formed. The objective is to find two **group classification functions**, D_1, and D_2,

$$\begin{aligned} D_1 &= a_1 X_1 + a_2 X_2 + \dots \\ D_2 &= b_1 X_1 + b_2 X_2 + \dots \end{aligned} \tag{97}$$

such that

$$\begin{aligned} D_1 &> D_2 \quad \forall 1,\dots,k \\ D_1 &< D_2 \quad \forall k+1,\dots,n \end{aligned} \tag{98}$$

The coefficients a_i and b_j are the discriminant weights obtained by a multiple regression procedure. Classification functions define a line, plane, or, in general, a surface (hyperplane) between the groups. The difeference between the two group classification functions is called the linear discriminant function, D_{12}.

$$D_{12} = D_1 - D_2 \tag{99}$$

The similarity subspaces can be selected using different methodologies, connected with the **selection of variables technique**.

The **subspaces of maximal variance** are selected according to the amount of variance explained by the subspace, determined by the eigenvalues ordering. The criterion to use this selection corresponds to the assumption that maximal variance dimensions provide the optimal subspace. These subspaces reflect the more external structural differences among the compared molecules. If a real grouping into classes is obtained connected with the discretisation of the property, the property can be related to a specific type of substituent or other characteristics of the structure of the molecules.

The **subspaces of optimal variables** utilise the parameters selected by the MPVM variables selection technique used in classical scaling to explain the data.

Besides, **other criteria** to select the subspaces have also been developed. Among them, a method based on the search of the bidimensional space that minimises the number of crossings among objects of different classes [207] can be mentioned.

REFERENCES

1. *QSAR for prediction of fate and effects of chemicals in the environment.* Environmental Technologies RTD Programme (DGXII/D-1). Contract EV5V-CT92-0211. Commission of the European Union: Brussels, **1995**.
2. *Fate and activity modeling of environmental pollutants using structure-activity relationships (FAME).* Contract number ENV4-CT96-0221. Environment and Climate Programme of the European Union: Brussels, **1999**.
3. United States Environmental Protection Agency (EPA). Integrated Risk Information System (IRIS). Environmental Criteria and Assessment Office, Office of Health and Environmental Assessment, Cincinnati OH, **1994**.
4. Barratt, M.D.; Castell, J.V.; Chamberlain, M.; Combes, R.D.; Dearden, J.C.; Fentem, J.H.; Gerner, I.; Giuliani, A.; Gray, T.J.B.; Livingstone, D.J.; Provan, W.M.; Rutten, F.A.J.J.L.; Verhaar, H.J.M.; Zbinden, P. The Integrated Use of Alternative Approaches for Predicting Toxic Hazard. The Report and Recommendations of ECVAM Workshop 8. *ATLA, 23,* **1995**, 410-429.
5. Balbes, L.M. *Guide to Rational (Computer-Aided) Drug Design*. Research Triangle Institute: Research Triangle Park.
6. Van de Waterbeemd, H. *Quant- Struct.-Act. Relat.; 11*, **1992**, 200-204.
7. Van de Waterbeemd, H. *Drug Des. Disc., 9*, **1993**, 277-285.
8. 3D Virtual Chemistry Library. Imperial College of Science, Technology & Medicine. Available at http: www.ch.ic.ac.uk/vchemlib/mol/mol.html
9. LIGAND: Ligand Chemical Database for Enzyme Reactions and Chemical Compounds. Kyoto University. Available at http: www.genome.ad.jp/htbin/www_bfind?ligand
10. The Cambridge Structural Database System – from crystallographic data to protein-ligand applications. Stephen J. Maginn, avalaible in via CrystalWeb, ICSD-WWW, ConQuest, ReactionWeb or ISIS. Available at http: www.ccdc.cam.ac.uk/products/csd/
11. Spilker, B. *Multinational Drug Companies*. Issues in Drug Discovery and Development. Raven Press: New York, **1989.**
12. Topliss, J.G. (Ed.) Quantitative Structure-Activity Relationships of Drugs. Academic Press: New York, **1983.**
13. Stuper, A.J.; Brügger, W.E.; Jurs, P.C. *Computer-Assisted Studies of Chemical Structure and Biological unction*. John Wiley: New York, **1979.**
14. Martin, Y.C. Quantitative Drug Design. A critical introduction. Marcel Dekker: New York, **1978.**
15. Devillers, J.; Karcher, W. (Ed.) *Applied Multivariate Analysis in QSAR and Environmental Studies.* Kluwer Academic Publishers: Dordrecht, **1991**.
16. Tute, M.S. History and Objectives of Quantitative Drug Design. *Comprehensive Medicinal Chemistry. Vol 4. Quantitive Drug Design.* Hansch, C.; Sammes, P.G.; Taylor, J.B. (Eds.) Pergamon Press: Oxford, **1990**, 1-31.
17. Gubernator, K. (Ed.) *Structure-Derived Ligand Design. Methods and Principles in Medicinal Chemistry*; *Vol 4*. Mannhold, R.; Krogsgaard-Larsen, P.; Timmerman, H. (Eds.) VCH: Weinheim, **1995.**
18. Chaiken, I.M.; Janda, K.D. (Eds.) Molecular Diversity and Combinatorial Chemistry. Libraries and Drug Discovery. American Chemical Society: Washington, **1996**.
19. DeWitt, S.H.; Czarnik, A.W. (Eds.) *A Practical Guide to Combinatorial Chemistry.* American Chemical Society: Washington, **1997**.
20. Gillet, V.J. ; Wild, D.J.; Willett, P.; Bradshaw, J. Similarity and Dissimilarity Methods for Processing Chemical Structure Databases. *The Computer Journal, 41*, **1998**.
21. Martin, Y.C. Computer Assisted Rational Drug Design. In, *Methods in Enzymology*. Lilley, D.M.J.; Dahlberg, J.E. (Eds.) Academic Press: San Diego, **1991**, 587-613.
22. Turner, L.; Choplin, F.; Dugard, P.; Hermens, J.; Jaeckh, R.; Marsmann, M.; Roberts, D. Structure-activity relationships in toxicology and ecotoxicology: an assessment. *Toxicology in Vitro, 1,* **1987,** 143-171.
23. Karcher, W. Basic concepts and aims of QSAR studies. In *Quantitative Structure-Activity Relationships (QSAR) in Toxicology.* Coccini, T.; Giannoni, L.; Karcher, W.; Manzo, L.; Roi, R. (Eds.) CEC: Luxembourg, **1992,** 5-25.

24. Livingstone, D.J. Computational techniques for the prediction of toxicity. *Toxicology in Vitro, 8,* **1994,** 873-877.
25. Rebek, J. Molecular Recognition and Biophysical Organic Chemistry, *Acc. of Chem. Res., 23,* **1990**, 399-404.
26. Compendium of Medicinal Chemistry (ISIS)
27. NCI-3D Chemotherapy Screening (www)
28. Cambridge Structural Database X-Ray (Quest, ISIS)
29. CRC Properties of Organic Compounds (Windows, Microsoft Access)
30. Wold, S.; Dunn, W.J. Multivariate quantitative structure-activity relationships (QSAR): conditions for their applicability. *J. Chem. Inf. Comput. Sci., 23*, **1983,** 6-13.
31. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold S. *Multi- and Megavariate Data Analysis. Principles and Applications.* Umetrics: Umea, **2001**.
32. Wold, S.; Sjöström, M.; Carlson, R.; Lundstedt, T.; Hellberg, S.; Skagerberg, B. Multivariate design. *Anal. Chem. Acta*, 191, **1986,** 17-32.
33. Linusson, A.; Gottfries, J.; Lindgren, F.; Wold, S. Statistical molecular design of building blocks for combinatorial chemistry. *J. Med. Chem.*, *43*, **2000,** 1320-1328.
34. Cronin, M.T.D.; Schultz, T.W. Pitfalls in QSAR. *J. Mol. Struct., 633*, **2003,** 39-51.
35. Devillers, J. Statistical analyses in drug design and environmental chemistry: basic concepts. In, *Quantitative Structure-Activity Relationships (QSAR) in Toxicology.* Coccini, T.; Giannoni, L.; Karcher, W.; Manzo, L.; Roi, R. (Eds.) CEC: Luxembourg, **1992,** 27-41.
36. Topliss, J.G. Utilization of operational schemes for analog synthesis in drug design. In, *Drug Design; Vol. 5.* Ariens, E.J. (Ed.) Academic Press: New York & London, 1-21.
37. Hansch, C.; Unger, S.H.; Forsythe, A.B. Strategy in drug design. Cluster analysis as an aid in the selection of substituents. *J. Med. Chem.*, *16*, **1973,** 1217-1222.
38. Austell, V. A manual method for systematic drug design. *Eur. J. Med. Chem., 17,* **1982,** 9-16.
39. Streich, W.J.; Dove, S.; Franke, R. On the rational selection of test series. I .Principal Component method combined with multidimensional mapping. *J. Med. Chem.*, 23, **1980,** 1452-1456.
40. Baroni, M. ; Clementi, S. ; Cruciani, G. ; Kettaneh Wold, N.; Wold, S. D-Optimal design in QSAR. *Quant. Struct.-Act. Relat.*, 12, **1993,** 225-231.
41. Pleiss, M.A.; Unger, S.H. The design of test series and the significance of QSAR relationships. In *Quantitative Drug Design.* Ramsden, C.A. (Ed.) Pergamon Press: Oxford**, 1990**, 561-587.
42. Workshop “Regulatory Acceptance of QSARs for Human Health and Environment Endpoints” Setubal, Portugal, **2002**.
43. Baroni, M.; Clementi, S.; Cruciani, C.; Kettaneth-Wold, M.; Wold, S. *Quant. Struct.-Act. Relat., 12*, **1993,** 225-231.
44. Box, G.E.P.; Hunter, W.G.; Hunter, J.S. *Statistics for Experimenters.* Wiley: New York, **1978.**
45. Box, G.E.P.; Draper, N.R. *Empirical Model-Building and Response Surfaces.* Wiley: New York, **1987.**
46. Mendeleev, D.I. *Principles of Chemistry*; *Vol. 2,* **1868–71,** tr. **1905**.
47. Borman, S. New QSAR Techniques Eyed for Environmental Assessments. *Chem. Eng. News, 68,* **1990,** 20-23.
48. Crum-Brown, A.; Fraser, T.R. On the connection between chemical constitution and physiological action. Part 1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans. Roy. Soc. Edinburgh, 25,* **1868-9**, 151-203.
49. Crum Brown, A.; Fraser, T.R. On the Connection between Chemical Constitution and Physiological Action. Part II. - On the Physiological Action of the Ammonium Bases derived from Atropia and Conia. *Trans. Roy. Soc. Edinburgh*, *25*, **1869,** 693-739.
50. Richardson, B.J. *Medical Times and Gazette*, *2*, **1868,** 703.
51. Mills, E.J. *Philosophical Magazine*, *17*, **1884,** 173-187.
52. Richet, C. *Compt. Rend. Soc. Biol.* (Paris), *45*, **1893**, 775-776.
53. Overton, E. *Z. Physik. Chem*., *22*, **1897**, 189.
54. Meyer, H. Theorie der Alkoholnarkose, welche Eigenschaft dir Anästhetica bedingt ihre narkotische Wirkung. *Arch. Exp. Pathol. Pharmakol*., *42*, **1899**, 109-118.
55. Overton, E. *Studien über dir Narkose*. Gustav Fischer: Jena, **1901**.
56. Lipnick, R.L. Charles Ernest Overton: Narcosis Studies and a Contribution to General Pharmacology. *Trends Pharmacol. Sci*., *7*., **1986,** 161-164.
57. Ferguson, J. The use of chemical potentials as indices of toxicity. *Proc. Roy. Soc. Lond. B. Biol. Sci., 127,* **1939**, 387–404.

58. Free, Jr. S.M.; Wilson, J.W. A mathematical contribution to structure-activity studies. *J. Med. Chem., 7,* **1964**, 395–399.
59. Hansch, C.; Fujita, T. *J. Am. Chem. Soc*., *86,* **1964,** 1616-1626.
60. Hammett, L.P. The effect of structures upon the reactions of organic compounds. Benzene derivatives. *J. Am. Chem. Soc*., *59,* **1937**, 96-103.
61. Hammett, L.P. *Physical organic chemistry*. McGraw-Hill: New York, **1940**.
62. Taft, R.W. *J. Am. Chem. Soc*., *74,* **1952**, 3120-3128.
63. Taft, R.W. Steric effects in organic chemistry. Newman, M.S. (Ed.) Wiley: New York, **1956**.
64. Swain, C.G.; Scott, C.B. *J. Am. Chem. Soc. 75,* **1953,** 141-147.
65. Lupton Jr. E.C. Field and ressonance components of substituent effects. *J. Am. Chem. Soc*., *90,* **1968**, 4328-4335.
66. Bruice, T.C.; Kharasch, N.; Winzler, R.J. *Arch. Biochem. Biophys., 62,* **1956,** 305-317.
67. Bocek, K.; Kopecký, J.; Krivucova, M. Vlachova, D. Chemical structure and biological activity of p-disubstituted derivatives of benzene. *Experientia, 20,* **1964**, 667-678.
68. Kopecký, J.; Bocek, K. Vlachova, D. *Nature, 207,* **1965**, 667-678.
69. Fujita, T.; Ban, T. Structure-activity studies of phenilethylamines as substrates of biosynthetic enzymes of sympathetic transmmiters. *J. Med. Chem*., *14,* **1971**, 148-152.
70. Pullman, A.; Pullman, B. *Rev. Sci., 84,* **1946**, 145.
71. Pullman, A.; Pullman, B. *Ad. Cancer Res. 3,* **1955**, 117.
72. Hansch, C.; Maloney, P.P.; Fujita,T.; Muir, R.M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature, 194,* **1962,** 178-180.
73. Fujita, T.; Iwasa, J.; Hansch, C. A new substituent constant, π, derived from partition coefficients. *J. Am. Chem. Soc*., *86,* **1964**, 5175-5180.
74. Hansch, C.; Fujita, T. ρ−σ−π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc*., *86,* **1964**, 1616-1626.
75. Hansch, C.A quantitative approach to biochemical structure-activity relationships. *Acct. Chem. Res., 2***, 1969**, 232-239.
76. Zahradnik, R.; Chvapil, M. *Experientia, 16,* **1960,** 511 – 512.
77. Zahradnik, R. *Arch. Int. Pharmacodyn. Ther., 135,* **1962,** 311 – 329.
78. Zahradnik, R. *Experientia, 18,* **1962,** 534 – 536.
79. Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc*., *69,* **1947**, 17-22.
80. Kier, L.B.; Hall, L.H.; Murray, W.J.; Randić, M. Molecular connectivity. I: Relationship to non-specific local anaesthesia. *J. Pharma. Sci., 64, 1975*, 1971-1974.
81. Kier, L.B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research*. Academic Press: New York, **1976**.
82. Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc*., *97,* **1975,** 6609-6615.
83. Hosoya, H. *Bull. chem. Soc. Jpn*., *44,* **1971**, 2332-2337.
84. Trinajstić, N. *Chemical graph theory*. CRC Press: Boca Raton, **1992**.
85. Rum, G.; Herndon, W.C. Molecular similarity concepts. 5. Analysis of steroid-protein binding constants. *J. Am. Chem. Soc*., *113,* **1991**, 9055- 9060.
86. Herndon, W.C.; Rum, G. Three-dimensional topological descriptors and similarity of molecular structures. In QSAR and molecular modeling: concepts, computational tools and biological applications: proceedings of the 10th european symposium on structure-activity relationships, QSAR and molecular modeling. Sanz, F.; Giraldo, J.; Manaut, F. (Eds.) Prous Science, Madrid, **1995**, 380-384.
87. Cramer III, R.D.; Paterson, D.E.; Bunce, J.D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc., 110,* **1988**, 5959-5967.
88. Kubinyi, H. (Ed.) 3D QSAR in Drug Design. Theory, Methods and Applications. Leiden: ESCOM, **1993**.
89. Good, A.C.; So, S.S.; Richards, W.G. Structure-activity relationships from molecular similarity matrices. *J. Med. Chem*., *36,* **1993**, 433-438.
90. Good, A.C.; Peterson, S.J.; Richards, W.G. QSAR's from similarity matrices. Technique validation and application in the comparison of different similarity evaluation methods. *J. Med. Chem., 36,* **1993**, 2929- 2937.
91. Good, A.C.; Richards, W.G. The extension and application of molecular similarity to drug design. *Drug Information Journal, 30,* **1996**, 371-388.

92. Vedani, A.; McMasters, D.R.; Dobler, M. Multi-conformational ligand representation in 4D–QSAR: Reducing the bias associated with ligand alignment. *Quant. Struct.-Act. Relat.*, *19*, **2000**, 149-161.
93. Vedani, A.; Briem, H.; Dobler, M.; Dollinger, H.; McMasters, D.R. Multiple conformation and protonation-state representation in 4D–QSAR: The neurokinin–1 receptor system. *J. Med. Chem.*, *43*, **2000**, 4416-4427.
94. Vedani, A.; Dobler, M. Multidimensional QSAR: Moving from three- to five-dimensional concepts. *Quant. Struct.-Act. Relat.*, *21*, **2002**, 382-390.
95. Vedani, A.; Dobler, M. 5D-QSAR: The key for simulating induced fit? *J. Med. Chem.*, *45*, **2002**, 2139-2149.
96. Downs, G.M. Molecular Descriptors. In *Computational Medicinal Chemistry for Drug Discovery.* Bultinck, P.; De Winter, H.; Langenaeker, W.; Tollenaere, J. P. (Eds.). Marcel Dekker; New York, **2004**, 515-538.
97. Devillers, J.; Balaban, A.T. *Topological Indices and Related Descriptors in QSAR and QSPR.* Gordon Breach Scientific Publishers: Amsterdam, **1999**, 811.
98. Karelson, M. *Molecular Descriptors in QSAR/QSPR*. Wiley-InterScience; New York, **2000**.
99. Todeschini, R.; Consonni, V.; Pavan, M. DRAGON-Software for the Calculation of Molecular Descriptors. Release 1.12 for Windows. **2001.** Available: http://www.disat.unimib/chm [accessed 25 March 2002].
100. Jurs, P.C. ADAPT-Automated Data Analysis and Pattern Recognition Toolkit. University Park, PA: Pennsylvania State University. **2002.** Available: http://research.chem.psu.edu/pcjgroup/ADAPT.html [accessed 23 April 2002].
101. Stuper, A.J.; Jurs, P.C. ADAPT: A computer system for auto-mated data analysis using pattern recognition techniques. *J. Chem. Inf. Comput. Sci.*, *16*, **1976**99–105.
102. Mekenyan, O.; Bonchev, D. OASIS method for predicting bio-logical activity of chemical compounds. *Acta Pharm. Jugosl.*, *36*, **1986**, 225–237.
103. Katritzky, A.R.; Lobanov, V.S.; Karelson, M. CODESSA, Reference Manual. Gainesville, FL University of Florida. **1994** Available: http://www.semichem.com/codessarefs.html [accessed 19 April 2002].
104. Jaffé, H.H. *Chem. Rev.*, *54*, **1953,** 191-261.
105. Roberts, J.D.; Moreland, W.T. *J. Am. Chem. Soc.*, *75*, **1953,** 2167-2173.
106. van de Waterbeemd, H.; Testa, B. The Parametrization of Lipophilicity and other Structural Properties in Drug Design. In *Advances in Drug Research*; *Vol 16*. Testa, B. (Ed.) Academic Press: New York, **1987**, 85-225.
107. Leo, A.J. Methods of Calculating Partition Coefficients. *Comprehensive Medicinal Chemistry. Vol 4. Quantitive Drug Design*. Hansch, C. Sammes, P.G.; Taylor, J.B. (Eds.) Pergamon Press: New York, **1990**, 295-320.
108. Taft, R.W. Separation of Polar, Steric, and Resonance Effects in reactivity. In *Steric effects in Organic Chemistry*. Newman, M.S. (Ed.) Wiley: New York, **1956**, 556-675.
109. Charton, M. *J. Am. Chem. Soc.*, *41*, **1976**, 2217-2220.
110. Verloop, A. The STERIMOL Approach to Drug Design. Marcel Dekker: New York,**1987.**
111. van de Waterbeemd, H.; Testa, B. Adv. Drug. Res. 16, 85-225, **1987.**
112. Purcell, W.P. , Bass, G.E.; Clayton, J.M.; Strategy of Drug Design , Wiley, New York, **1973**.
113. Kamlet, M.J.; Taft, R.W.; *J. Am. Chem. Soc.*;98, 377-383, **1976.**
114. Atkins, P.W. *Physical Chemistry*. Freeeman: New York, **1982.**
115. Lien, E.J.; Guo, Z-R.; Li, R-L.; Su, C.-T. *J. Pharm. Sci.*, *71*, **1982,** 641-655.
116. Martin, Y.C. The Quantitative Relationships between pk$_a$, Ionization and Drug Potency: Utility of Model-Based Equations. In *Physical Chemical Properties of Drugs*. Yalkowsky, S.H.; Sinkula, A.A.; Valvani, S.C. (Eds.) Marcel Dekker: New York, **1980,** 49-110.
117. Kier, L.B. Molecular Orbital Theory in Drug Research. Academic Press: New York, **1971**.
118. Pearlman, R.S. Molecular Surface Areas and Volumes and Their Use in Structure-Activity Relationships. In *Physical Chemical Properties of Drugs*. Yalkowsky, S.H.; Sinkula, A.A.; Valvani, S.C. (Eds.) Marcel Dekker: New York, **1980**, 321-345.
119. Lee, B.; Richards, F.M. *J. Mol. Biol.*, *55*, **1971**, 379-400.
120. Hermann, R.B. *J. Phys. Chem.*, *76*, **1972,** 2754-2759.
121. Pearlman, R. SAREA Quantum Chemistry Program Exchange, University of Indiana, Bloomington
122. Grigoras, S. *J. Comp. Chem.*, *11*, **1990,** 493-510.

123.Silipo, C. Vittoria, A. Three-Dimensional Structure of Drugs. In *Comprehensive Medicinal Chemistry. Vol 4. Quantitive Drug Design*. Hansch, C. Sammes, P.G.; Taylor, J.B.; eds. Pergamon Press, New York, **1990**, 154-204.

124.Kier, L.B.*Quant. Struct.-Act. Relat.* Pahrmacol.; Chem. Biol.; 4, 109-116 (1985)

125.Kier, L.B.*Quant. Struct.-Act. Relat.* Pahrmacol.; Chem. Biol.; 5, 1-7 (1986)

126.Kier, L.B.*Quant. Struct.-Act. Relat.* Pahrmacol.; Chem. Biol.; 5, 7-12 (1986)

127.Höskuldsson, A. Prediction Methods in Science and Technology. Thor Publishing: Copenhagen, **1996**.

128.Xu, L.; Zhang, W.J. Comparison of different methods for variable selection. *Anal. Chim. Acta, 446,* **2001,** 477–483.

129.Montgomery, D.C.; Peck, E.A. *Introduction to linear regression analysis*. Wiley: New York, **1992**

130.Wold, S.; Johansson, E.; Cocchi, M. PLS—partial least squares projections to latent structures. In *3D-QSAR in Drug Design, Theory, Methods, and Applications.* Kubinyi, H. (Ed). ESCOM Science Publishers: Leiden, **1993,** 523–550.

131.Albano, C; Dunn, WG, III; Edlund, U; Johansson, E; Nordén, B; Sjöström, M; et al. Four levels of pattern recognition. *Anal. Chem. Acta, 103,* **1978,** 429–443.

132.Wold, S; Albano, C; Dunn, WJ, III; Esbensen, K; Hellberg, S; Johansson, E; et, al. Pattern recognition: finding and using regularities in multivariate data. In *Food Research and Data Analysis* Martens, H.; Russwurm, H. (Eds) Applied Science Publishers: Essex, **1983,** 147–188.

133.Tou, J.T.; González, R.C. *Pattern Recognition Principles*. Addison-Wesley: Reading, **1974**.

134.Hand, D.J. *Discrimination and classification*. Wiley: Chichester, **1993.**

135.Myers, R.H. *Classical and modern regression with applications*. PWS-KENT Publishing company: Boston, **1990**.

136.Jolliffe, I.T. *Principal Component Analysis*. Springer-Verlag: New York, **1986**.

137.Geladi, P.; Kowalsky, B.R. Partial Least Squares: a tutorial. *Anal. Chem. Acta*, *185*, **1986**, 1-17.

138.Wold, S.; Sjöström, M. ; Eriksson, L. Partial least squares projections to latent structures (PLS) in chemistry. In *Encyclopedia of Computational Chemistry*; *Vol.4.* Schleyer, P.R.; Allinger, N.L.; Clark, T.; Gasteiger, J.; Kollman, P.A.; Schaefer III, H.F.; Screiner, P.R. (Eds.). John Wiley and Sons Ltd.: Chichester, **1994**, 2006-2021.

139.Neter, J.; Wasserman, W.; Kutner, M. H. *Applied linear regression models*. Homewood, I.L: Irwin, **1989**.

140.Darlington, R.B. *Regression and linear models*. McGraw-Hill: New York**, 1990**.

141.Topliss, J.G.; Edwards, R.P. Chance factors in studies of quantitative structure-activity relationships. *J. Med. Chem., 22,* **1979,** 1238–1244.

142.Cuadras, C.M. *Métodos de análisis multivariante*. UAB: Barcelona, **1996**.

143.Mardia, K.V.; Kent, J.T.; Bibby, J.M. *Multivariate Analysis*. Academic Press: London, **1979**.

144.Cox, T.F.; Cox, M.A.A. *Multidimensional Scaling*. Chapman and Hall: London, **1994**.

145.Van de Waterbeemd, H. Chemometric Methods in Drug Discovery. In *Structure-Property Correlations in Drug Research*. Van de Waterbeemd, H. (Ed.); VCH: New York, **1995**, 65.

146.Borg, I.; Groenen, P. Modern Multidimensional Scaling: Theory and Applications. Springer: New York, **1997.**

147.Searle, S.R. Matrix algebra useful for statistics. Wiley: New York, **1982.**

148.Gower, J.C.; Legendre, P. Metric and Euclidean properties of dissimilarity coefficients. *J. Classification, 3,* **1986**, 5-48.

149.De Leeuw, J.; Heiser, W. Theory of multidimensional scaling. In *Handbook of Statistics*; *Vol. 2.* Krishnaiah, P.R.; Kanal, L.N. (Eds.) North-Holland: Amsterdam, **1982**, 285-316.

150.Cattell, R.B. The scree test for the number of factors. *Multivariate Behavioral Research*, *1,* **1966**, 245-276.

151.Kruskal, J.B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* **1964**, 1-27.

152.Kruskal, J.B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29,* **1964,** 115-129.

153.De Leeuw, J.; Heiser, W.J. Multidimensional scaling with restrictions on the configuration. In *Multivariate Analysis; Vol 5*. Krishnaiah, P.R. (Ed.) North-Holland: Amsterdam, **1980**, 501-522.

154.Borg, I.; Lingoes, J.C. A model and algorithm for multidimensional scaling with external constraints on the distances. *Psychometrika, 45,* **1980**, 25-38.

155.Kendall, M. *Multivariate Analysis*. Charles Griffin&Co.: London, **1975**.

156.Manly, B. F. J. *Multivariate Statistical Methods*. Primer, A. (Ed.) Chapman and Hall: London and New York, **1986**.

157.Nilsson, J. *Multiway calibration in 3D QSAR: applications to dopamine receptor ligands.* University Library Groningen: Groningen, **1998**, Online Resource.

158.Fisher, R.; MacKensie, W. *J. Agric. Sci., 13,* **1923**, 311-320.

159.Wold, H. *Research papers in Statistics*. David, F. (Ed.) Wiley & Sons: New York, **1966**, 411-444.

160.Wold, H. Soft modeling by latent variable, the non-linear iterative partial least squares (NIPALS) algorithm. In *Perspectives in probability and statistics*. Gani, J. (Ed.) Academic Press: London, **1975**, 117-142.

161.de Jong, S. SIMPLS: An Alternative Approach to Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, *18,* **1993**, 251-263

162.Mandel, J. *American Statistician*, *36*, **1982,** 15-24

163.Wold, S.; Ruhe, A.; Wold, H.; Dunn III, W.J. The collinearity problem in linear regression. The partial least squares ({PLS}) approach to generalized inverses. *J. Sci. Stat. Comput., 3*, **1984**, 735-743.

164.Martens, H.; Naes, T. *Multivariate calibration*. John Wiley and Sons: New York, **1989.**

165.Tetko, I.V.; Alessandro, E.; Villa, P.; Livingstone, D.J. Neural network studies. 2. Variable selection. *J. Chem. Inf. Comput. Sci., 36,* **1996**, 794-803.

166.Novič, M.; Nikolovska-Coleska, Z.; Solmajer, T. Quantitative structure-activity relationship of flavonoid p56 protein tyrosine kinase inhibitors. A neural network approach. *J. Chem. Inf. Comput. Sci., 37,* **1997**, 990-998.

167.Duprat, A.F.; Huynh, T.; Dreyfus, G. Toward a principled methodology for neural network design and performance evaluation in QSAR. Application to the prediction of log P. *J. Chem. Inf. Comput. Sci., 38,* **1998**, 586-594.

168.Wilde, D.J. *Optimum seeking methods*. Prentice-Hall Inc.: Englewood Cliffs, **1965**.

169. Tabak, D.; Kuo, B.C. *Optimal control by mathematical programming*. Prentice-Hall Inc.; Englewood Cliffs, **1971.**

170.Bard, Y. *Non-linear parameter estimation*. Academic Press: New York, **1974**.

171.Bukietyński, B. (Ed.) *Mathematical programming.* Akademia Ekonomiczna: Wroclaw, **1976**.

172.Carroll, J.; Chang, J.J. Analysis of individual diferences in multidimensional scaling with an n-way generalization of the Eckart-Young decomposition. *Psycometrika, 35,* **1970**, 283-319.

173.Harshman, R.A. Foundations of the PARAFAC procedure: models and conditions for an explanatory multimodal factor analysis. *UCLA Working Papers on Phonetics, 16,* **1970**, 1-84.

174.Bro, R. Multiway calibration. Multilinear PLS. *J. of Chemometrics, 10,* **1996**, 47-61.

175.De Jong, S.; Kiers, H.A.L. Principal covariates regression Part I: theory. *Chemom. and Intell. Lab. Syst., 14,* **1992**, 155-164.

176.Kshirsagar, A.M. Multivariate Analysis. Marcel Dekker: New York, **1972.**

177.Cuadras, C.M.; Arenas, C. A distance based regression model for prediction with mixed data. *Commun. Statist. Theor. Meth., 19,* **1990**, 2261-2279.

178.Cuadras, C.M.; Arenas, C.; Fortiana, J. Some computational aspects of a distance-based model for prediction. *Commun. Statist. Simula*., *25,* **1996**, 593-609.

179.Besalú, E.; Vera, L. On the optimal selection of principal components in QSPR studies. *J. Math. Chem., 1,* **2001**, 21-34.

180.Livingstone, D.J.; Rahr, E. Corchop - An interactive routine for the dimension reduction of large QSAR data sets. *Quant. Struct.-Act. Relat., 8,* **1989**, 103-108.

181.Carbó, R.; Besalú E. Nested Summation Symbols and Perturbation Theory. *J. Math. Chem., 13,* **1993**, 331-342.

182.Carbó, R.; Besalú, E. Definition, mathematical examples and quantum chemical applications of nested summation symbols and logical Kronecker deltas. *Comput. Chem., 18,* **1994**, 117-126.

183.Carbó, R.; Besalú, E. Definition and quantum chemical applications of nested summations symbols and logical functions: pedagogical artificial intelligence devices for formulae writing sequential programming and automatic parallel implementation. *J. Math. Chem., 18,* **1995**, 37-72.

184.Carbó, R.; Besalú, E. Application of Nested Summation Symbols to Quantum Chemistry: Formalism and Programming Techniques. In *Strategies and Applications in Quantum Chemistry: from Astrophysics to Molecular Engineering.* Defranceschi, M.; Ellinger, Y. (Eds.). Kluwer Academic Publishers: Amsterdam, **1996**, 229-248.

185.Shi, L.M.; Fan, Y.; Myers, T.G.; O'Connor, P.M.; Paull, K.D.; Friend, S.H.; Weinstein, J.N. Mining the NCI anticancer drug discovery databases: genetic function approximation to QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci., 39,* **1998**, 189-199.

186.Partek Pro 2000. Partek Incorporated, Partek analysis and recognition technologies, **1993-1999**.

187.Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-based PLS analysis of calcium channel antagonists. *J. Chem. Inf. Comput. Sci., 37*, **1997**, 306-310.

188.Hasegawa, K.; Miyashita, Y.; Funatsu, K. GA strategy for variable selection in QSAR studies: GA-PLS and D-optimal designs for predictive QSAR model. *J. Mol. Struct. (Teochem), 425,* **1998**, 255-262.

189.Cho, S.J.; Zheng, W.; Tropsha, A. Rational combinatory library design. 2. Rational design of targeted combinatorial peptide libraries using chemical similarity prove and the inverse QSAR approaches. *J. Chem. Inf. Comput. Sci., 38,* **1998**, 259-268.

190.Zheng, W.; Tropsha, A. Novel variable selection quantitative structure-property relationships approach based on the k-nearest neighbor principle. *J. Chem. Inf. Comput. Sci., 40,* **2000**, 185-194.

191.So, S.S.; Karplus, M. Genetic neural networks for quantitative structure-property relationships: improvement and application of benzodiazepine affinity for benzodiazepine/GABA A receptor. *J. Med. Chem., 39,* **1996**, 5246-5256.

192.Wessel, M.D.; Jurs, P.C.; Tolan, J.W.; Muskal, S.M. Prediction of human intestinal absorption of drug compounds from molecular structure. *J. Chem. Inf. Comput. Sci., 38,* **1998**, 726-735.

193.Edwards, A.L. *An introduction to linear regression and correlation.* W. H. Freeman and Company: New York, **1984**.

194.Pecka, J.; Ponec, R. Simple Analytical Method for Evaluation of Statistical Importance of Correlations in QSAR Studies. *J. Math. Chem*., *23*, **2000,** 13-22.

195.Allen, D.M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics, 16*, **1974**, 125-127.

196.Wold, S. Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics, 20,* **1978**, 397-405.

197.Wold, S. Validation of QSARs. *Quant. Struct.-Act. Relat., 10,* **1991**, 191-193.

198.Allen, D.M. The prediction sum of squares as a criterion for selecting variables. Technical report 23, Department of Statistics, University of Kentucky, **1971**.

199.Besalú, E. Fast computation of cross-validated properties in full linear leave-many-out procedures. *J. Math. Chem*., *29,* **2001**, 191-204.

200.Wold, S.; Eriksson, L. Statistical validation of QSAR results. In *Chemometric methods in molecular design*. van de Waterbeemd, H. (Ed.). VCH: New York, **1995,** 309-318.

201.Topliss, J.G.; Costello, R.J. Chance correlations in structure-activity studies using multiple regression analysis. *J. Med. Chem., 15,* **1972**, 1066-1068.

202.Besalú, E.; Vera, L. Internal Test Sets (ITS) Method: A new cross-validation technique to assess the predictive capability of QSAR models. *SAR QSAR Environ. Res*.; in press.

203.Hartigan, J.A. *Clustering algorithms*. Wiley: New York, **1975**.

204.Kaufman, L.; Rousseeuw, R. Finding groups in data: an introduction to cluster analysis. Wiley: New York, **1990.**

205.Kruskal, J.B.; Wish, M. *Multidimensional Scaling*. Sage: Beverly Hills, **1978.**

206.Borg, I.; Shye, S. *Facet Theory: form and content*. Sage: Newbury Park, **1995.**

207.Besalú, E. Un algorisme per al reconeixement automatitzat de grafs dicotomitzats. *Scientia Gerundensis, 20,* **1994**, 87-93.

208.Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. TQSARSIM. Institut de Química Computacional, **1997.**

209.Amat, L.; Robert, D.; Besalú, E.; Carbó-Dorca, R. Molecular Quantum Similarity Measures Tuned QSAR: An Antitumoral Family Validation Study. *J. Chem. Inf. Comput. Sci., 38,* **1998**, 624-631.

Applications of Quantum Similarity Measures to QSAR

The tangible, the real, the solid,
is explained by
the intangible, the unreal, the purely mental.
Yet that is what we chemists are always doing,
wave-mechanically or otherwise.

J. chem. Soc. 1955, 2096
C.A. Coulson

1 INTRODUCTION

In this chapter, several application examples of the previously exposed methodology are presented in order to illustrate the feasibility of the implementation of theoretical abstract principles and methods to the resolution of practical issues in chemistry-based applications.

On one hand, in connection with different methods to obtain structural descriptors for molecules, this section presents some cases using **Quantum Similarity Indices (QSI)**, obtained from Quantum Similarity Theory (QST). Molecular descriptors are expressed by means of Quantum Self-Similarity Measures (QS-SM), Overlap MQSM, Coulomb MQSM, and Kinetic MQSM, whereas descriptors for particular functional groups are defined using Fragment Quantum Self-Similarity Measures (QS-SM). Besides, several studies dealing with different **three-dimensional topological quantum similarity indices** (3D-TQSI) derived from graph theory are discussed.

On the other hand, attending to the classification of the specific action of studied molecular families, this chapter has been organised in relation to the type of function exerted. On one hand, the studies related to molecular toxicities have been compiled in a section devoted to **Quantitative Structure-Toxicity Relationships (QSTR)**, i.e. estimation of the percutaneous absorption of carcinogenic compounds, and study of the aquatic toxicity of environmental pollutants. On the other hand, the cases dealing with biological acivities of pharmacological interest, such as antimalarial, estrogenic, and antituberculotic activities, have been gathered in the **Quantitative Structure-Activity Relationships (QSAR)** section. Sometimes, a particular molecular set has been studied by using different methods.

In addition, studies can be **qualitative**, **quantitative** or **semi-quantitative**, depending on the type and expression of activity values. Concerning the purposes, studies can be intended for drug discovery, environmental risk or health hazard assessment, etc. For all the models, the results are displayed with the corresponding statistical parameters that account for their quality, accuracy and reliability.

In the present work, the protocol for the development of a QSAR model with predictive capabilities for any study case follows the following steps:

1) **Compilation of the information** related to the molecular set and the experimental properties or activities of study, extracted from the literature. At this point, it must be remarked that experimental data should be carefully handled, regarding not only the biological meaning, but also paying attention to the accomplishment of the first SETUBAL principle [1], and adequately referencing the bibliographic sources.

2) **Molecular modelling** of compounds, using graphic assistants for the edition and initial cleaning of structures. In this section, structures for molecular data sets have been displayed in tables in order to provide a graphical insight and to picture the molecules. In this work, the software used for this purposes is WebLab ViewerPro [2], and HyperChem [3].

3) **Geometry optimisation** for each molecule of the family at different levels of theory. According to the complexity of the method, there are several possibilities, i.e. molecular mechanics force field, approximate semi-empirical methods, *ab initio* optimisation level, etc. Due to the size of the studied molecules and the specific purposes, it has been demonstrated that in most cases an optimisation at the semi-empirical AM1 level [4] carried out with Ampac [5] or Mopac [6] packages, is enough to perform the comparison between density functions. Even in some particular studies the Sybil molecular mechanics force field calculated with PC-Spartan [7] is sufficient.

4) From the optimised molecular geometry, the **calculation of the approximated Density Function (DF)** is performed. As exposed in the methodological section, usually the first-order Atomic Shell Approximation (ASA) DF fitted to a 3-21G basis set is sufficiently accurate [8-9].

5) **Molecular alignment** may be required for the calculation of descriptors. In this thesis, the superposition process has been carried out by using programs entirely developed in the laboratory, i.e. the Maximum Similarity Superposition Algorithm [10], and the Topo-Geometrical Superposition Algorithm (TGSA) [11].

6) **Computation of descriptors**, employing programs that read the output files of the optimization program, and define the type of atoms, exponents, *xyz* coordinates, effective atomic charges, topological matrices, etc. The programs also involve the gathering of molecular indices in matrices of descriptors, where each row represents a molecule, whether each column symbolises an index. As remarked before, two different classes of indices have been used, namely Quantum Similarity Indices (QSI) and Topological Quantum Similarity Indices (TQSI). QSI have been calculated using either the first version [12] or the graphical interface version of MOLSIMIL [13], developed by Amat et al., while TQSI have been computed with TOPO [14]. In addition, in one study, TSAR [15] was used for the authomatic generation of indicator variables and other structural descriptors.

7) **Selection of descriptors and building of the statistical model**. In this phase, besides from computational details of the statistical protocol, i.e. correlation, dimensions reduction, and variables selection methods, relevant data such as the number of molecules, number of parameters, the equation, and other statistical parameters accounting for the quality of adjustment and goodness-of-fit of the model must be reported in order to achieve a transparent, reproducible model. Also, statistical plots can provide valuable graphical and intuitive information. In most cases, self-developed software by the Institute has been used [16-17], although in specific studies commercial statistical packages such as MINITAB [18] have been employed.

8) **Validation of the model**, by means of internal validation, and, if there are further available data, external validation techniques. In this project the Leave-One-Out Cross-Validation (LOO-CV) process, along with random tests have been carried out and, in some cases, external test sets of untested molecules have been used for further predictions. Such validation techniques have been used using the programs aforementioned, in the previous point.

9) Finally, the last but not less important step is involved with the **interpretation of the model**, and the deduction of conclusions, which should provide chemical insight into the poblem.

2 QUANTITATIVE STRUCTURE–TOXICITY RELATIONSHIPS

2.1 Dermal penetration of Polycyclic Aromatic Hydrocarbons

Introduction

Polycyclic aromatic hydrocarbons (PAHs) are a group of over hundred different chemicals, formed during the incomplete burning of coal [19], oil and gas [20], garbage [21], or other organic substances like tobacco or charbroiled meat [22]. PAHs are usually found in the environment [23] as a mixture containing two or more of these compounds, such as soot. The exposure to polycyclic aromatic hydrocarbons usually occurs by breathing air contaminated by wild fires or coal tar, or by eating foods that have been grilled.

Existing an early documented evidence [24**]** of their mutagenic and carcinogenic properties, as a result of their impact in human health, the evaluation of dermal penetration is of great importance in hazard assessment programs [25-28]. Thus, quantitative structure-activity relationships (QSAR) have been constructed to estimate PAHs carcinogenic power, using different empirical and theoretical methods [29-34].

The study is focused on the application of similarity matrices to the study of the carcinogenic power of two different sets of commercially available PAHs, for which QSAR models were constructed. Two reported properties, discrete levels of carcinogenic activity [35] in the first example, and *in vitro* percutaneous absorption in rat skin [36], in the second one, have been examined for correlation.

2.1.1 Semiquantitative Classical Study of 78 PAHs

In the first application example, the molecular set was made of 78 PAHs, divided into two subsets responding to structural criteria, that is, the presence or absence of methyl substitutions. The splitting of the whole set resulted in two subsets of 32 non-methylated and 46 methylated PAHs.

Provided that the sources did not collect experimental data obtained in homologous conditions, a discrete classification into classes, attending to the carcinogenic power, was adopted [37-41]. Thus, the property discretization leads to a semiquantitative study, which only discriminates between active (A) and inactive (I) compounds.

Table 1. Structures and carcinogenic activities for 78 PAHs. A: Active; I: Inactive; NA: not available.

Non-methylated PAHs

N^o	Compound	Act	N^o	Compound	Act
1	Dibenzo[3,4:9,10]pyrene	A	17	Benzo[1,2]pyrene	I
2	Benzo[3,4]pyrene	A	18	Phenanthrene	I
3	Dibenzo[3,4:8,9]pyrene	A	19	Triphenylene	I
4	Dibenzo[3,4:6,7]pyrene	A	20	Benzo[1,2]naphthacene	I
5	Dibenzo[1,2:3,4]pyrene	A	21	Dibenzo[3,4:5,6]phenanthrene	I
6	Naphto[2,3:3,4]pyrene	A	22	Picene	I
7	Dibenzo[1,2:5,6]anthracene	A	23	Tribenzo[1,2:3,4:5,6]anthracene	I
8	Tribenzo[3,4:6,7:8,9]pyrene	A	24	Dibenzo[1,2:5,6]pyrene	I
9	Dibenzo[1,2:3,4]phenanthrene	A	25	Phenanthra[2,3:1,2]anthracene	I
10	Tribenzo[3,4:6,7:9,10]pyrene	A	26	Benzo[1,2]pentacene	I
11	Dibenzo[1,2:5,6]phenanthrene	I	27	Anthanthrene	I
12	Benzo[1,2]anthracene	I	28	Benzene	I
13	Chrysene	I	29	Naphtalene	I
14	Benzo[3,4]phenanthrene	I	30	Pyrene	I
15	Dibenzo[1,2:7,8]anthracene	I	31	Benzo[*ghi*]perylene	I
16	Dibenzo[1,2:3,4]anthracene	I	32	Coronene	I

Methylated PAHs

N^o	Compound	Act	N^o	Compound	Act
33	7,12-dimethylbenz[*a*]anthracene	A	58	3-methylbenzo[*c*]phenanthrene	I
34	6,12-dimethylbenz[*a*]anthracene	A	59	6-methylbenzo[*c*]phenanthrene	I
35	6,8,12-trimethylbenz[*a*]anthracene	A	60	6-methylbenz[*a*]anthracene	I
36	2-methylbenzo[*a*]pyrene	A	61	12-methylbenz[*a*]anthracene	I
37	4-methylbenzo[*a*]pyrene	A	62	6-methylanthanthrene	I
38	11-methylbenzo[*a*]pyrene	A	63	6,12-dimethylanthanthrene	I
39	12-methylbenzo[*a*]pyrene	A	64	1-methylbenzo[*c*]phenanthrene	I
40	1-methylbenzo[*a*]pyrene	A	65	2-methylbenzo[*c*]phenanthrene	I
41	4,5-dimethylbenzo[*a*]pyrene	A	66	10-methylbenzo[*a*]pyrene	I
42	3-methylbenzo[*a*]pyrene	A	67	6-methylchrysene	I
43	1,2-dimethylbenzo[*a*]pyrene	A	68	3-methylbenz[*a*]anthracene	I
44	2,3-dimethylbenzo[*a*]pyrene	A	69	1-methylbenz[*a*]anthracene	I
45	3,12-dimethylbenzo[*a*]pyrene	A	70	11-methylbenz[*a*]anthracene	I
46	1,3-dimethylbenzo[*a*]pyrene	A	71	9-methylbenz[*a*]anthracene	I

Methylated PAHs					
N^o	Compound	Act	N^o	Compound	Act
47	1,4-dimethylbenzo[*a*]pyrene	A	72	2-methylbenz[*a*]anthracene	I
48	5-methylbenzo[*c*]phenanthrene	A	73	5-methylbenz[*a*]anthracene	I
49	5-methylchrysene	A	74	8-methylbenz[*a*]anthracene	I
50	6,8-dimethylbenz[*a*]anthracene	A	75	2-methylpyrene	I
51	7-methylbenz[*a*]anthracene	A	76	4-methylpyrene	I
52	5-methylbenzo[*a*]pyrene	A	77	1-methylpyrene	I
53	7-methylbenzo[*a*]pyrene	A	78	7,10-dimethylbenzo[*a*]pyrene	I
54	6-methylbenzo[*a*]pyrene	A	79	6,10-dimethylbenzo[*a*]pyrene	NA
55	1,6-dimethylbenzo[*a*]pyrene	A	80	8-methylbenzo[*a*]pyrene	NA
56	3,6-dimethylbenzo[*a*]pyrene	A	81	9-methylbenzo[*a*]pyrene	NA
57	4-methylbenzo[*c*]phenanthrene	I			

Table 2. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Polycyclic Aromatic Hydrocarbons (PAHs)
Type of Compounds	32 non-methylated PAHs and 46 methylated PAHs
Number of Compounds	78
Activity	Binary activities {0,1}
Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	Semiempirical AM1 level [4], using Ampac-6.55 [5]
Density Function	Fitted first-order Promolecular ASA (PASA), 3-21G basis set
MQSM Operator	Overlap operator
Molecular Alignment	Maximum similarity superposition algorithm
Reduction of Dimensions	Principal Components Analysis (PCA)
Selection of Variables	Most Predictive Variables Method (MPVM)
Correlation Method	Multiple Linear Regression (MLR)
Validation	Internal Leave-One-Out Cross-Validation (LOO-CV) External test set

Results and discussion

Table 3. Results of the optimal QSAR model for the non-methylated PAHs subset.

32 non-methylated PAHs subset				
	Adjustment		**Cross-Validation**	
Number of PCs	**% Correct Classification**	**% Correct classification for carcinogenic compounds**	**% Correct Classification**	**% Correct classification for carcinogenic compounds**
4	93.8	90.0	84.4	80.0
Missclassified molecules: (2/32) ⇔ 7,22				

Table 4. Results of the optimal QSAR model for the methylated PAHs subset.

46 methylated PAHs subset				
	Adjustment		**Cross-Validation**	
Number of PCs	**% Correct Cl assification**	**% Correct classification for carcinogenic compounds**	**% Correct Classification**	**% Correct classification for carcinogenic compounds**
5	87.0	83.3	84.8	79.2
Missclassified molecules: (6/46) ⇔ 33,48,50,51,61,67				

Table 5. Results of the optimal QSAR model for the entire PAHs set.

78 entire PAHs set				
	Adjustment		**Cross-Validation**	
Number of PCs	**% Correct Classification**	**% Correct classification for carcinogenic compounds**	**% Correct Classification**	**% Correct classification for carcinogenic compounds**
3	82.1	70.6	80.8	70.6
Missclassified molecules: (14/78) ⇔ 4,7,13,22,25,33,34,35,38,48,49,50,51,67				
Missclassified molecules correctly predicted in their subset: 4,13,25,34,35,38,49				

In this particular case, integer binary numbers {0,1} were arbitrarily assigned to the active and inactive classes, respectively, and thereafter the multilinear regression was carried out. The adjusted and cross-validated activities were classified into a category attending to a pre pre-established threshold of $r^2=0.5$. Besides, instead of the continuous r^2 and q^2 coefficients, the quality of the models was assessed by means of the percentage of correct classifications, and the percentage of correctly classified carcinogenic compounds.

The common backbone, made of fused benzenes, without the presence of 5-member rings or bonds connecting rings, facilitated exact atom-atom superpositions, more favorable for overlap MQSM. For all the subsets valuable semi-quantitative SAR models were obtained. Indeed, a great number of misclassified compounds were found to be misclassified by other methods. In the case of the non-methylated and the methylated subsets, comparable results were found with four and five principal components for the optimal model, respectively. The general trend is a slight decrease in the percentages of the cross-validated results.

However, for the entire set, there is a significative difference of the percentage of correct classification between the totality of compounds and the carcinogenic ones. This evidences that the model concentrates the misclassifications in the active compounds. Besides, the lessening in the predictivity is expected because of the structural heterogeneity of the molecular set, negatively influencing similarity-based QSAR approaches. The same set had been studied with different techniques, i.e., electronic index methodology (EIM), principal component analysis (PCA), and neural networks (NN), based on the local density of states (LDOS) theory [42-45]. Comparisons with the different studies show that the MQSM improve the description of the system, indicating that a global density approach encodes relevant information for the characterization of PAHs carcinogenicity.

Table 6. Comparison among different QSAR methodologies. NM: non-methylated PAHs; M: methylated PAHs.

Method	32 NM PAHS	46 M PAHS	Full set of 78 PAHS
EIM	84.4	73.9	78.2
PCA	84.4	78.3	80.8
NN	93.8	78.3	84.6
MQSM	93.8	87.0	82.1

Finally, three compounds without experimentally measured activity were predicted using MQSM. Two of them were found to be unequivocally active, in agreement with the previously reported assignments.

2.1.2 Quantitative Classical Study of 60 PAHs

In the second case, a quantitative study of continuous data was performed. The molecular set consisted of 60 polycyclic aromatic hydrocarbons, made up of three to seven fused aromatic rings, in a particularly rigid conformation.

The toxicity, i.e. the percutaneous absorption, measured *in vitro* on rat skin sections, was expressed as the percentage of applied dose (PADA) penetrating the skin following the protocol of application [35,46-48]. The activity ranges from 0,7 to 50% for the very active and inactive compounds, respectively.

It has been found that dermal penetration of PAHs with 1 or 2 aromatic rings was difficult to measure because of the volatility and loss from the skin surface during the biological essays. For this reason, interest was primarily focused on the carcinogenic PAHs, mainly comprised within 4 and 6 ring structures.

Table 7. Molecular data set: structures for the 60 Polycyclic Aromatic Hydrocarbons

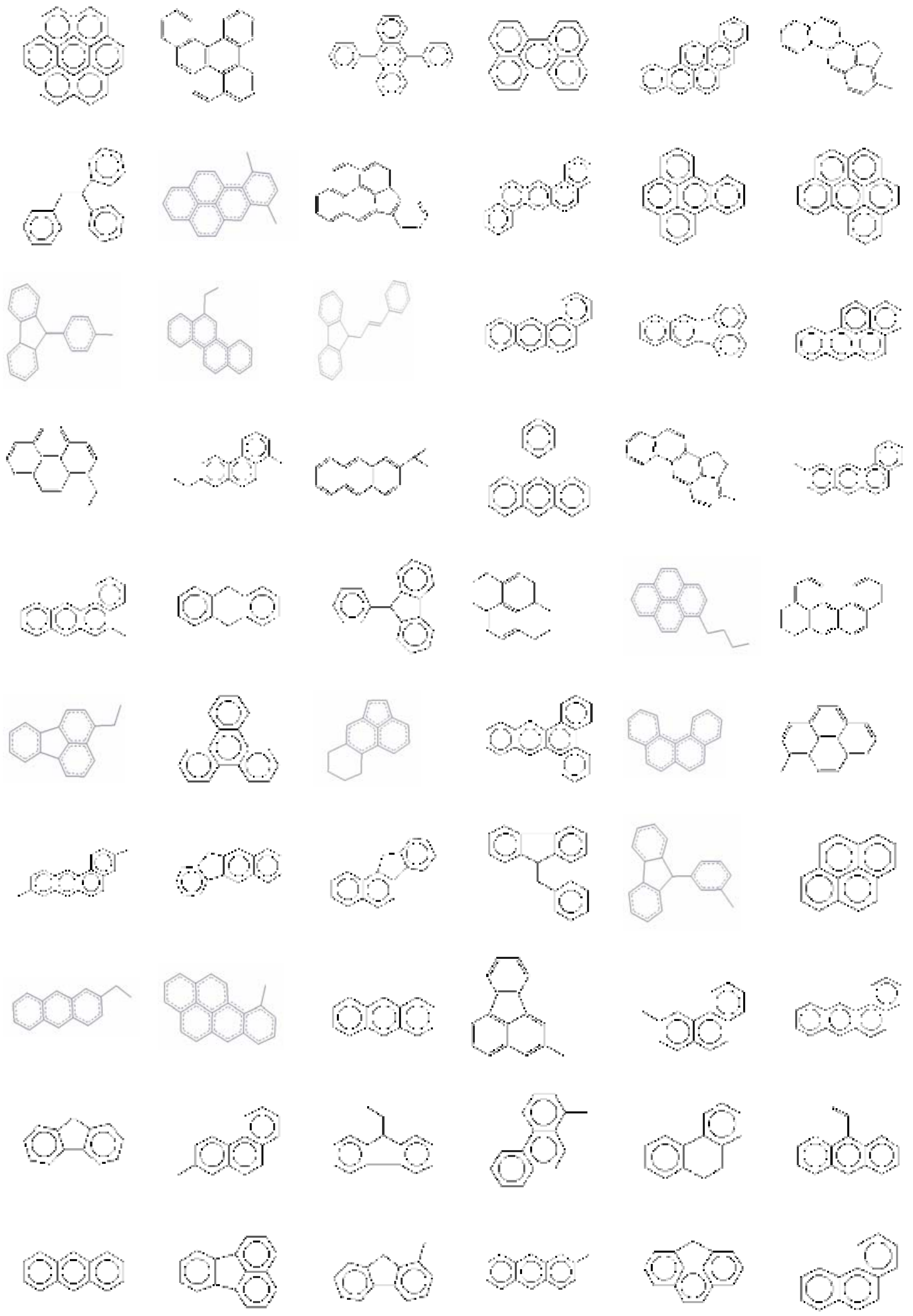


Table 8. Structures and dermal penetration (PADA) values (%) for 60 PAHs.

N°	Compound	%PADA	N°	Compound	%PADA
1	Coronene	0.7	31	3-ethylfluoranthene	20
2	Dibenzo[a,l]pyrene	2	32	Triphenylene	20
3	9,10-diphenylanthracene	6	33	7,8,9,10-tetrahydroacephenanthrene	20
4	Perylene	7	34	2,3-benztriphenylene	20
5	Dibenzo[*a*,*i*]pyrene	8	35	Benzo[*c*]phenanthrene	20
6	3-methylcholanthrene	8	36	1-methylpyrene	22
7	9-benzylidenefluorene	8	37	3,9-dimethylbenz[*a*]anthracene	24
8	7,10-dimethylbenzo(*a*)pyrene	8.3	38	2,3-benzofluorene	25
9	Indeno(1,2,3-*cd*)pyrene	9	39	1,2-benzofluorene	25
10	Dibenz[*a*,*h*]anthracene	9.4	40	9-benzylfluorene	26
11	Benzo[*e*]pyrene	10	41	9-m-tolylfluorene	29
12	Benzo[*g*,*h*,*i*]perylene	10	42	Pyrene	30
13	9-p-tolylfluorene	10	43	2-ethylanthracene	30
14	6-ethylchrysene	10	44	10-methylbenzo[*a*]pyrene	32
15	9-cynnamylfluorene	11	45	1-methylanthracene	33
16	6- methylbenz[*a*]anthracene	14	46	2-methylfluoranthene	33
17	Benzo[*k*]fluoranthene	14	47	3,6-dimethylphenanthrene	33
18	Benzo[*a*]pyrene	15	48	Benzo[*a*]anthracene	35
19	3-ethylpyrene	18	49	Fluorene	36
20	1-methyl-7-isopropylphenanthrene	20	50	2-methylphenanthrene	38
21	2-(tert-butyl)anthracene	20	51	9-ethylfluorene	38
22	9-phenylanthracene	20	52	1-methylphenanthrene	40
23	3-methylcholanthrene	20	53	9,10-dihydrophenanthrene	40
24	10-methylbenz[*a*]anthracene	20	54	9-vinylanthracene	40
25	5-methylbenz[*a*]anthracene	20	55	Anthracene	42
26	9,10-dihydroanthracene	20	56	Fluoranthene	42
27	9-phenylfluorene	20	57	1-methylfluorene	49
28	1,2,3,6,7,8-hexahydropyrene	20	58	2-methylanthracene	50
29	n-butylpyrene	20	59	4H-cyclopenta(*d*,*e*,*f*)phenanthrene	50
30	5,6-dihydro-4H-dibenz[*a*,*k*,*l*]anthracene	20	60	Phenanthrene	50

At first sight, the carcinogenicity is primarily concentrated on the structures comprising from four to six aromatic rings, according to the postulates/assumptions of the K-L-M “bay region” theory [49-54].

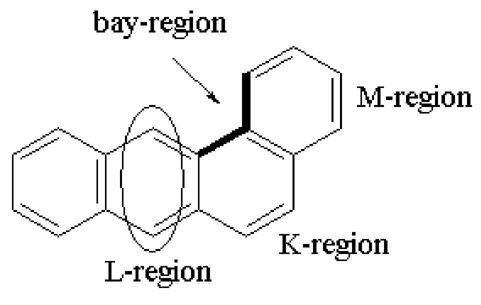


Figure 1. The K, L, M and bay regions of a polycyclic aromatic hydrocarbon: benzo[a]anthracene.

Table 9. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Polycyclic Aromatic Hydrocarbons (PAHs)
Type of Compounds	3 to 7 fused aromatic rings
Number of Compounds	60
Activity	% of penetrating applied dose (PADA)
Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	Semiempirical AM1 level [4], using Ampac-6.55 [5]
Density Function	Fitted first-order Promolecular ASA (PASA), 3-21G basis set
MQSM Operator	Coulomb operator
Molecular Alignment	Maximum similarity superposition algorithm
Reduction of Dimensions	Principal Components Analysis (PCA)
Selection of Variables	Most Predictive Variables Method (MPVM)
Correlation Method	Multiple Linear Regression (MLR)
Validation	Internal Leave-One-Out Cross-Validation (LOO-CV) Random Test

Results and discussion

Concerning the choice of the operator, the presence of five-membered rings and internal single bonds hinders an exact intermolecular atom-atom matching, so that the similarity contribution of the Overlap operator is very low. Therefore, the Coulomb operator has been chosen.

Table 10. Optimal QSAR model for the dermal penetration of 60 PAHs (3 PCs).

No PCs	Selected PCs	r^2	q^2
3	1, 2, 13	0.684	0.634
Equation	$\mathbf{y} = -1.263\mathbf{x}_1 + 0.489\mathbf{x}_2 + 0.914\mathbf{x}_{13} + 23.173$		

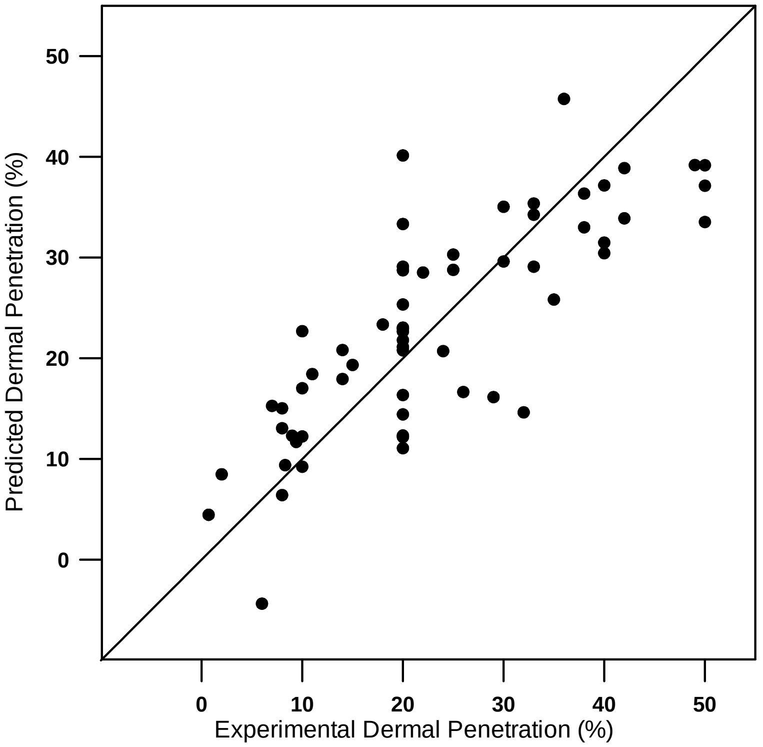


Figure 2. Predicted vs. Experimental. Cross-validated vs experimental percutaneous absorption values.

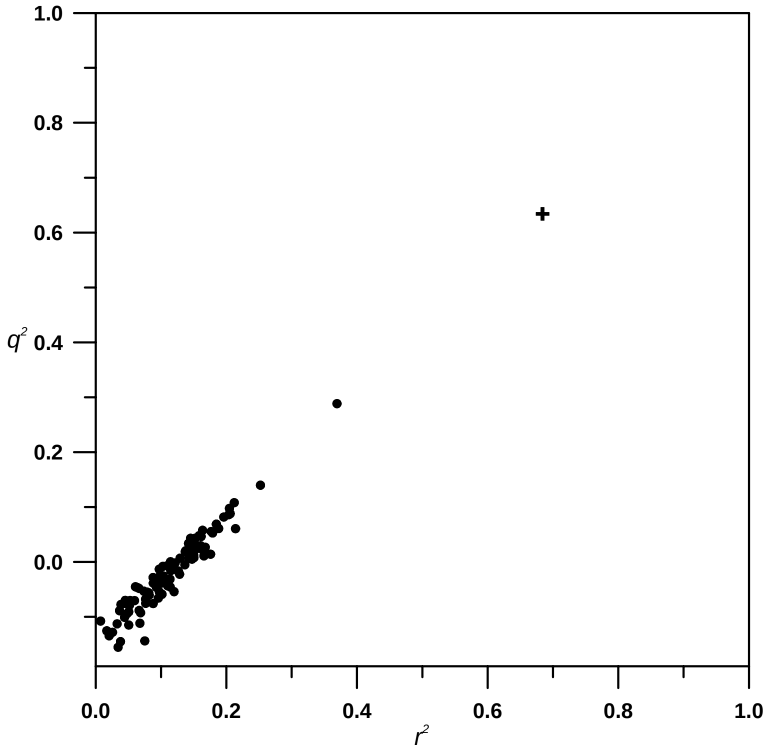


Figure 3. Randomization test for the optimal model. The randomized responses (100) have been marked with circles, and the correctly ordered activity has been marked with a cross. (r^2_{cv}/q^2 vs r^2)

Most of the experimental measurements of dermal penetration are expressed in integer percentages, preventing from the estimation of slight differences in the activity. Indeed, 16 compounds presenting exactly the same PADA value have the highest residuals in the model, which might result in a decrease of the quality of the prediction model.

However, in the random test examination, the clear separation between real data and the random ones ensures a reliable structure-property relationship. Besides, comparisons with other QSAR approaches show that the application of MQSM to QSAR produces comparable results [55-58].

Appended Contribution

The results of Sections 2.1.1 and 2.1.2 have been gathered in the following contribution, which has been appended in the annex.

►Gallegos, A.; Robert, D.; Gironés, X.; Carbó-Dorca, R. Structure-Toxicity Relationships of Polycyclic Aromatic Hydrocarbons using Molecular Quantum Similarity. *J. Comput.-Aid. Mol. Des., 15(1),* **2001**, 67-80.

In order to test the application of topological quantum similarity measures to QSAR, topological similarity-based descriptors were computed for three different families, which exhibit different kinds of toxicity.

In the first case (Section 2.1.3), the dermal penetration of the set of 60 commercially available polycyclic aromatic hydrocarbons studied in the preceding section (Section 2.1.2) was correlated with the computed TQSI. Thereafter, the Inhibitory Growth Concentration (IGC) of two families, one composed by 30 aliphatic alcohols and amines and, the other, by 48 selected anilines, was also studied, in Sections 2.2.1, and 2.2.2, respectively.

For the calculation of the so-called Topological Quantum Similarity Indices (TQSI), it must be reminded that the classical construction according to the theoretical graph theory framework, has been used, but replacing the classical topological matrices by matrices derived from Quantum Similarity (QS) calculations, due to the connection between chemical graph theory and quantum similarity. TQSI, which also account for further three-dimensional information, have been computed using the program TOPO [14]. Concretely, the traditionally defined integer topological matrix has been substituted by the interatomic Quantum Similarity Measure with a similarity weight operator, which is calculated between each pair of atoms of a given molecule. Similarly, the valence vector also has been computed from the entries of the similarity matrix. In addition to the use of QSM, the topological distance has also been replaced by the three-dimensional Euclidean distance between every pair of atoms. Thus, the obtained three-dimensional indices obtained with the real matrix include spatial information.

2.1.3 Topological Study of 60 PAH

2.1.4

Table 11. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Polycyclic Aromatic Hydrocarbons (PAHs)
Type of Compounds	3 to 7 fused aromatic rings
Number of Compounds	60
Activity	% of penetrating applied dose (PADA)
Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	Semiempirical AM1 level [4], using Ampac-6.55 [5]
Density Function	Fitted first-order ASA DF, 3-21G basis set
MQSM Operator	Coulomb interatomic QSM
Molecular Alignment	Not needed
Correlation Method	Partial Least Squares (PLS)
Validation	Leave-One-Out Cross Validation (LOO-CV)

Results and discussion

Table 12. Statistical results obtained in the optimal model.

Total N^{o} of Indices	**N^{0} Descriptors**	r^2	q^2
13	5	0.6940	0.6524

In this case, after computing a total of 13 indices including the above mentioned up to third order, the best model was chosen as the one with 5 descriptors. The optimal model, obtaining the values of r^2=0.694 and $q^2=0.652$ yielded satisfactory results.

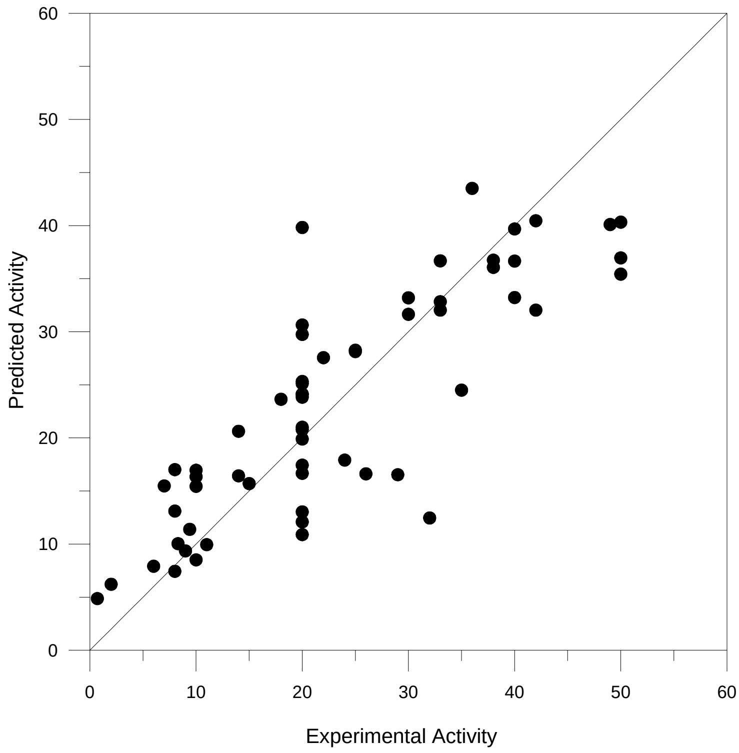


Figure 4. Cross-validated versus experimental percutaneous absorption values.

As the plot evidences, as in the previous study a possible cause for the poor results in correlation and prediction of this set can be due to the peculiar distribution of the experimental data. In most cases, the measurement of the dermal penetration is expressed by an integer number, precluding the appreciation of slight differences within the studied activity. For example, the 16 molecules with exactly the same PADA value of 20%, present the highest residuals in the model.

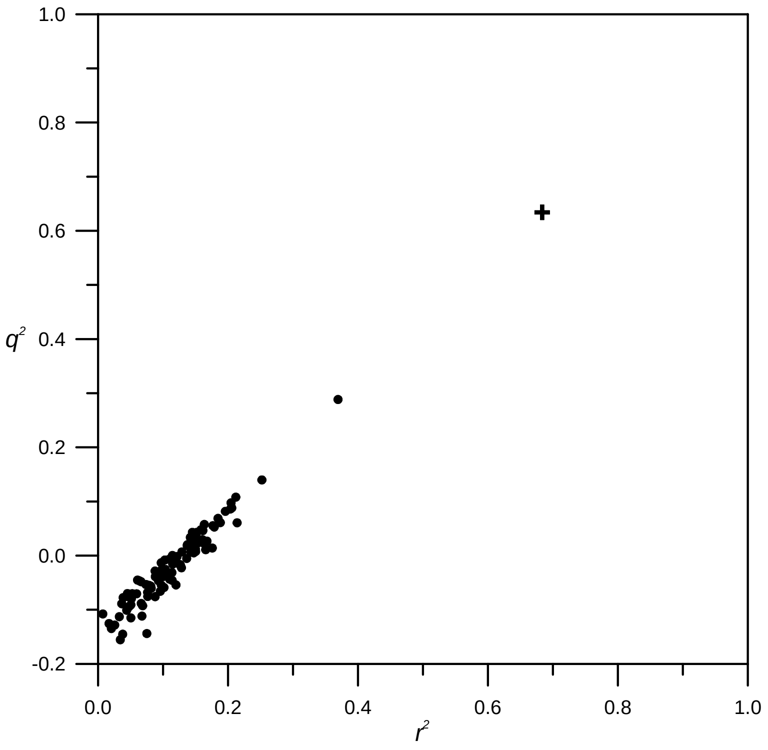


Figure 5. Randomization test for the optimal model. The randomized responses (100) have been marked with circles, and the correctly ordered activity have been marked with a cross in bold face.

The same series of PAHs had also been previously studied using only Quantum Similarity Theory, employing overlap and Coulomb measures. When analyzing the results, it was observed that the quality of the models with Coulomb MQSM was notably better than those built employing overlap MQSM, because of the particular structure of the studied set, which did not allow an exact intermolecular atom-atom matching.

2.2 Aquatic toxicity of Environmental Pollutants

2.2.1 Topological Study of 30 aliphatic alcohols

The second application example of TQSI was constituted by 30 aliphatic alcohols and amines [60]. Toxicity tests were performed using the Tetrahymena pyriformis population growth assay, which includes several parameters, such as the initial pH, the temperature, the shape and the volume of the culture system, the amount of medium and the age and volume of inoculums. Further references concerning the method can be found in some reviews by Holz, Cameron and Levy [61-63].

For each toxicant, the IGC_{50} (50% inhibitory growth concentration) was determined for three replicates at different concentration, having into account that specific absorbance and concentration are directly proportional.

QSARs were examined using the logarithm of the inverse of the IGC_{50}. The structures and the activity values of the first molecular set, the family composed by 30 aliphatic alcohols and amines, are shown in Table 13.

Table 13. Relative toxicity for 30 aliphatic alcohols and amines.

Nº	Compound	$(LogIGC_{50})^{-1}$	Nº	Compound	$(LogIGC_{50})^{-1}$
1	Methanol	-2.77	16	3-pentanol	-1.33
2	Ethanol	-2.41	17	2-methyl-1-butanol	-1.13
3	1-propanol	-1.84	18	3-methyl-1-butanol	-1.13
4	1-butanol	-1.52	19	3-methyl-2-butanol	-1.08
5	1-pentanol	-1.12	20	(tert)pentanol	-1.27
6	1-hexanol	-0.47	21	(neo)pentanol	-0.96
7	1-heptanol	0.02	22	1-propylamine	-0.85
8	1-octanol	0.5	23	1-butylamine	-0.7
9	1-nonaol	0.77	24	1-maylamine	-0.61
10	1-decanol	1.1	25	1-hexylamine	-0.34
11	1-undecanol	1.87	26	1-heptylamine	0.1
12	1-dodecanol	2.07	27	1-octylamine	0.51
13	1-tridecanol	2.28	28	1-nonylamine	1.59
14	2-propanol	-1.99	29	1-decylamine	1.95
15	2-pentanol	-1.25	30	1-unidecylamine	2.26

Table 14. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Aromatics
Type of Compounds	Aliphatic alcohols and anilines
Number of Compounds	30
Activity	$(\text{LogIGC}_{50})^{-1}$ (IGC : Inhibitory Growth Concentration)
Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	Semiempirical AM1 level [4], using Ampac-6.55 [5]
Density Function	Fitted first-order ASA DF, 3-21G basis set
MQSM Operator	Coulomb interatomic QSM
Molecular Alignment	Not needed
Correlation Method	Partial Least Squares (PLS)
Validation	Leave-One-Out Cross Validation (LOO-CV)

Results and discussion

Table 15. Statistical results obtained in the optimal model.

Total N^o of Indices	**N^0 Descriptors**	r^2	q^2
13	4	0.8875	0.8587

In this case, the model described with 4 parameters was chosen as the optimal. As can be easily regarded, good results were obtained, with coefficient values above *0.8*, which achieve significant results.

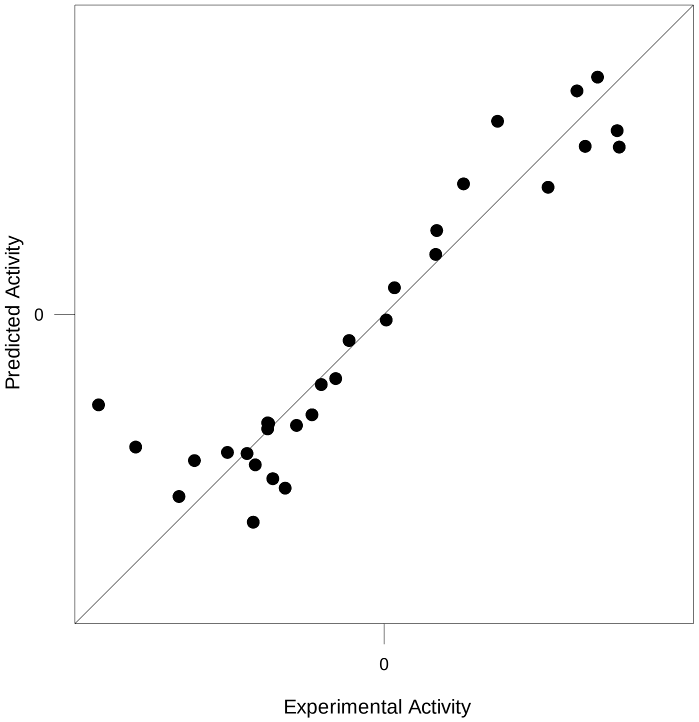


Figure 6. Cross-validated versus experimental inhibition growth concentration values, for the set of 30 aliphatic alcohols and amines.

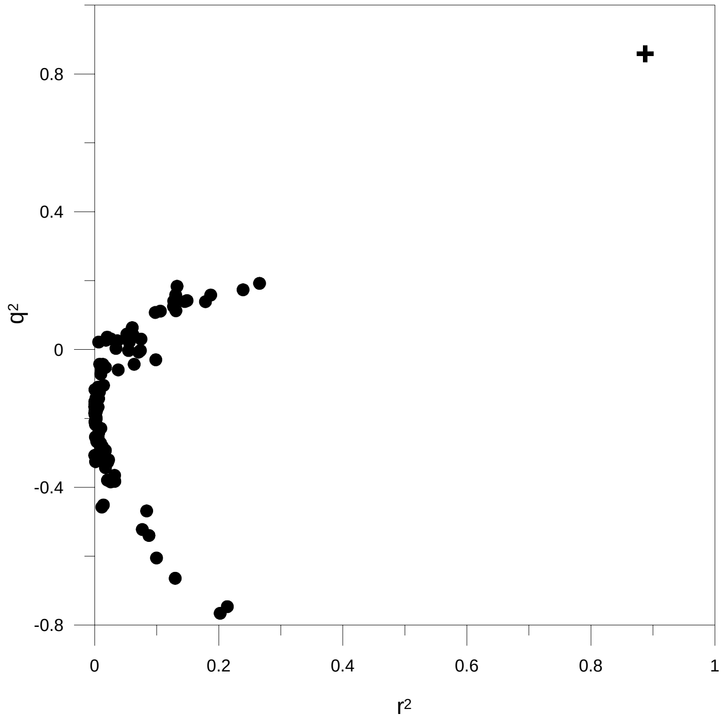


Figure 7. Randomization test for the optimal model.

The cross-validated versus adjusted plot reveals that the normal aliphatic amines are slightly more toxic than the aliphatic alcohols. This probably reflects a pH effect due to the basicity of the amines. Besides, the randomization test shows that results corresponding to altered data do not reach statistically significant levels, with q^2 lower than *0.3* for all cases.

2.2.2 Topological Study of 48 amines

The third application example of TQSI was constituted by 48 anilines. This set was studied in the same conditions as the previous one.

Table 16. Relative toxicity for 48 anilines.

Nº	Compound	LogIGC$_{50}$$^{-1}$	Nº	Compound	LogIGC$_{50}$$^{-1}$
1	2-methylanilline	-0.55	25	4-phenylaniline	0.95
2	2-ethylanilline	-0.25	26	2,4-dimethylaniline	-0.30
3	2-propylanilline	0.06	27	2,5-dimethylaniline	-0.35
4	2-isopropylanilline	0.10	28	2,6-dimethylaniline	-0.43
5	2-phenylanilline	0.86	29	3,4-dimethylaniline	-0.29
6	2-fluoroanilline	-0.31	30	3,5-dimethylaniline	-0.37
7	2-chloroanilline	-0.09	31	2,3-dichloroaniline	1.02
8	2-bromoanilline	0.46	32	2,4-dichloroaniline	0.56
9	2-iodoanilline	0.35	33	2,5-dichloroaniline	0.58
10	3-methylanilline	-0.43	34	2,6-dichloroaniline	0.33
11	3-ethylanilline	-0.12	35	3,4-dichloroaniline	1.14
12	3-phenylanilline	0.78	36	3,5-dichloroaniline	0.71
13	3-fluoroanilline	0.04	37	2-chloro-4-methylaniline	0.24
14	3-chloroanilline	0.09	38	2-chloro-5-methylaniline	0.20
15	3-bromoanilline	0.52	39	2-chloro-6-methylaniline	0.12
16	3-iodoanilline	0.61	40	3-chloro-2-methylaniline	0.45
17	4-methylanilline	-0.02	41	3-chloro-4-methylaniline	0.45
18	4-ethylanilline	0.04	42	4-chloro-2-methylaniline	0.35
19	4-propylanilline	0.49	43	5-chloro-2-methylaniline	0.51
20	4-isopropylanilline	0.21	44	2,3,4-trichloroaniline	1.35
21	4-butylaniline	1.05	45	2,4,5-trichloroaniline	1.30
22	4-pentylaniline	1.67	46	2,4,6-trichloroaniline	1.01
23	4-hexylaniline	2.04	47	3,4,5-trichloroaniline	1.51
24	4-octylaniline	2.34	48	2,6-dichloro-3-methylaniline	0.69

Table 17. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Anilines
Number of Compounds	48
Activity	$(\mathrm{LogIGC}_{50})^{-1}$
Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	Semiempirical AM1 level [4], using Ampac-6.55 [5]
Density Function	Fitted first-order Promolecular ASA (PASA), 3-21G basis set For systems with Iodine, Huzinaga basis set was used
MQSM Operator	Coulomb interatomic QSM
Molecular Alignment	Not needed
Correlation Method	Partial Least Squares (PLS)
Validation	Leave-One-Out Cross Validation (LOO-CV)

Results and discussion

Table 18. Statistical results obtained in the optimal model.

Total N^0 of Indices	**N^0 Descriptors**	r^2	q^2
10	7	0.8229	0.7904

In this case, attending to the larger size of the set, a greater number of descriptors had to be considered in order to appropriately describe the molecular system. Concretely, seven descriptors have to be considered in order to satisfactorily describe the system.

In this set, molecule **24** was removed, because the model depicted a log IGC_{50}^{-1} of *5.70* for the measured *2.34*. This was considered as an outlier, because of the great difference between the measured and the predicted value.

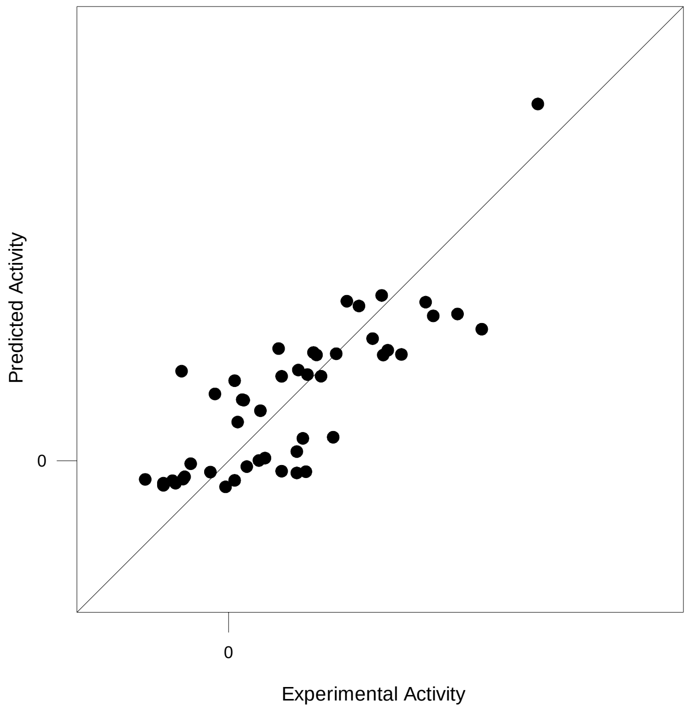


Figure 8. Cross-validated versus experimental inhibition growth concentration values, for the set of 48 anilines.

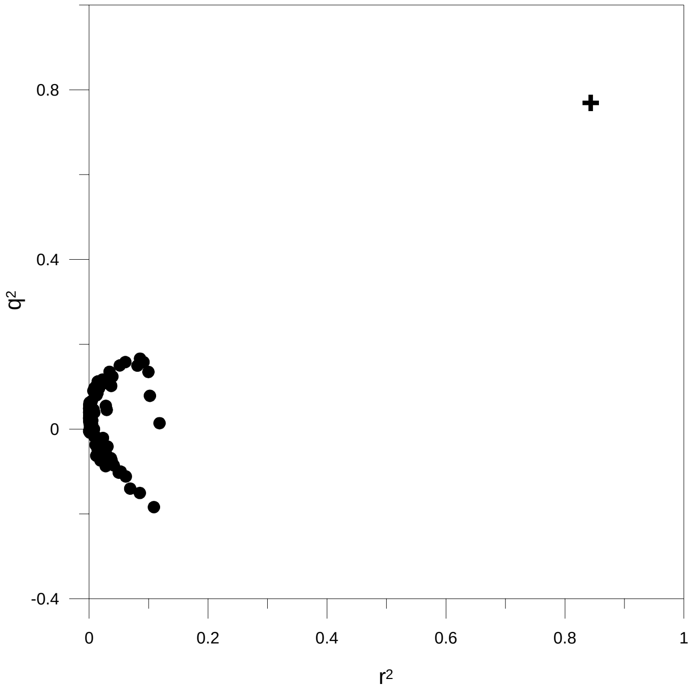


Figure 9. Randomization test for the optimal model.

In the two latter cases, a study of Wayne et al. [64] performed QSAR calculations modelling the data using least squares regression (general linear model procedure) and measuring the model adequacy with the coefficient of determination and the root of the mean square for error.

In comparison with this work, those models resulted in a better quality; however, as no prediction studies were made, the comparison cannot be considered as complete. In the second set, a value of r^2=0.952 was achieved, while in the third one a r^2=0.872 was obtained. It has to be taken on account that, in both cases, the inverse of the logarithm of IGC_{50} was correlated to the logarithm of K_{ow}, the logarithm of 1-octanol/water partition coefficient, as the independent variable.

Contribution

► Gallegos, A.; Gironés, X.; Carbó-Dorca, R. Topological Quantum Similarity Measures: applications in QSAR. In *Proceedings of the 5th GSMS*. Sen, K. (Ed.) Nova Press. *In press.*

3 QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS

3.1 Antimalarial Activity

Introduction

Malaria is an infectious disease endemic in many parts of the world [65], caused by protozoan parasites of the genus Plasmodium. Mostly located in tropical and subtropical areas, it is estimated that hundreds of million people worldwide are affected by malaria and it is considered a major cause of death [66].

Several problems for controlling malaria in these regions, aggravated by inadequate health structures and poor socioeconomic conditions, have propitiated an increasing resistance to the drugs used to combat the Plasmodium parasite.

The spread of drug-resistant Plasmodium Falciparum clones, the most widespread and dangerous, present an increasing immunity against traditional therapies used to inhibit the synthesis of the parasite [67], such as chloroquine [68]. To such an extent, the need for new antimalarial drugs with unconventional structures and novel modes of action to be used for the treatment of pervasive strains of drug-resistant P. Falciparum have impelled the periodic introduction of new antimalarial drugs [69-70].

In present times, two natural peroxides, artemisinin - a naturally occurring peroxidyc cadinane sesquiterpene [71-72]- and yingzhaosu –endoperoxide- [73-75], both of them possessing potent antimalarial activity, have been discovered to be active against chloroquine resistant strains of Plasmodium falciparum. Thus, the total synthesis [76] and structure-activity relationships studies [76-81] of these compounds have opened a new era in chemotherapy of malaria. Due to the complex structure of the natural products, structurally simpler 1,2,4-trioxanes and cyclic peroxy ketals have been synthesized and tested for antimalarial activity [82-84].

The first study is devoted to the establishment of QSAR models for different sets of potential antimalarial drugs, using quantum similarity measures. The first application example tests the correlation of two series of synthetic 1,2,4-trioxanes with different biological responses of the parasite Plasmodium Falciparum (IC_{90} and ED_{90}).

The second application case deals with the activiy (IC_{50}) of set of cyclic peroxy ketals, and the third application case treats two molecular sets composed by artemisinin derivatives, in which the 50% inhibition of synthesis and reduction of hidrofolate (IC50) in different Plasmodium Falciparum clones were studied, applying Kinetic Energy-based Quantum Similarity Measures to QSAR.

In the fourth application case and so forth, molecular topological indices capable to grasp complex information coming from the three-dimensional (3D) molecular structure were described. In order to illustrate the application of the TQSI, the results for the QSAR models related to five molecular families of antimalarial agents are presented.

3.1.1 Classical Study of 20 and 7 Synthetic Trioxanes

In the first case, quantitative QSAR models for two molecular sets of 1,2,4-trioxanes were built using molecular quantum similarity measures (MQSM). The molecular sets were composed by 20 and 7 1,2,4-trioxanes, respectively. Besides, the QSAR results for the antimalarial set composed by 20 1,2,4- Trioxanes were qualitatively analyzed.

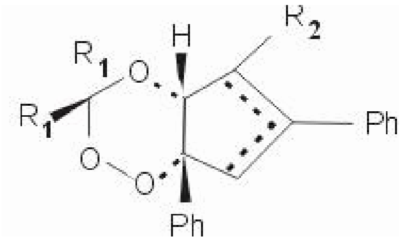


Figure 10. Parent structure of 1,2,4-trioxanes.

For the set composed by 20 antimalarial compounds, the analyzed properties consist of the concentration (in ng/ml) of the drug able to inhibit 90% of synthesis and reduction of hydrofolate (IC_{90}) in the parasite. In vitro experiments have been performed studying two specimens: the P. Falciparum *W2*, from Indochina, and *D6* clone, from Sierra Leone. Due to the wide range of values present in reference [81], a logarithmic scaling has been adopted to uniform the activity data.

For the smaller set formed by 7 compounds, the biological activity is the same (but in mg/kg). However, it has been measured in the P. Berghei in vivo (ED_{90}). In this case, as the range of values is narrower, no scaling was performed.

Table 19. Structures and observed in vitro (log IC90) activities of the set of 20 1,2,4-trioxanes.

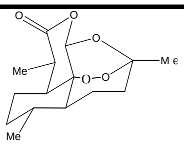
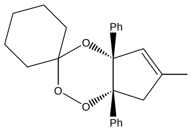
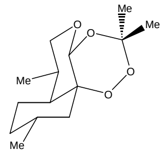
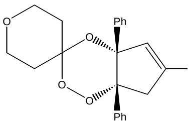
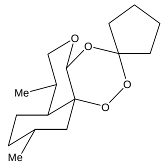
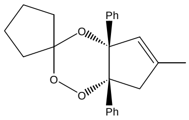
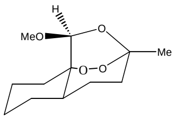
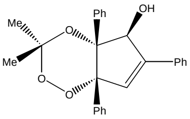
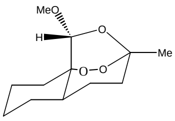
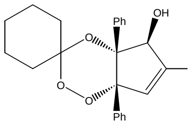
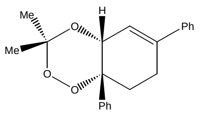
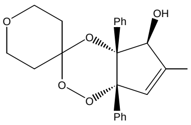
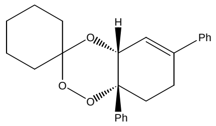
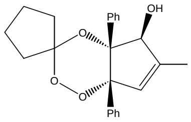
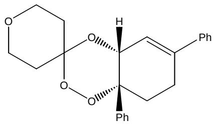
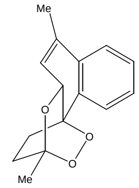
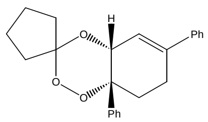
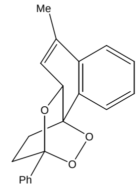
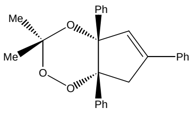
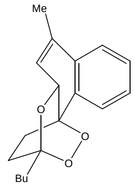
n	Structure	P.falciparum Indo-China W2 log IC_{90}	P.falciparum Sierra Leone D6 log IC_{90}	n	Structure	P.falciparum Indo-China W2 log IC_{90}	P.falciparum Sierra Leone D6 log IC_{90}
1		0.0414	0.3617	11		0.6990	1.0864
2		1.4116	1.7767	12		-0.3979	1.0969
3		0.5185	1.4829	13		0.0414	0.3617
4		0.5798	1.0607	14		0.9731	1.0086
5		1.2279	2.9420	15		0.3010	0.6990
6		3.0734	3.1004	16		1.0864	1.3054
7		2.8082	3.4125	17		-0.2218	1.0253
8		2.7482	3.1483	18		3.0550	2.8209
9		2.3483	2.3856	19		2.7543	2.6884
10		0.9243	1.2856	20		1.4771	1.5441

Table 20. Structures and observed in vivo (ED90) activities of the set of 7 1,2,4-trioxanes.

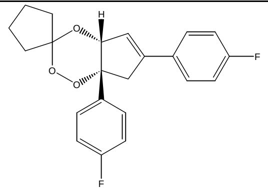
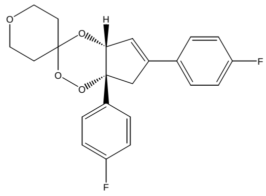
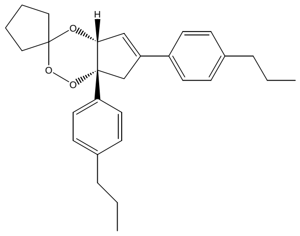
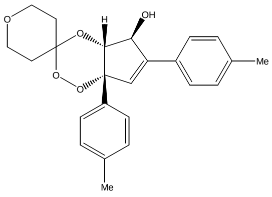
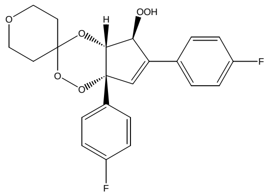
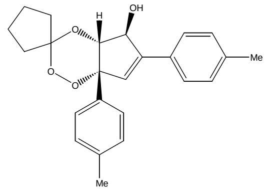
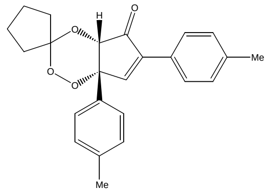
n	Structure	P.beghei ED$_{90}$
1		6.8
2		19.5
3		22.5
4		17.0
5		13.2
6		15.0
7		16.0

Table 21. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	1,2,4-trioxanes
Number of Compounds	20
Activity	Log IC_{90} Plasmodium Falciparum for Indo-China (W2) strain Log IC_{90} Plasmodium Falciparum for Sierra Leone (D6) strain
Compounds	1,2,4-trioxanes
Number of Compounds	7
Activity	ED_{90} Plasmodium Berguei
Computational Details	
Molecular Modelling	PC Spartan software package [7]
Geometry Optimization	built-in Sybyl Molecular Mechanics force-field
Density Function	Fitted first-order Promolecular ASA (PASA), 3-21G basis set
MQSM Operator	Overlap operator
Index transformation	Carbó index
Molecular Alignment	Topo-Geometrical Superposition Algorithm (TGSA)
Reduction of Dimensions	Principal Components Analysis (PCA)
Selection of Variables	Most Predictive Variables Method (MPVM)
Correlation Method	Multiple Linear Regression (MLR)
Validation	Internal Leave-One-Out Cross-Validation (LOO-CV) Random Test

Results and discussion

In this case, the optimization was carried out using the built-in Sybyl Molecular Mechanics (MM) force-field. This choice was made according to a previous study [81], where a comparison between molecular mechanics and *ab initio* procedures was carried out, proving the superiority of MM in these molecular sets, and thus saving computational time.

Taking into account that the studied molecular sets share common structural features, the Topo-Geometrical Superposition Algorithm (TGSA) is used as it performs pairwise superpositions according to the molecular backbones. This molecular superposition method overlays the molecules according the maximal common substructure shared by the analyzed molecules.

The normalized scaling of the MQSM has been done by means of the Carbó index transformation.

Table 22. Statistical parameters for P. falciparum Indo-china W2 clone. Optimal model marked in italics nd bold face.

#PCs	r^2	q^2	σ_N	PCs used
1	0.524	0.446	0.771	2
2	0.631	0.489	0.679	1, 2
3	0.722	0.578	0.589	1, 2, 7
4	***0.757***	***0.589***	***0.550***	***1, 2, 3, 7***
5	0.797	0.592	0.504	1, 2, 3, 5, 7
6	0.828	0.646	0.462	1, 2, 3, 5, 7, 11
Equation	$\log IC_{90}^{W2} = 2.912\mathbf{x_1} - 7.166\mathbf{x_2} - 2.083\mathbf{x_3} + 4.948\mathbf{x_7} + 1.272$			

Table 23. Statistical parameters for P. falciparum Sierra Leone D6 clone. Optimal model marked in italics nd bold face.

#PCs	r^2	q^2	σ_N	PCs used
1	0.519	0.423	0.663	2
2	0.575	0.427	0.623	1, 2
3	0.757	0.629	0.472	1, 2, 7
4	***0.789***	***0.662***	***0.439***	***1, 2, 7, 11***
5	0.795	0.621	0.433	1, 2, 4, 7, 11
6	0.801	0.540	0.426	1, 2, 4, 5, 7, 11
Equation	$log\ IC_{90}^{D6} = 1.806\mathbf{x_1} - 6.102\mathbf{x_2} + 5.966\mathbf{x_7} - 3.420\mathbf{x_{11}} + 1.730$			

Table 24. Statistical parameters for P. Berghei. Optimal model marked in italics nd bold face.

#PCs	r^2	q^2	σ_N	PCs used
1	0.643	0.270	2.76	6
2	0.852	0.428	1.78	4, 6
3	***0.929***	***0.708***	***1.23***	***4, 5, 6***
4	0.976	0.789	0.721	3, 4, 5, 6
Equation	$ED_{90} = -16.53\mathbf{x_4} - 11.67\mathbf{x_5} - 37.52\mathbf{x_6} + 15.79$			

As evidenced in the quoted set of equations, the most descriptive PCs are not necessarily those accounting for maximal variance.

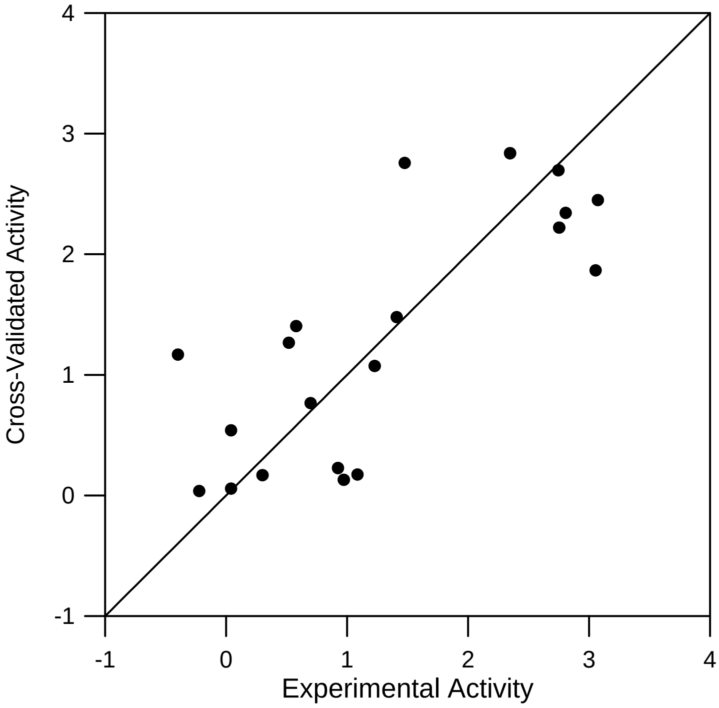


Figure 11. Cross-validated versus experimental antimalarial activity values over P. falciparum W2.

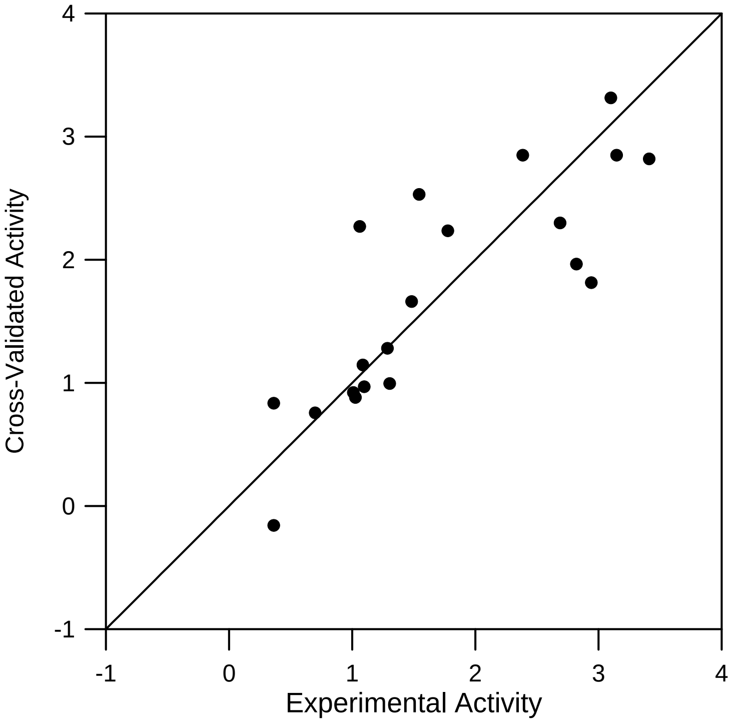


Figure 12. Cross-validated versus experimental antimalarial activity values over P. falciparum D6.

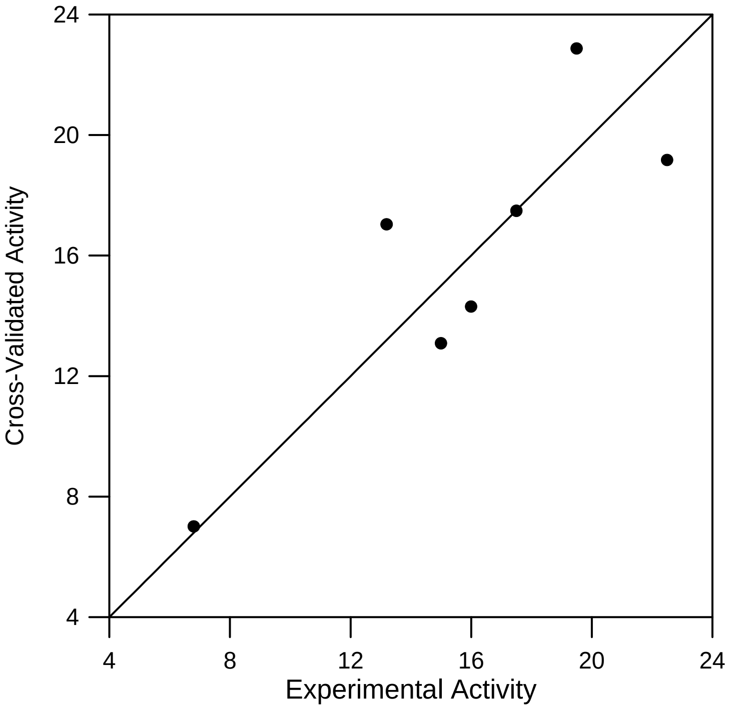


Figure 13. Cross-validated versus experimental antimalarial activity values over P. Berguei.

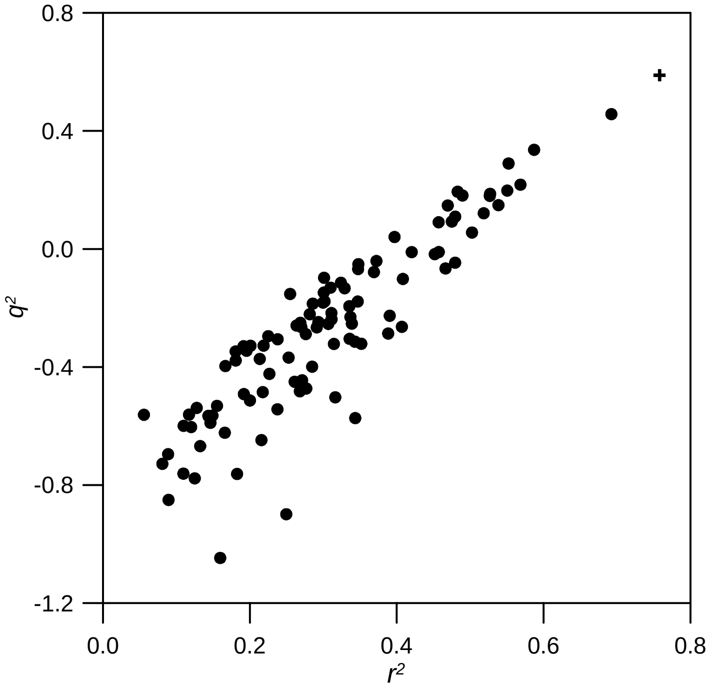


Figure 14. Randomization test for the optimal P. falciparum W2 model. The randomized responses have been marked with dots, and the correctly ordered activity with a cross.

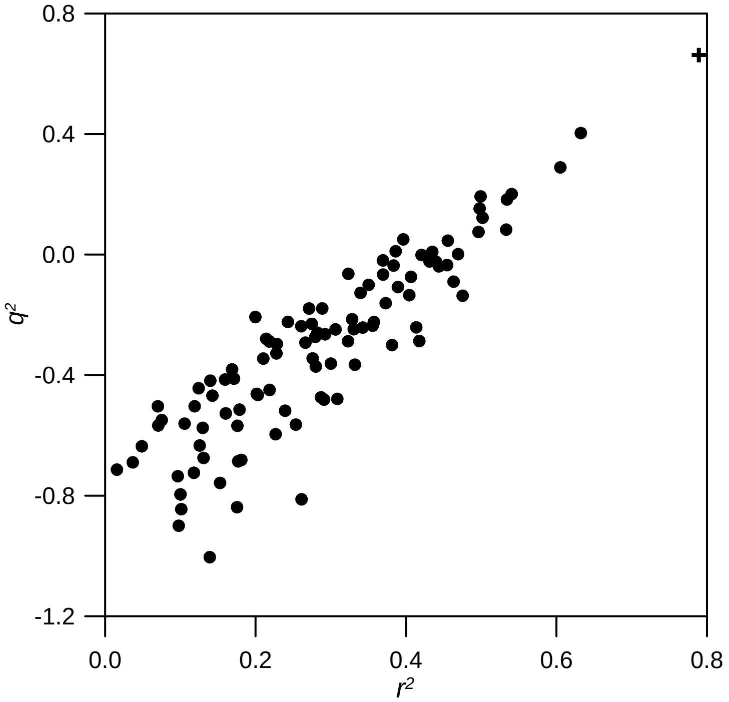


Figure 15. Randomization test for the optimal P. falciparum D6 model.

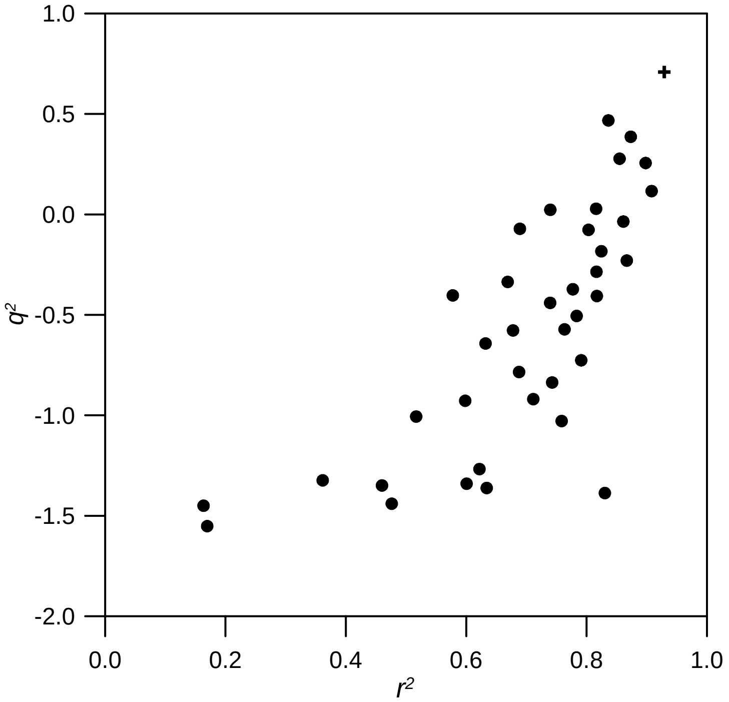


Figure 16. Randomization test for the optimal P. Berguei model. In this case, 61 dots, which were q^2 *< 1.5*, were eliminated.

As observed, a clear separation is present between the actual models and the permuted ones, none of them yielding to $q^2 > 0.5$, or furthermore being negative. As a matter of fact, the last random test lacks of 61 points that were removed because $-300 < q^2 < -1.5$, and no clear representation could be obtained. In this way, it can be concluded that real QSAR have been discovered, and that no fortuitous correlations or overparameterizations exist in the reported models.

A further analysis of the results obtained in the equations may help in the interpretation of QSAR models. The first set is ideal to carry out this analysis provided that it is composed by assorted different substitutions. This molecular set includes two different biological activities, which correlate fairly well (*r^2=0.796*), when applying the logarithmic transformation. This fact explains both the resemblance in the statistical results, and the elections and relevance of the optimal PCs, which were chosen in the same order, except for the last one.

The subsequent step involves an analysis of the PCs chosen to see how the molecular point clouds are distributed in the lower dimensional space. As example, the first and second PCs are plotted in *Figure 17*.

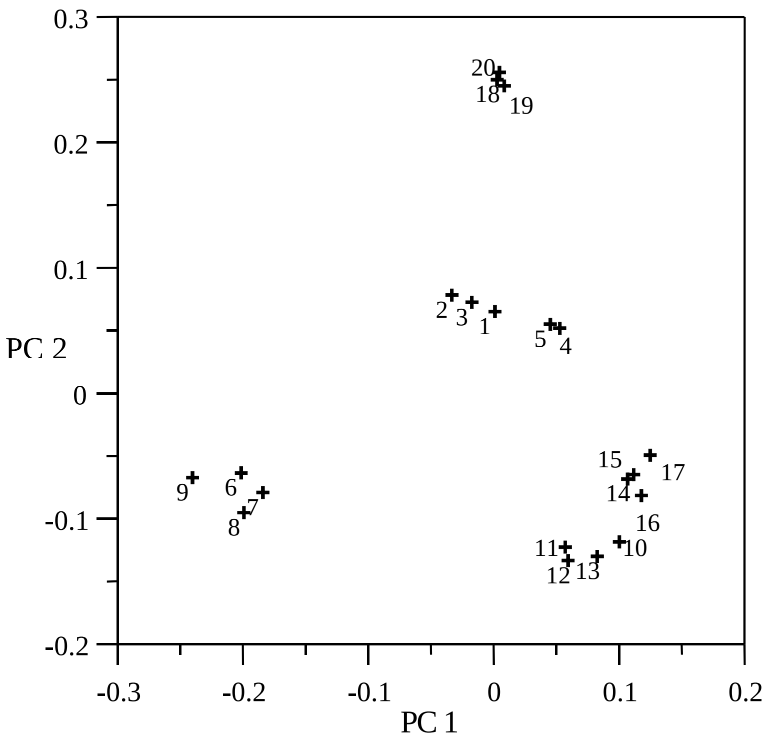


Figure 17. Plot of the first versus de second PC for the molecular set of 20 1,2,4-trioxanes.

As it can be observed, a clear 4-cluster pattern is present, grouping molecules according to clear substructural features:

- Molecules **1**–**5**, which present small aliphatic, but no aromatic, substitutions, present high activity.
- Molecules **6**–**9**, which have an aromatic substitution and a phenyl group in the region where the two main ring fuse, present low activity.
- Molecular **10**–**17**, which present phenyl groups at both sides of the fusion region, present high or very high activity.
- Molecular **18**–**20**, which have a benzene ring fused to the main backbone, present low activity.

Thus, it can be seen the influence of these phenyl groups in the biological activity and how it is reflected in this simple two-dimensional space. From this observation, it can be deduced that any novel structure falling in the first, or better in the third, cluster would possess a high activity against both breeds of P. Falciparum. In this way, computational design should be guided to those structures substituted with two phenyl groups at both sides of the ring fusion area and tested with different functional groups.

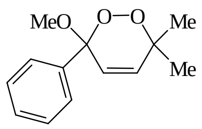
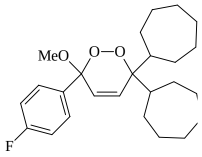
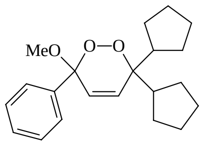
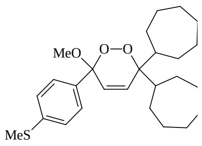
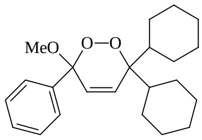
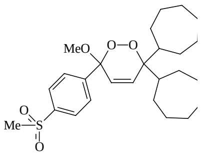
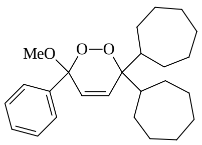
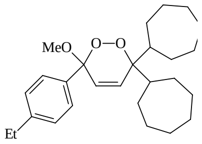
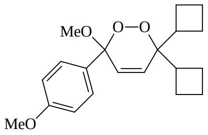
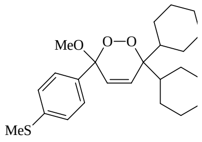
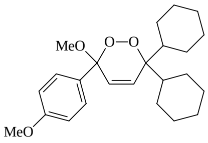
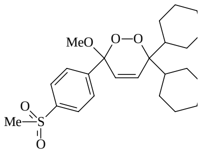
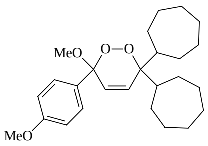
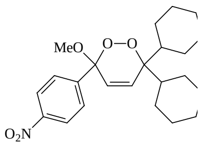
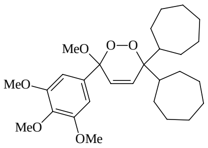
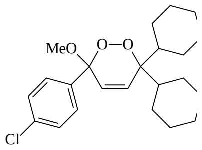
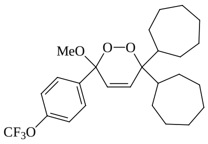
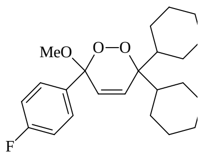
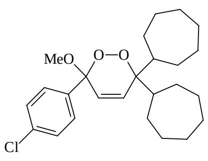
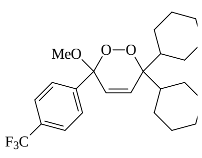
A a conclusion, in the present study, molecular quantum similarity measures were applied to correlate systematically the antimalarial activity of 1,2,4-trioxanes. Satisfactory quantitative models were obtained using a small number of descriptors based on Principal Components Analysis, achieving also good results in Leave-One-Out Cross Validations (LOO-CV) and random tests. In addition, a qualitative analysis of the results for the antimalarial set composed by 20 1,2,4- Trioxanes was carried out, revealing structural information about the data set. The molecules were clustered according to common structural features, which in turn explained the biological activity. When two phenyl substitutions are present in the molecule, it seems that the biological activity tends to increase.

3.1.2 Classical Study of 20 Cyclic Peroxy Ketals

In the second antimalarial study case, a quantitative QSAR models for a series of 20 cyclic peroxy ketals was built using molecular quantum similarity measures (MQSM).

The biological property studied in the 20 cyclic peroxy ketals set [85] also consists of inhibition of the metabolism to hydrofolate, but 50% (IC_{50}) in nM concentration units. Similarly to the first antimalarial set, activity values were taken in logarithmic scale.

Table 25. Structures and observed (log IC_{50}) activities of 20 cyclic peroxy ketals.

n	Structure	log IC_{50}	n	Structure	log IC_{50}
1		3.041	11		1.929
2		2.279	12		1.892
3		2.447	13		1.491
4		2.342	14		2.255
5		2.204	15		2.204
6		2.255	16		1.748
7		2.322	17		1.663
8		2.079	18		2.000
9		1.785	19		2.301
10		1.763	20		2.146

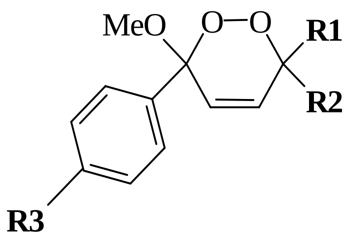


Figure 18. Parent structure of cyclic peroxy ketals.

Table 26. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Cyclic peroxy ketals
Number of Compounds	20
Activity	Log IC_{50}
Computational Details	
Molecular Modelling	PC Spartan software package [7]
Geometry Optimization	built-in Sybyl Molecular Mechanics force-field
Density Function	Fitted first-order ASA DF, 3-21G basis set
MQSM Operator	Coulomb operator
Index transformation	Carbó index
Molecular Alignment	Topo-Geometrical Superposition Algorithm (TGSA)
Reduction of Dimensions	Principal Components Analysis (PCA)
Selection of Variables	Most Predictive Variables Method (MPVM)
Correlation Method	Multiple Linear Regression (MLR)
Validation	Internal Leave-One-Out Cross-Validation (LOO-CV) Random Test

Results and discussion

Table 27. Statistical parameters for P. falciparum. The optimal QSAR model has been marke in italics and bold face.

#PCs	r^2	q^2	σ_N	PCs used
1	0.546	0.414	0.225	2
2	0.592	0.418	0.214	1,2
3	0.738	0.607	0.171	1,2,6
4	***0.778***	***0.691***	***0.158***	***1,2,3,6***
5	0.795	0.656	0.151	1,2,3,4,6
Equation	$\log IC_{50} = 1.739\mathbf{x}_1 - 10.188\mathbf{x}_2 - 4.042\mathbf{x}_3 - 11.379\mathbf{x}_6 + 2.107$			

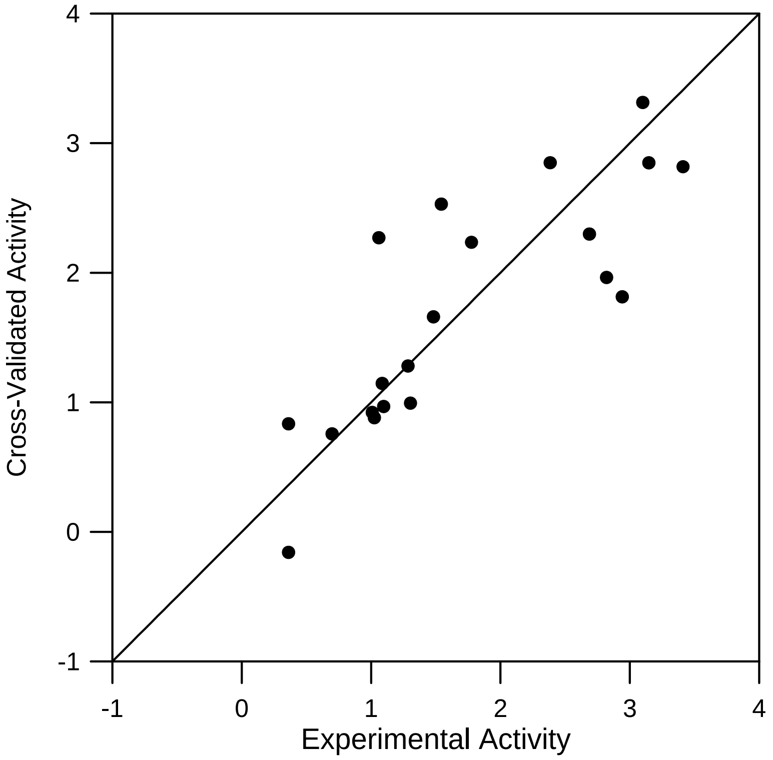


Figure 19. Cross-validated versus experimental antimalarial activity values.

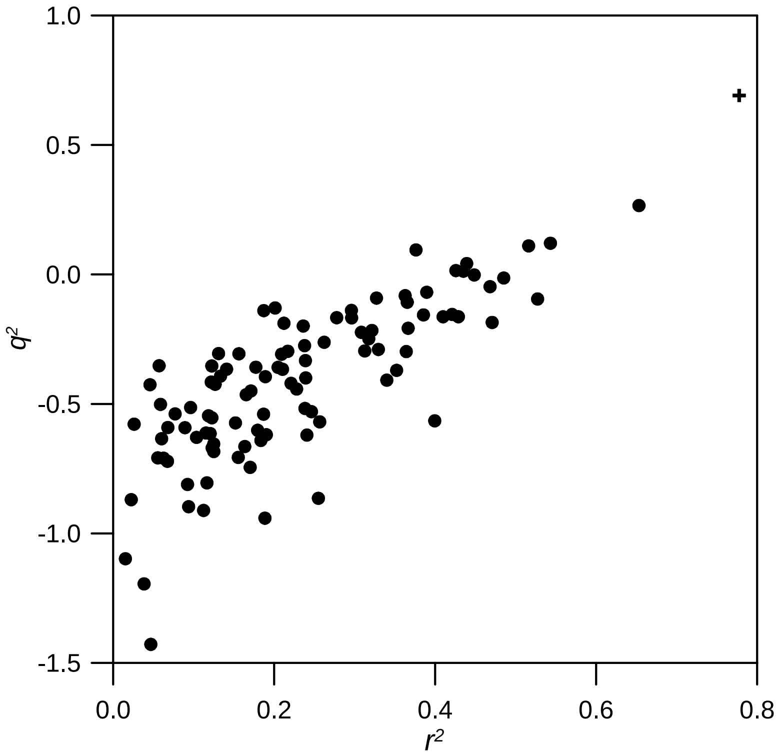


Figure 20. Randomization test for the optimal P. falciparum. The randomized responses have been marked with dots, and the correctly ordered activity with a cross.

The results are comparable to the previous study case.

Contribution

► Gironés, X.; Gallegos, A.; Carbó-Dorca, R. Antimalarial Activity of Synthetic 1,2,4-Trioxanes and Cyclic Peroxy Ketals, a Quantum Similarity Study. *J. Comput.-Aid. Mol. Des., 15(12),* **2001**, 1053-1063.

3.1.3 Study of 18 and 15 artemisinin derivatives

In this study, quantitative QSAR models were built for two molecular sets, composed by 18 and 15 artemisinin derivatives, respectively.

For the set composed by 18 antimalarial compounds, the analyzed property werethe nanomolar concentration of the drug able to inhibit 50% synthesis and reduction of hydrofolate in the NF54 strain (choloroquine sensitive) of P. falciparum in vitro (IC_{50}) [86], as reported in a previous work [87].

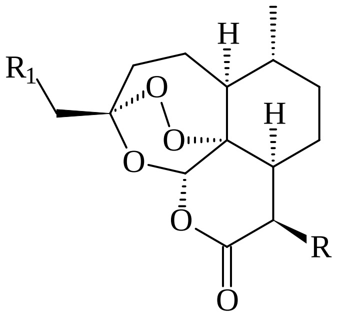
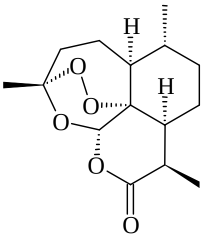
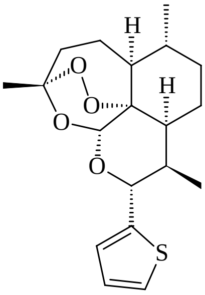
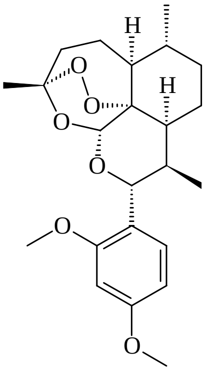
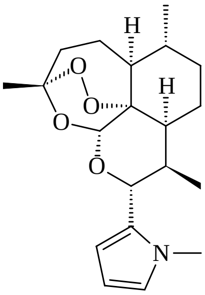
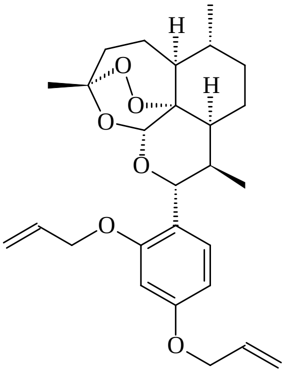
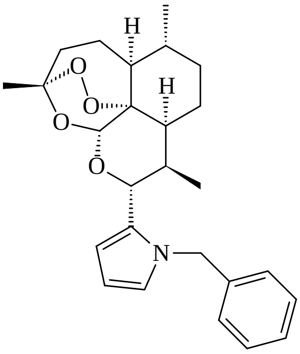
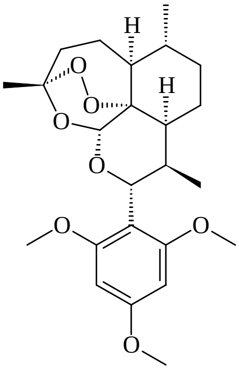
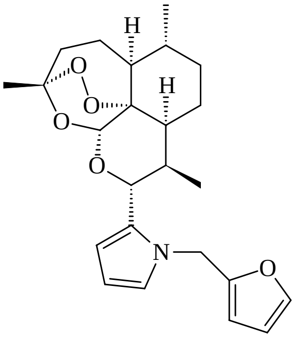


Figure 21. Parent structure of artemisinin derivatives.

For the second series, the studied property is also IC_{50}; however, the activity were measured relative to the artemisinin value when tested in vitro in human blood over P. falciparum Indochina (W2) and Sierra Leone (D6) clones [88], following a procedure proposed by Desjardins [89-90]. These clones present an interesting resistance to drugs: whereas the W2 strain is chloroquine-resistant and mefloquine-sensitive, the D6 breed is sensitive to chloroquine and resists mefloquine. This fact allows evaluating the effect and mechanism of action of a novel drug in this widely mutated parasite. The relative potency of these compounds has been adjusted according to IC_{50} values and then multiplied by the ratio between molecular weight of the analog and molecular weight of artemisinin. A logarithmic scaling has been applied to the last set of activities due to the wide range of values present in reference [88].

Table 28. Molecular structures and biological activities of a set of 18 artemisinin derivatives.

nº	Structure	IC_{50}	nº	Structure	IC_{50}
1		9.908	10		5.105
2		4.198	11		4.609
3		6.607	12		15.996
4		7.798	13		9.397

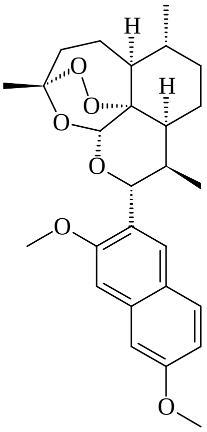
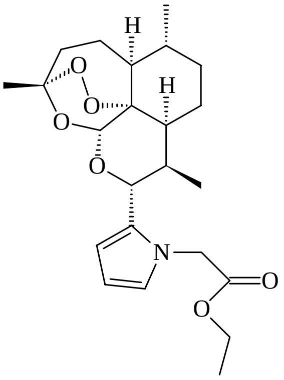
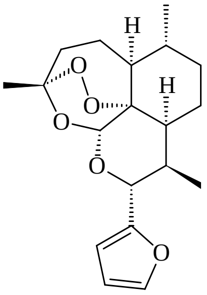
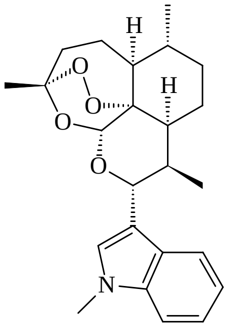
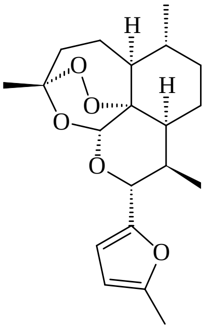
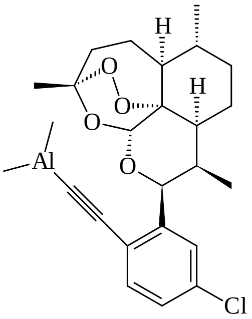
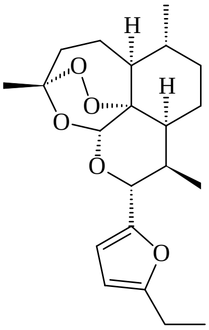
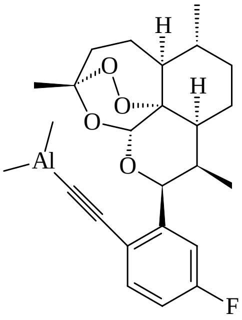
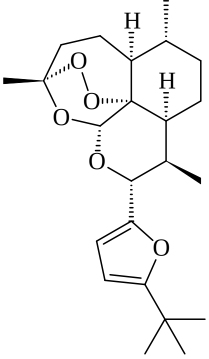
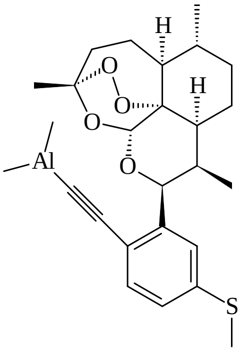
5		8.995	14		9.099
6		1.400	15		4.000
7		5.200	16		10.990
8		8.590	17		8.299
9		10.000	18		8.395

Table 29. Biological activities of a set of 15 artemisinin derivatives.

N	R_1	R	Log IC_{50} (D6)	Log IC_{50} (W2)
1	H	CH_3	2.000	2.000
2	CH_3	H	1.944	2.049
3	CH_3CH_2	H	3.323	2.828
4	$CH_3(CH_2)_2$	H	1.301	1.255
5	$(CH_3)_2CH$	H	1.724	1.653
6	EtO_2CCH_2	H	2.365	2.365
7	$C_6H_5CH_2$	H	0.477	0.000
8	p-$ClC_6H_4(CH_2)_2$	H	2.057	2.104
9	$C_6H_4(CH_2)_3$	H	2.342	2.449
10	CH_3	$CH_3(CH_2)_3$	2.265	2.410
11	$CH_3(CH_2)_2$	$CH_3(CH_2)_3$	1.447	1.519
12	$C_6H_5CH_2$	$CH_3(CH_2)_3$	0.000	0.000
13	p-$ClC_6H_4(CH_2)_2$	$CH_3(CH_2)_3$	1.633	1.724
14	$C_6H_4(CH_2)_3$	$CH_3(CH_2)_3$	1.591	1.681
15	EtO_2CCH_2	$CH_3(CH_2)_3$	3.141	3.359

Table 30. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Artemisinin derivatives
Number of Compounds	18
Activity	IC_{50} Reduction of hydrofolate in the NF54 strain
Compounds	Artemisinin derivatives
Number of Compounds	15
Activity	Log IC_{50} Plasmodium Falciparum for Indo-China (W2) strain Log IC_{50} Plasmodium Falciparum for Sierra Leone (D6) strain

Computational Details	
Molecular Modelling	PC Spartan software package [7]
Geometry Optimization	Built-in Sybyl Molecular Mechanics force-field
Density Function	Fitted Promolecular ASA (PASA) KE DF, 3-21G basis set
MQSM Operator	Kinetic Energy operator
Index transformation	Carbó index
Molecular Alignment	Topo-Geometrical Superposition Algorithm (TGSA)
Reduction of Dimensions	Principal Components Analysis (PCA)
Selection of Variables	Most Predictive Variables Method (MPVM)
Correlation Method	Multiple Linear Regression (MLR)
Validation	Internal Leave-One-Out Cross-Validation (LOO-CV) Random Test

Results and discussion

The choice of the best geometry optimization method was made after performing a comparative analysis between different optimization methodologies over the artemisinin molecule. The comparison was carried out using the following methodologies: Sybyl Molecular Mechanics (MM) force field and AM1, both included in PC Spartan, and using a direct *ab initio* method, which in this case corresponds to a restricted Hartree-Fock with the 3-21G* basis set, implemented in Gaussian 98 [91]. From the optimized molecular coordinates of artemisinin, the electronic DF was constructed within the previously discussed promolecular ASA, and MQSM involving the Coulomb operator have been carried out. The quantum similarity results over the artemisinin molecule were given in terms of Carbó Indices, resulting from the comparison of the different geometries of artemisinin.

Table 31. Upper triangle of the Carbó Index matrix for artemisinin used to compare the different computational optimization methodologies.

Sybyl	**AM1**	**HF/3-21g***
1.000	0.978	0.985
	1.000	0.980
		1.000

From the results, it is evidenced that in this case all methods lead to very close structures. However, there is a very important difference in computational time required to complete the calculations: whereas the Sybyl process was completed in a few seconds, AM1 and ab initio methods lasted some minutes and several hours respectively. In this way, the Sybyl methodology was chosen to optimize the geometry due to its accuracy and efficiency regarding these molecular sets.

Promolecular ASA Kinetic Energy (KE) DF with atomic densities fitted to a 3-21G basis set has been used. It has been shown that, at low and high DF values, the electronic and the KE DF behave almost in the same way, reflecting the molecular shape and the atomic locations respectively. However, at intermediate values some relevant differences become visible, like an oversize of the heavier atoms and the appearance of interatomic maxima. These features preclude a different description of the electronic distribution based on KE concepts.

Table 32. Statistical parameters for the proposed QSAR models.

P. falciparum strain	# PCs	PCs used	% Variance explained	r^2	q^2	σ_N
NF54	4	2,4,9,16	27.775	0.754	0.561	2.140
D6	4	4,6,14,15	19.503	0.767	0.520	0.580
W2	4	1,4,14,15	37.693	0.821	0.576	0.575

Table 33. QSAR equations for the different models.

$$\mathrm{IC}_{50}^{\mathrm{NF54}} = 11.956\mathbf{x}_2 - 12.179\mathbf{x}_4 - 16.280\mathbf{x}_9 - 74.446\mathbf{x}_{16} + 7.700$$

$$\log \mathrm{IC}_{50}^{\mathrm{D6}} = 1.900\mathbf{x}_4 + 4.620\mathbf{x}_6 - 32.563\mathbf{x}_{14} + 5.965\mathbf{x}_{15} + 1.826$$

$$\log \mathrm{IC}_{50}^{\mathrm{W2}} = 4.276\mathbf{x}_1 + 2.507\mathbf{x}_4 - 30.241\mathbf{x}_{14} + 5.822\mathbf{x}_{15} + 1.841$$

As it can be seen seen, not always the PCs accounting for the maximal variance are those related to the activity. In the studied molecular sets, even the chosen PCs explain less than 40% of the whole variance, they are able to provide an acceptable description.

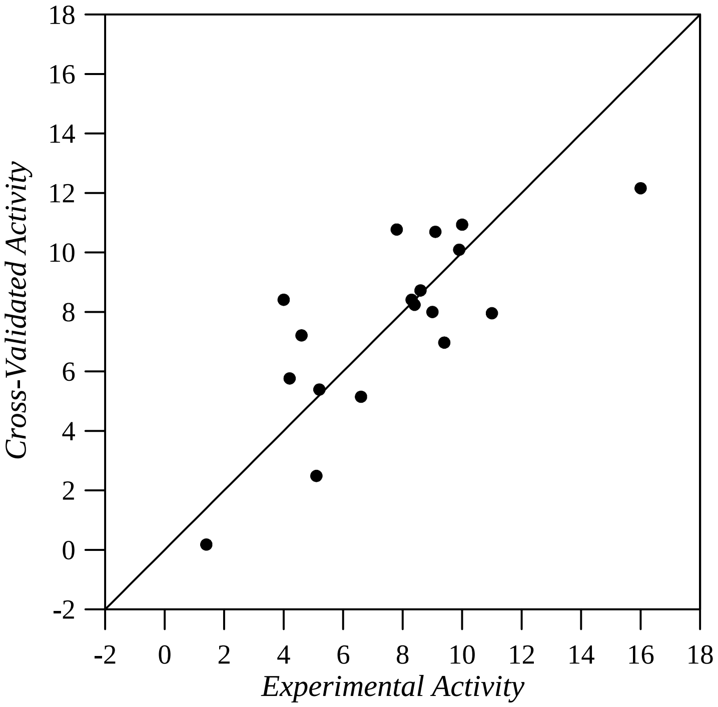


Figure 22. Cross-validated versus experimental antimalarial activity values over P. falciparum NF54.

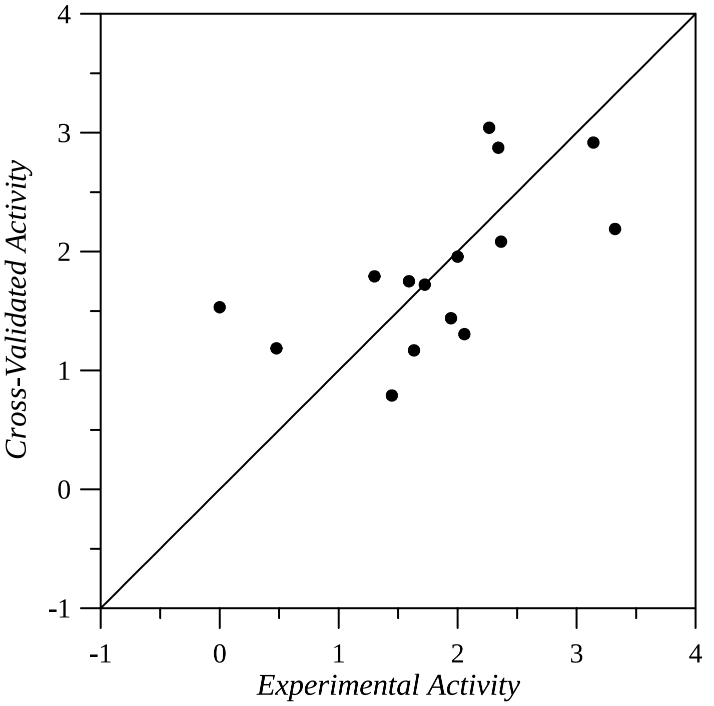


Figure 23. Cross-validated versus experimental antimalarial activity values over P. falciparum D6.

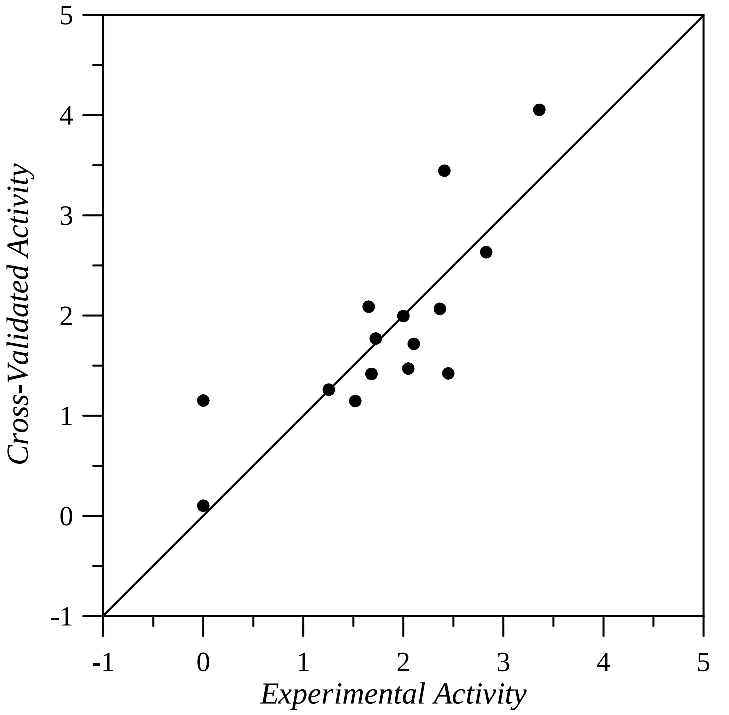


Figure 24. Cross-validated versus experimental antimalarial activity values over P. falciparum W2.

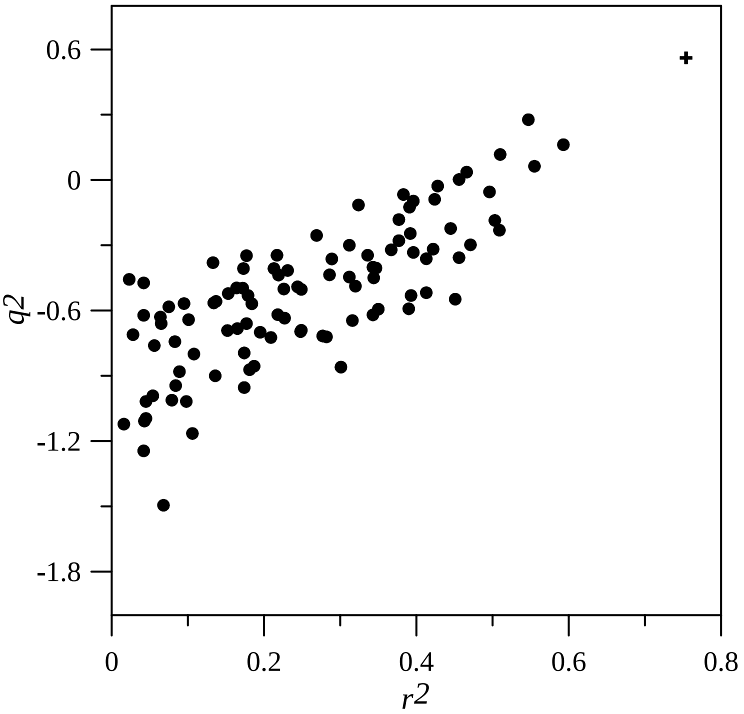


Figure 25. Randomization test for the optimal P. falciparum NF54 model. The randomized responses have been marked with dots, and the correctly ordered activity with a cross.

The proposed models were subjected to a random test procedure to assess that they were not built up from an excess of parameters. Thus, in each molecular set, a hundred activity vectors were generated from randomized permutations of the original ordered one. All proposed models presented a clear separation between the original solution and the randomized ones, which clearly achieve values of q^2 below 0.5, or furthermore being negative. As an example, the compilation of the results for the P. falciparum NF54 set is presented. Similar results are obtained for the remaining studied systems.

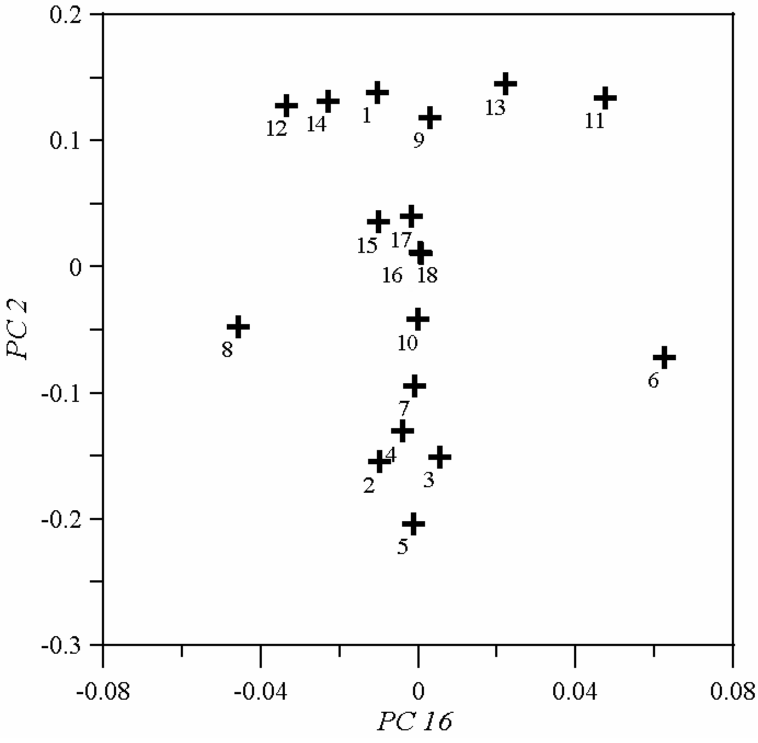


Figure 26. 2D distribution of the antimalarial compounds according to the first two chosen PCs.

The first two most predictive PCs spread the molecules of the first molecular set in the 2D space are presented. As it can be seen, both PC 16 and 2 behave differently. PC 16 scatters the molecules according to the weight of the substitution, roughly collapsing the heavy substitutions nearby the origin and distributing the light ones along the sides. PC 2 mostly acts as a substituent discriminator. In the negative part of PC 2, most of the compounds presenting oxygen are present (**2, 3, 4, 5, 6, 7, 8**), those compounds having an aluminum atom are approximately located in the middle of the axis (**16, 17, 18**) and finally the nitrogenated substitutions are present in the positive part of PC 2 (**11, 12, 13, 14**).

Exceptions to this rule are compound **1**, which only contains a ketonic substitution, compound **9**, which contains oxygen but is located with the nitrogen group, and compound **15**, which contains a nitrogen and it is located too close to the aluminum containing substitutions. Compound **10** contains a sulfur and is located between the oxygen containing rings and the heavier substitutions with aluminum.

Kinetic Energy based molecular quantum similarity measures were applied to correlate sistematically the antimalarial activity of various artemisinin derivatives, yielding satisfactory correlations for all antimalarial activities in all studied molecular sets.

Contribution

► Gironés, X.; Gallegos, A.; Carbó-Dorca, R. Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR. *J. Chem. Inf. Comput. Sci, 40,* **2000**, 1400-1407.

3.1.4 Topological study of five series of antimalarial compounds

A series of five antimalarial sets of compounds, were studied following the same statistical protocol.

Table 34. Summary of the protocol.

Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	Semiempirical AM1 level [4], using Ampac-6.55 [5]
Density Function	Atomic densities described by *1S* GTO basis functions
MQSM Operator	Coulomb or Cioslowski interatomic QSM
Molecular Alignment	Not needed
Reduction of Dimensions	Nested Summation Symbols (NSS) Algorithm
Correlation Method	Multiple Linear Regression (MLR)
Validation	Leave-One-Out Cross Validation (LOO-CV)

Once given the molecular geometry, for practical purposes, a simple set of *1S* GTO basis functions was employed in order to describe atomic densities, instead of ASA DF:

$$g_i = g(\boldsymbol{r} - \boldsymbol{R}_i, \zeta_i) = N_i e^{-\zeta_i (\boldsymbol{r} - \boldsymbol{R}_i)^2}$$

The term N_i is a normalisation factor,

$$N_i = \left(\frac{2\zeta_i}{\pi} \right)^{\frac{3}{4}} \left(v_i^T \right)^{\frac{1}{2}}$$

being v_i^T the topological atomic valence of the atom *i*. The function exponent ζ_i parameter can be numerically modulated for every atom, acoording to the following table:

Table 35. Atomic exponents (in a.u.)

Atom	Exponent
H	3.436350
C	0.467380
N	0.497510
O	0.530530
F	0.548240
S	0.249715
Cl	0.268040
Br	0.196090

The classical topological matrix was computed for all the systems. In addition, similarity operators were used to compute the Atomic Quantum Similarity Measures (AQSM); in particular, the Coulomb, and the Cioslowski operators. The matrices resulting from the similarity measures allow to reproduce, among others, the classical TI formulation for indices such as: Wiener (W) and Wiener Path Number (WPN), Randic (χ), Schultz (MTI), Balaban (B) and Hosoya (Z) indices, Harary Number (H), the generalised connectivity indices (${}^{m}\chi_{t}$) of Kier and Hall, and so on. Of course, the topological distance matrices were substituted in the TQSM case by the three dimensional euclidean distances, in such a way that the resulting indices included information about the molecular 3D structure.

The optimized structures were sent to the program package developed in our laboratory where the TM and TQSI were computed [92]. During the phase of molecular indices generation, some additional restrictions were considered in order to reduce the amount of data to be analysed: the generalised connectivity indices of Kier and Hall were computed only up to order 9 and only contributions up to also order 9 were considered to obtain the Hosoya index.

Once the TQSI matrices were obtained, each array was sent to a multiple linear regression program. All the combinations of 2, 3 and 4 descriptors were generated and the ones attached to the highest values of the q(2) coefficient have been reported. The linear correlation coefficients arising from a true leave-one-out cross-validation procedure, r^2_{cv}, and the data fitting, r^2, are reported. Except for particular cases which are specifically indicated, the statistical significance parameter coming from the Snedecor *F*-test, *p*, is lesser than 0.0001 for all the correlation coefficients attached to the cross-validation processes.

3.1.4.1 Topological study of 15 artemisinin analogs

This molecular family is composed by a set of 15 3-alkyl substituted analogs of artemisinin [88]. In vitro activities against W2 and D6 strains of Plasmodium falciparum are reported in the original article in terms of relative IC_{50} value. The relative activity was computed as the relative quantity

$$100\frac{(IC_{50})_{artemisinin}(W)_{ana\log}}{(IC_{50})_{ana\log}(W)_{artemisinin}}$$

where W stands for the molecular weight [88].

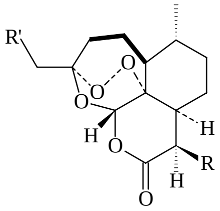


Figure 27. General molecular structure of the 3-alkyl substituted analogs of artemisinin molecules of system 1.

Table 36. Activities and structures of the molecules of system 1.

Molecule	R'	R	Relative activity	
			D6	W2
1	H	CH_3	100	100
2	CH_3	H	88	112
3	CH_3CH_2	H	2102	673
4	$CH_3(CH_2)_2$	H	20	18
5	$(CH_3)_2CH$	H	53	45
6	EtO_2CCH_2	H	232	232
7	$C_6H_5CH_2$	H	3	1
8	*p*-$ClC_6H_4(CH_2)_2$	H	114	127
9	$C_6H_5(CH_2)_3$	H	220	281
10	CH_3	$CH_3(CH_2)_3$	184	257
11	$CH_3(CH_2)_2$	$CH_3(CH_2)_3$	28	33
12	$C_6H_5CH_2$	$CH_3(CH_2)_3$	1	1
13	*p*-$ClC_6H_4(CH_2)_2$	$CH_3(CH_2)_3$	43	53
14	$C_6H_5(CH_2)_3$	$CH_3(CH_2)_3$	39	48
15	EtO_2CCH_2	$CH_3(CH_2)_3$	1382	2285

Table 37. Summary of the molecular data set.

Molecular Data Set	
Compounds	3-alkyl substituted analogs of artemisinin
Number of Compounds	15
Activity	Relative Activity (IC_{50} Plasmodium Falciparum, D6 strain) Relative Activity (IC_{50} Plasmodium Falciparum, W2 strain)

Results and discussion

Table 38. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the D6 activity.

Nº descr.	r^2	r^2_{cv}	Linear model equation (log D-6 activity)
2	0.590	0.422	$y = 0.0107418\ WPN^T - 0.00570827\ p^T(3) + 1.09179$
3	0.839	0.720	$y = -14.2616\,{}^4\chi_P^S + 6.81012\,{}^4\chi_P^C + 20.4000\,{}^8\chi_{CH}^C - 13.0016$
4	0.933	0.851	$y = 13.9258\,{}^4\chi_{PC}^T - 5.88954\,{}^5\chi_{PC}^T - 19.1486\,{}^6\chi_P^S + 8.31544\,{}^8\chi_P^C - 0.961919$

Table 39. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the W2 activity.

Nº descr.	r2	r2cv	Linear model equation (log D-6 activity)
2	0.631	0.476	$y = 0.0121877\ W^T - 6.46274\cdot10^{-3}\ p^T(3) + 0.262345$
3	0.854	0.744	$y = -3.98021\,{}^8\chi_P^T + 1.88598\ {}^4\chi_P^C - 13.6583\,{}^6\chi_{CH}^C + 2.47056$
4	0.961	0.897	$y = 2.54094\chi^T - 6.07153\,{}^5\chi_P^S - 52.5154\,{}^8\chi_P^S + 11.4215\,{}^8\chi_P^C + 4.32928$

In both series of results a qualitative improvement of the model was obtained when 3 descriptors were considered. The linear equations involving 4 descriptors were also indicated but the possibility to deal with an over-parameterised model should be taken into account. This idea can be also applied to other families presented in this study.

3.1.4.2 Topological study of 17 artemisinin analogs

This family is composed by a set of 17 analogs of 10-deoxoartemisinin substituted at positions C-3 and C-9 [93]. In vitro molecular activities are of the same nature as the ones reported for the previous system.

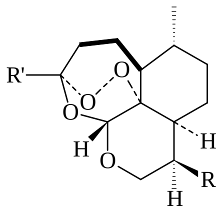


Figure 28. General molecular structure of the analogues of 10-deoxoartemisinin antimalarial molecules of system 2.

Table 40. Activities and structures of the molecules of system 2. Activity obtained in reference [79] in the same way as previously.

Molecule	R'	R	Relative activity	
			D6	W2
1	CH_3	CH_3	659	567
2	CH_3	H	237	190
3	CH_3	CH_3CH_2	914	466
4	CH_3	$CH_3(CH_2)_2$	473	550
5	CH_3	$CH_3(CH_2)_3$	5826	2090
6	CH_3	$CH_3(CH_2)_4$	170	145
7	CH_3	$C_6H_5(CH_2)_3$	5073	2506
8	CH_3	*p*-$ClC_6H_4(CH_2)_3$	6991	3317
9	CH_3CH_2	H	10	10
10	$CH_3(CH_2)_2$	H	722	685
11	$CH_3(CH_2)_3$	H	653	556
12	$(CH_3)_2CHCH_2$	H	183	250
13	$C_6H_5(CH_2)_4$	H	336	380
14	$C_6H_5(CH_2)_2$	H	6	2
15	*p*-$ClC_6H_4(CH_2)_3$	H	13	28
16	$(CH_2)_2CO_2Et$	H	422	506
17	$(CH_2)_2CO_2H$	H	0.09	0.09

Table 41. Summary of the molecular data set.

Molecular Data Set	
Compounds	analogues of 10-deoxoartemisinin
Number of Compounds	17
Activity	Relative Activity (IC_{50} Plasmodium Falciparum, D6 strain) Relative Activity (IC_{50} Plasmodium Falciparum, W2 strain)

Results and discussion

Table 42. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the D6 activity.

Nº descr.	r^2	$\mathrm{r}^2_{\mathrm{cv}}$	Linear model equation (log D-6 activity)
3	0.711	0.486	$y = 44.6304\,{}^{4}\chi_{C}^{T}$ - $5.62376\,{}^{3}\chi_{C}^{S}$ + $2.12991\,{}^{5}\chi_{PC}^{C}$ - 18.6223
4	0.808	0.639	y = $33.6116\,{}^{3}\chi_{C}^{T}$ - $7.93018\,{}^{4}\chi_{PC}^{T}$ - $37.1546\,{}^{3}\chi_{C}^{S}$ + $2.65114\,{}^{5}\chi_{PC}^{C}$ - 6.38758

Table 43. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the W2 activity.

Nº descr.	r^2	$\mathrm{r}^2_{\mathrm{cv}}$	Linear model equation (log D-6 activity)
3	0.699	0.478	y = -6.40997 ${}^{7}\chi_{P}^{T}$ + 2.43942 ${}^{4}\chi_{P}^{C}$ + 1.27781 ${}^{7}\chi_{PC}^{C}$ – 14.8043
4	0.824	0.657	y = 31.5817 ${}^{3}\chi_{C}^{T}$ – 7.79599 ${}^{4}\chi_{PC}^{T}$ - 34.9313 ${}^{3}\chi_{C}^{S}$ + 2.77878 ${}^{5}\chi_{PC}^{C}$ - 6.50120

Table 44. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the D6 activity. The models were obtained without considering the molecule number **17**.

Nº descr.	r^2	$\mathrm{r}^2_{\mathrm{cv}}$	Linear model equation (log D-6 activity)
2	0.611	0.460	y = 2.32173·10^{-5} $p^T(9)$ - 22.0545 ${}^{7}\chi_{CH}^{C}$ + 12.0052
3	0.682	0.541	y = 6.33065·10^{-6} Z^T – 1.03067 ${}^{7}\chi_{P}^{C}$ - 29.0119 ${}^{7}\chi_{CH}^{C}$ +19.2279
4	0.837	0.744	y = -6.99820 ${}^{3}\chi_{P}^{T}$ + 4.42378·10^{-3} ${}^{3D}MTI^{S}$ - 3.17224 ${}^{3}\chi_{C}^{C}$ + 1.57109 ${}^{7}\chi_{PC}^{C}$ + 24.2764

Table 45. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the W2 activity. The models were obtained without considering the molecule number **17**.

Nº descr.	r^2	$\mathrm{r}^2_{\mathrm{cv}}$	Linear model equation (log D-6 activity)
2	0.613	0.443	y = 2.18192·10^{-5} $p^T(9)$ - 20.7077 ${}^{7}\chi_{CH}^{C}$ + 11.3316
3	0.762	0.627	y = 33.2729 B^T -7.59291 ${}^{4}\chi_{PC}^{S}$ +5.73675·10^{-3} MTI^C –48.8439
4	0.881	0.784	y = -7.09264 ${}^{3}\chi_{P}^{T}$ + 4.58703·10^{-3} ${}^{3}MTI^{S}$ - 7.39280 ${}^{3}\chi_{C}^{S}$ + 1.57086 ${}^{7}\chi_{PC}^{C}$ + 26.5073

In the experimental paper [93] the authors reported some results related to qualitative SAR but no satisfactory QSAR equations were obtained even using, among others, topological and connectivity/shape indices from the Tsar [94] program. The present results show how the combination of several kinds of TQSI activates a synergic effect which leads to acceptable linear models.

Molecule **17** exhibited a very low value of both experimental values. The point lies quite far away from the remaining molecular data. The model was enhanced just removing the analogue; such an effect can be related to the fact that activity of molecule number **17** is substantially different from the ones corresponding to other molecules of the same family and having similar chemical structures.

3.1.4.3 Topological study of 21 β-metoxyacrylates

The antimalarial system 3 is constituted by 21 β-metoxyacrylates having different linkers against chloroquine-sensitive (NF54) and chloroquine-resistant (K1) P. falciparum in vitro [95]. Activity was reported as IC_{50} in nmol l^{-1}, a quantity derived from the original data of reference [95].

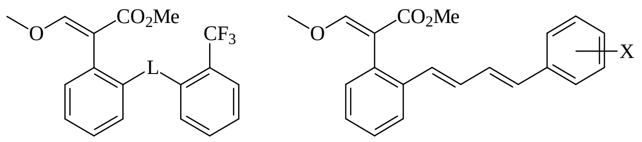


Figure 29. Molecular structures of the molecules of system 3.

Table 46. Molecular structures and activities of the antimalarial molecules of system 3.

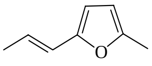
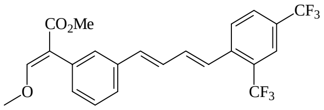
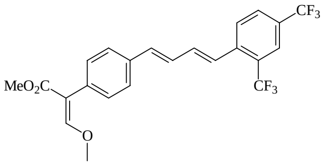
Molecule	Structure	Activity	
		NF54	K1
1	L = $CH_2CH_2SCH_2$	14.4	75.8
2	L = $CH_2CH{=}CHCH_2$	8.31	21.5
3	L = $CH_2CH{=}CHCH_2SCH_2$	4.29	6.15
4	L = $CH_2CH_2ON{=}C(CH_3)CH_2$	0.42	1.6
5	L = $CH_2CH{=}CHCH{=}CHCH_2$	0.15	0.39
6	L = $(CH_2)_6$	1.38	3.9
7	L = 	0.91	4.2
8	X = H	2.5	11.5
9	X = 2-Cl	0.25	1.01
10	X = 2-CN	1.51	4.63
11	X = 3-F	4.43	24.8
12	X = 3-CF_3	20.1	43.
13	X = 3-Br	15.3	78.6
14	X = 4-Cl	1.47	5.64
15	X = 2,4-di-CF_3	0.13	0.28
16	X = 2,4-di-Cl	0.08	0.26
17	X = 2,4-di-Me	0.09	0.14
18	X = 2-Cl, 4-F	0.16	0.51
19	X = 3-MeO, 2-NO_2	0.27	1.40
20		385.4	868.6
21		>11000	>11000

Table 47. Summary of the molecular data set.

Molecular Data Set	
Compounds	β-metoxyacrylates
Number of Compounds	21
Activity	Log IC_{50} for chloroquine-sensitive NF54 strain
	Log IC_{50} for chloroquine-resistant K1strain

Results and discussion

Table 48. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the NF54 activity. [a]The statistical significance parameter is p=0.00062.

Nº descr.	r^2	r^2_{cv}	Linear model equation (log D-6 activity)
2	0.652	0.508[a]	$y = -4.26130\,{}^{3}\chi_{P}^{S} + 14.9392\,{}^{7}\chi_{P}^{S} + 8.06932$
3	0.797	0.726	$y = -2.62207\,{}^{3}\chi_{P}^{T} + 9.55966\,{}^{7}\chi_{P}^{T} - 6.38591\,{}^{8}\chi_{P}^{C} + 8.78029$
4	0.815	0.750	$y = -2.45384\,{}^{3}\chi_{P}^{T} + 10.6961\,{}^{7}\chi_{P}^{T} - 7.62112\,{}^{8}\chi_{P}^{C} - 0.328296\,{}^{2}\chi_{P}^{S} + 9.22398$

Table 49. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the K1 activity.

Nº descr.	r^2	r^2_{cv}	Linear model equation (log D-6 activity)
3	0.759	0.676	$y = -1.03123\,p^{T}(1) + 12.5640\,{}^{7}\chi_{P}^{T} - 9.69714\,{}^{8}\chi_{P}^{C} + 13.4701$
4	0.870	0.766	$y = 1.99351\,{}^{0}\chi_{P}^{T} - 13.4758\,{}^{3}\chi_{P}^{S} + 71.7140\,{}^{9}\chi_{P}^{S} - 50.3532\,{}^{8}\chi_{C}^{C} + 12.3526$

Linear models obtained for the two reported activities (molecules number **20** and **21** were discarded). For the last molecule only a unique semiquantitative value is available. Apparently, good results are obtained when considering molecule number **21** but this could be due to the artificial and ambiguous extrapolated experimental value attached to it. On the other hand, the low activity of molecule number **20** also seemed to distort the molecular data cloud. If the first **19** molecules are taken into account, the molecular activity distribution becomes more uniform and satisfactory models are obtained.

3.1.4.4 Topological study of 17 flavines

This molecular group consists on a series of 17 3-methyl-10-(substituted-phenyl)flavins. The activity was reported as the action versus the lethal parasite Plasmodium vinckei in mice [82-100]. In particular, activity was given as the effective dose (in mmol kg^{-1} 10^{-3}) required obtaining a parasitemia of 40% in 48h (ED$_{40}$) [82].

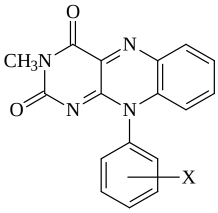


Figure 30. General structure of the 3-methyl-10-(substituted-phenyl)flavins of molecular system 4.

Table 50. Molecules and biological activity conforming the antimalarial system number 4.

Molecule	Structure	ED$_{40}$
1	4-Br	38.4
2	4-Cl	38.8
3	3,5-di-Cl	40.2
4	3-CF$_3$	79.3
5	3-Cl, 5-Me	85.7
6	4-F	103
7	3,5-di-Me	105
8	4-CF$_3$	135
9	4-OMe	138
10	3-Br	148
11	4-Cl, 3-Me	182
12	3,4-di-Me	210
13	3,5-di-OMe	219
14	3-Cl	229
15	H	248
16	4-Et	281
17	3-Cl, 4-Me	456

Table 51. Summary of the molecular data set.

Molecular Data Set	
Compounds	3-methyl-10-(substituted-phenyl)flavins
Number of Compounds	17
Activity	ED_{40} Plasmodium vinckei

Results and discussion

Table 52. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the ED_{40} activity. [a]Statistical significance p=0.0013.

Nº descr.	r^2	r^2_{cv}	Linear model equation (log ED_{40} activity)
3	0.679 [a]	0.510[a]	$y = 0.0443328\, p^T(2) – 16.3798\, {}^9\chi^T_P + 3.17847\, {}^3\chi^S_P + 8.07947$
4	0.777	0.673	$y = 9.44469{\cdot}10^{-3}\, MTI^T - 1.41134\, {}^5\chi^T_P - 0.0111852\, {}^{3D}MTI^S - 967.435\, {}^7\chi^S_C + 77.8123$

3.1.4.5 Topological study of 27 phenothiazine derivatives

This family was constituted originally by 27 phenothiazine derivatives [97] with capacity to inhibite the Plasmodium falciparum cysteine protease falcipain activity. IC_{50} for inhibition of falcipain activity measured as the hydrolysis of Z-Phe-Arg-AMC [83].

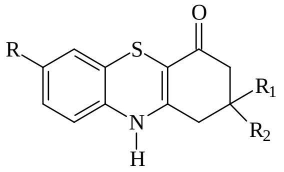


Figure 31. Molecular structures of the 16 molecules taken form reference [83] which conform the system number 5.

Table 53. Biological activities and codification of the molecules of system number 5.

Molecule	R	R_1	R_2	IC_{50}
1	Cl	H	H	40
2	Cl	CH_3	CH_3	30
3	Cl	H	C_6H_5	10
4	Cl	H	3-$CH_3OC_6H_4$	20
5	Cl	H	4-$CH_3OC_6H_4$	10
6	Cl	H	2,3-$(CH_3O)_2C_6H_3$	10
7	Cl	H	3,4-$(CH_3O)_2C_6H_3$	30
8	Cl	H	4-ClC_6H_4	4
9	Cl	H	2,4-$Cl_2C_6H_3$	10
10	F	CH_3	CH_3	60
11	F	H	C_6H_5	20
12	F	H	3-$CH_3OC_6H_4$	20
13	F	H	2,3-$(CH_3O)_2C_6H_3$	20
14	F	H	2,4-$Cl_2C_6H_3$	10
15	F	H	4-ClC_6H_4	5
16	F	H	3,4-$(CH_3O)_2C_6H_3$	20

Table 54. Summary of the molecular data set.

Molecular Data Set	
Compounds	phenothiazine derivatives
Number of Compounds	16
Activity	Log IC_{50} Plasmodium Falciparum cysteine falcipain

Results and discussion

Table 55. Linear models having a maximal r^2_{cv} value for every set of descriptors. The variable y stands for the logarithm of the ED_{40} activity. [a]Statistical significance p=0.0053.

Nº descr.	r^2	r^2_{cv}	Linear model equation (log ED_{40} activity)
1	0.575[a]	0.437[a]	$y = -7.18658\,{}^7\chi^C_{CH} + 3.91614$
2	0.765	0.671	$y = 24.7079\,{}^8\chi^S_{CH} - 0.680162\,{}^2\chi^C_P + 1.67282$
3	0.884	0.811	$y = 1.35265\,{}^3\chi^C_P - 16.0903\,{}^7\chi^C_{CH} - 8.27100\,{}^9\chi^C_P + 6.72932$

Several trials were carried out among this molecular set and no satisfactory results were yet obtained except for one case. If the set of 11 molecules having a sulphur atom is considered, a good linear model was obtained with 2 descriptors. Results could apparently be improved by means of the construction of models involving 3 or more descriptors but this could be a spurious result arising from the system over-parameterisation.

General Conclusions

For each family of n members a matrix of descriptors of dimension $n \times m$ (where n is the number of molecules and m is number of descriptors) was obtained. Originally, for all the systems m was $46 \times 3 = 138$ because the matrix of indices was obtained by juxtaposition of the tree kinds of available, classical topological, Coulomb, and Cioslowski matrices. Then, after removing null or other kind of irrelevant columns (for instance, those originating linear dependencies), for the studied system 4 the parameter m became 121 and $m = 126$ for the rest.

From the obtained linear models it can be seen that connectivity indices are used many times and, among them, these of higher order are commonly requested. Indices derived from the classical topological matrix are used but many indices coming from matrices S and C are also employed too. These facts indicate that this approach really contributes to improve the old methodological capabilities. Nevertheless, at least for the families studied here, only in some cases the Hosoya index or its contributions are used and, when requested, only those coming from the classical definition occur in the models. The presence or absence of indices in the linear models is related to the nature of the studied molecular sets and activities

Contribution

► Besalú, E; Gallegos, A.; Carbó-Dorca, R. Topological Quantum Similarity Indices and Their Use in QSAR: Application to Several Families of Antimalarial Compounds. In *MATCH-Communications in Mathematical and in Computer Chemistry (Special issue dedicated to Prof. Balaban)*. Diudea, M.; Ivanciuc, O. (Eds.) *MATCH-Commun. Math. CO, 44,* **2001**, 41-64.

3.2 Estrogenic Activity

Introduction

Endocrine disruptive (ED) chemicals [98] are present in the aquatic environment as pollutants, and in food and water as antioxidants and metabolites of other anthropogenic chemicals [99]. These environmental estrogens may result either from naturally occurring compounds, i.e. plants, phytoestrogens and agricultural products, or synthetically produced chemicals, such as pesticides, plastics, combustion by-products.

Endocrine disrupting chemicals can bind to estrogen receptors (ER), thus interfering with genetic functions such as sexual development and reproductive fecundity [101]. Due to the deleterious effects on human health and the environment [102], the U.S. Environmental Protection Agency (EPA), and the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC) [103] have run several screening and testing programs to identify such compounds and develop various computational tools to model ligand binding to the ER [104-106].

Thus, SAR studies have been used extensively since the 1930s [107] to model the interactions between a ligand and receptor and to estimate the effects of estrogenic compounds for hazard identification, human health risk assessments [108] and wildlife exposure studies [109], in order to avoid costly empiric evaluations based on in vivo and in vitro bioassays [110-112]. In particular, several 3D-QSAR similarity studies predict ligand-hormone receptor binding affinities, i.e. the Comparative Molecular Field Analysis (CoMFA) [113-114] approach, and the Common Reactive Pattern approach (CoRePa) [115-116].

3.2.1 Semiquantitative study of 120 Aromatic Compounds

This study is based on the analysis of the reported in vitro estrogenicity data of a source of 120 structurally diverse aromatic chemicals that constitute estrogenic endocrine disruptors. The study attempted to develop structure-based methods to evaluate predict the ability of compounds to promote an estrogenic effect, and identify potential ligands.

The main molecular features required for the pharmacophore to bind to the estrogenic receptor [117-118] have clear analogies with the the estradiol molecule, considered to be one of the most potent estrogens:

- A hydroxy (phenolic) group at the C-3 position of the aromatic A-ring
- A ketone or alcohol functional group at site 17
- Four hydrophobic centers, corresponding to the A to D rings of estradiol
- A hydrophobic group at the para position relative to the phenolic hydroxy group
- The shape of the ligand must be constrained to fit the estrogen receptor pocket

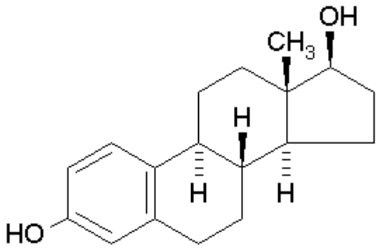
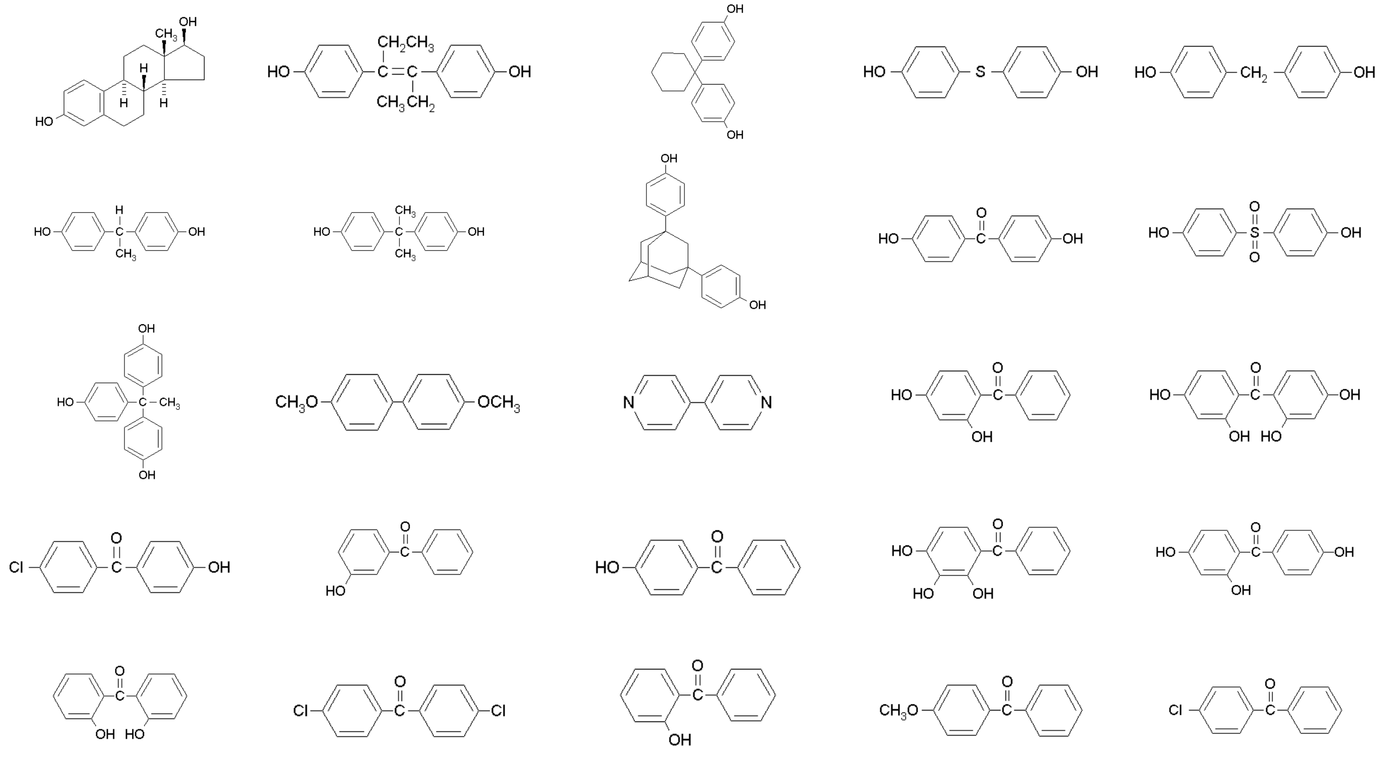
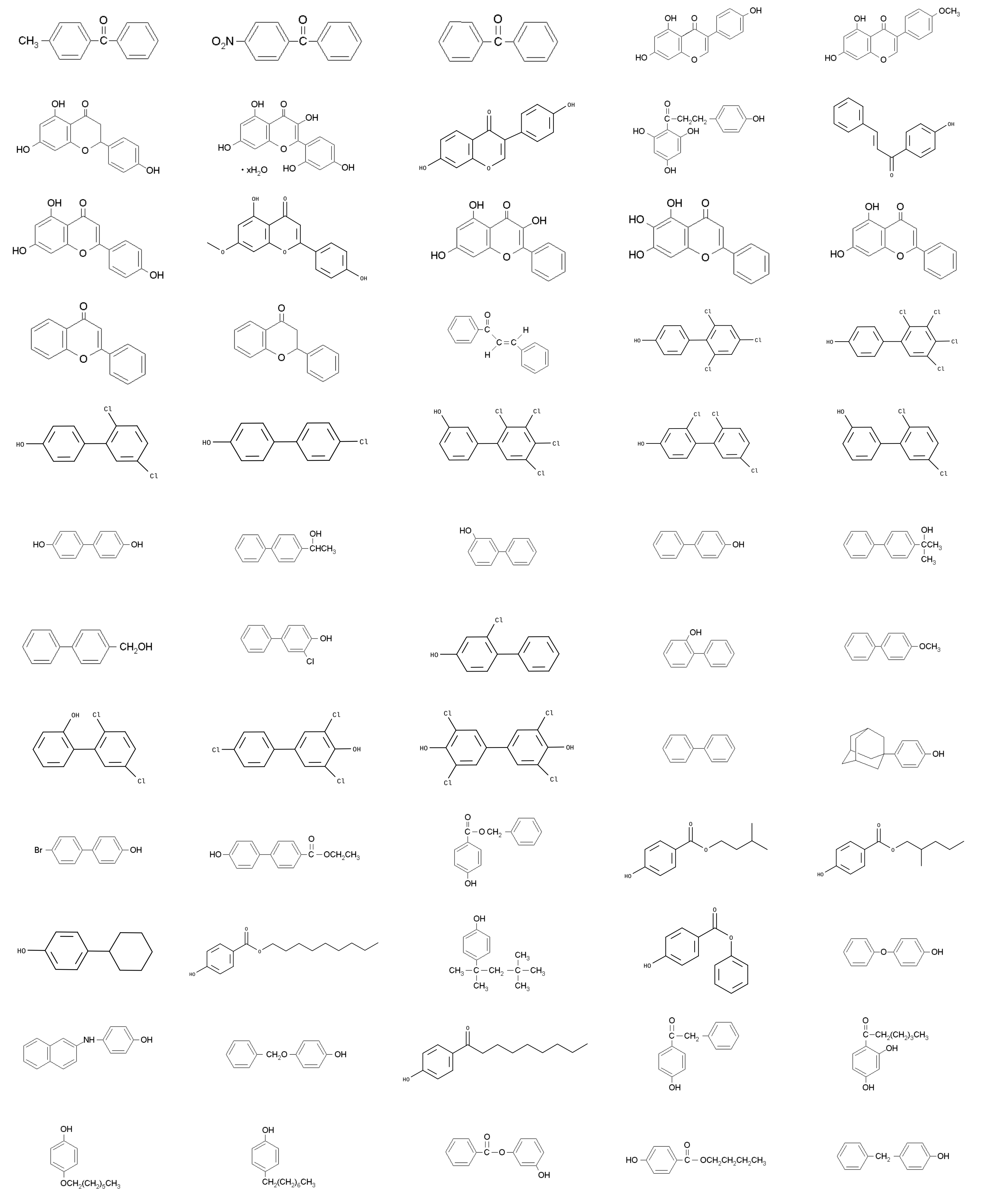


Figure 32. Structure of estradiol molecule

The studied compounds consisted of a group of 120 assorted aromatic compounds, covering a wide range of chemical classes, with a broad degree of structural diversity, including bisphenols, benzophenones, flavonoids, biphenyls, phenols and other aromatic and bi-aromatic chemicals.




CH$_3$ C O
O$_2$N C O
C O
OH O OH HO O
OH O OCH$_3$ HO O
OH O HO O OH
OH O OH HO O · xH$_2$O HO OH
O OH HO O
O C CH$_2$CH$_2$ OH HO OH OH
OH O
OH O HO O OH
OH O O O OH
OH O OH HO O
OH O HO HO O
OH O HO O
O O
O O
O C C=C H H
HO Cl Cl Cl
HO Cl Cl Cl Cl
Cl HO Cl
HO Cl
HO Cl Cl Cl Cl
Cl Cl HO Cl
HO Cl Cl
HO OH
OH CHCH$_3$
HO
OH
OH CCH$_3$ CH$_3$
CH$_2$OH
OH Cl
Cl HO
OH
OCH$_3$
OH Cl Cl
Cl Cl OH Cl
Cl Cl HO OH Cl Cl
OH
Br OH
HO O C OCH$_2$CH$_3$
O C O CH$_2$ OH
O O HO
O O HO
HO
O O HO
OH CH$_3$ CH$_3$ C CH$_2$ C CH$_3$ CH$_3$ CH$_3$
O O HO
O OH
NH OH
CH$_2$O OH
O HO
O C CH$_2$ OH
O C CH$_2$(CH$_2$)$_3$CH$_3$ OH OH
OH OCH$_2$(CH$_2$)$_5$CH$_3$
OH CH$_2$(CH$_2$)$_6$CH$_3$
O C O OH
HO O C OCH$_2$CH$_2$CH$_2$CH$_3$
CH$_2$ OH

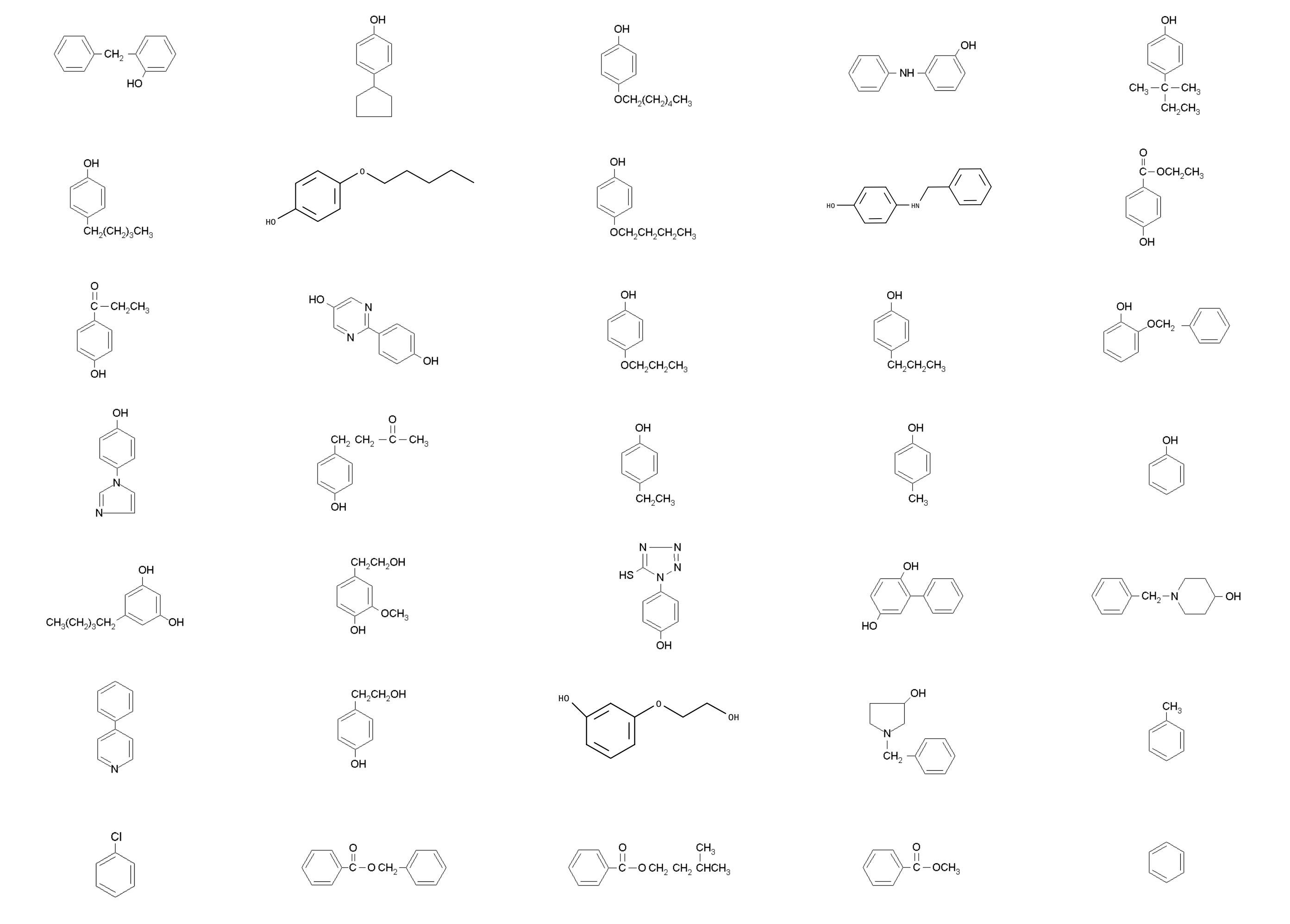


Figure 33. Structure for 120 compounds.

For the development of SARs for the estrogenic gene activation, the whole set was split into five classes attending to the chemical group/substitution, following the classification of Schultz [119].

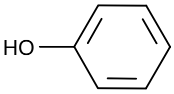
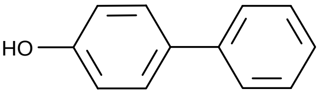
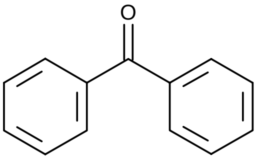
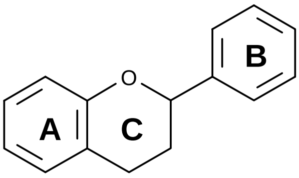
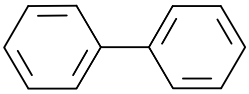
FUNCTIONAL GROUP CLASSIFICATION	STRUCTURES
Phenols	
Bisphenols	
Benzophenones	
Flavonoids	
Biphenyls	

Figure 34. **Functional group categories.**

Besides, a more fundamental structural basis classification criteria [120], was used to split the complete set into several categories, attending to their topological structure, that is, compounds with only one aromatic ring, biphenyl-like structures, two aromatic rings separated with one bridging atom, molecules with either two or three bridging atoms.

The reported biological activity, gene expression for the α-human estrogen receptor (hERα) was measured in vitro using the recombinant yeast assay [121], performed according to the protocol of Schultz et al [122]. For each compound, the concentration eliciting an activity equal to 50% of the positive control 17ß-estradiol was determined. Those compounds with a maximum gene expression at a concentration less than 50% of 17ß-estradiol were noted as being detectable, but that an EC_{50} could not be established. For modelling purposes, experimental data were classified as active (A) or inactive (I) [123].

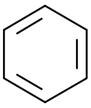
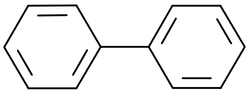
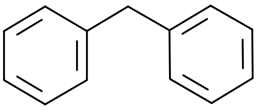
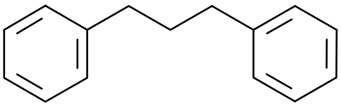
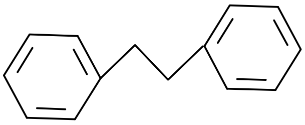
N. CLASS	Classification	STRUCTURES
1	A single aromatic ring	
2	Two directly bonded aromatic rings	
3	Two aromatic rings separated by one bridging atom	
4	Two aromatic rings two benzene rings separated by two or three bridging atoms	 

Figure 35. Fundamental structural basis for classification.

Table 56. Structures and relative estrogenic gene activation for 120 compounds.

N	Compound	EC_{50}	Binary Activity
1	17-β-estradiol	3.91e-11	A
2	4,4'-diethylethylene bisphenol	9.11e-11	A
3	4,4'-cyclohexylidene bisphenol	4.43e-08	A
4	4,4'-thiodiphenol	7.15e-08	
5	Bis (4-hydroxyphenyl) methane	8.15e-08	A
6	4,4'-ethylidene bisphenol	2.28e-07	A
7	Bisphenol A	4.28e-07	A
8	4,4'-(1,3-adamantanediyl) bisphenol	6.10e-07	A
9	4,4'-dihydroxybenzophenone	6.33e-07	A
10	Bis(4-hydroxyphenyl)sulphone	7.50e-05	A
11	1,1,1-tris(4-hydroxyphenyl)ethane	1.03e-03	A
12	4,4'-dimethoxybiphenyl	non active	I
13	4,4'-dipyridyl	non active	I
14	2,4-dihydroxybenzophenone	4.23e-08	A
15	2,2',4,4'-tetrahydroxyl benzophenone	1.98e-07	A
16	4-chloro-4'-hydroxy benzophenone	4.20e-07	A
N	**Compound**	**EC_{50}**	**Binary**

			Activity
17	3-hydroxybenzophenone	4.93e-07	A
18	4-hydroxybenzophenone	9.84e-07	A
19	2,3,4-trihydroxybenzophenone	1.27e-06	A
20	2,4,4'-trihydroxybenzophenone	1.41e-07	A
21	2,2'-dihydroxybenzophenone	non active	I
22	4,4'-dichlorobenzophenone	non active	I
23	2-hydroxybenzophenone	non active	I
24	4-methoxybenzophenone	non active	I
25	4-chlorobenzophenone	non active	I
26	4-methylbenzophenone	non active	I
27	4-nitrobenzophenone	non active	I
28	Benzophenone	non active	I
29	Genistein (4',5,7-trihydroxyisoflavone)	1.81e-07	A
30	Biochanin A (5,7-dihydroxy-4'-methoxy isoflavone)	6.87e-07	A
31	Naringenin (4',5,7-trihydroxyflavanone)	2.30e-05	A
32	Morin hydrate (3,3',5,5',7-pentahydroxyflavone)	8.80e-05	A
33	Daidzein (4',7-dihydroxyisoflavone)	4.92e-05	A
34	Phloretin (2',4,4',6'-tetrahydroxychalcone)	1.80e-05	A
35	4'-hydroxychalcone	1.40e-05	A
36	Apigenin (4',5,7-trihydroxyflavone)	no EC50	-
37	Genkwanin (4',5-dihydroxy-7-methoxyflavone)	no EC50	-
38	Galangin (3,5,7-trihydroxyflavone)	non active	I
39	Baicalein (5,6,7-trihydroxyflavone)	non active	I
40	Chrysin (5,7-dihydroxyflavone)	non active	I
41	Flavone	non active	I
42	Flavanone	non active	I
43	Trans-chalcone	non active	I
44	2',4',6'-trichloro-4-biphenylol	1.29e-09	A
45	2',3',4',5'-tetrachloro-4-biphenylol	6.30e-09	A
46	2',5'-dichloro-4-biphenylol	3.00e-08	A
47	4'-chloro-4-biphenylol	5.98e-08	A
48	2',3',4',5'-tetrachloro-3-biphenylol	1.58e-07	A
49	2,2',5'-trichloro-4-biphenylol	1.78e-07	A
50	2',5'-dichloro-3-biphenylol	2.04e-07	A
51	4,4'-biphenyldiol	2.63e-07	A
N	**Compound**	**EC_{50}**	**Binary**

			Activity
52	4-(1-hydroxyethyl) biphenyl	7.88e-06	A
53	3-hydroxybiphenyl	9.18e-06	A
54	4-hydroxybiphenyl	1.15e-06	A
55	4-(2-hydroxypropyl)biphenyl	1.84e-06	A
56	4-biphenylmethanol	2.12e-06	A
57	3-chloro-4-biphenylol	3.82e-06	A
58	2-chloro-4-biphenylol	3.82e-06	A
59	2-hydroxybiphenyl	1.84e-05	A
60	4-methoxybiphenyl	3.39e-05	A
61	2',5'-dichloro-2-biphenylol	5.23e-05	A
62	3,4',5-trichloro-4-biphenylol	non active	I
63	3,3',5,5'-tetrachloro-4,4'-biphenyldiol	non active	I
64	Biphenyl	non active	I
65	4-(1-adamantyl)phenol	8.55e-09	A
66	4-(4-bromophenyl)phenol	2.37e-08	A
67	Ethyl-4'-hydroxy-4-biphenyl carboxylate	5.03e-08	A
68	Benzyl-4-hydroxybenzoate	1.07e-07	A
69	Isoamyl-4-hydroxybenzoate	1.17e-07	A
70	2-ethylhexyl-4'-hydroxy benzoate	1.36e-07	A
71	4-cyclohexylphenol	1.39e-07	A
72	Nonyl-4-hydroxybenzoate	1.65e-07	A
73	4-(tert-octyl)phenol	1.77e-07	A
74	Phenyl-4-hydroxybenzoate	2.28e-07	A
75	4-phenoxyphenol	2.62e-07	A
76	N-(4-hyroxyphenyl)-2-naphthylamine	4.15e-07	A
77	4-(benzyloxy)phenol	5.43e-07	A
78	4-hydroxyoctanophenone	8.85e-07	A
79	Benzyl-4-hydroxyphenyl ketone	9.20e-07	A
80	4-hexanoyl resorcinol	9.38e-07	A
81	4-heptyloxyphenol	1.88e-06	A
82	4-octylphenol	1.89e-06	A
83	Resorcinol monobenzoate	1.95e-06	A
84	Butyl-4-hydroxybenzoate	2.01e-06	A
85	4-hydroxydiphenylmethane	2.12e-06	A
86	2-hydroxydiphenylmethane	2.12e-06	A
N	**Compound**	**EC_{50}**	**Binary**

			Activity
87	4-cyclopentyl phenol	2.41e-06	A
88	4-hexyloxyphenol	4.02e-06	A
89	3-hydroxydiphenylamine	4.20e-06	A
90	4-(tert-pentyl)phenol	4.76e-06	A
91	4-n-pentylphenol	9.50e-06	A
92	4-pentyloxyphenol	1.73e-05	A
93	4-butoxyphenol	1.88e-05	A
94	N-benzyl-4-hydroxyaniline	6.27e-05	A
95	Ethyl-4-hydroxybenzoate	7.52e-05	A
96	4-hydroxypropiophenone	8.32e-05	A
97	2-(4-hydroxyphenyl)-5-pyrimidinol	1.33e-04	A
98	4-propoxyphenol	1.64e-04	A
99	4-propylphenol	1.84e-04	A
100	2-(benzyloxy)phenol	2.50e-04	A
101	4-(Imidazol-1-yl)phenol	1.25e-03	A
102	4-(4-hydroxyphenyl)-2-butanone	1.22e-03	A
103	4-ethylphenol	No EC50	-
104	4-methylphenol	non active	I
105	phenol	non active	I
106	5-pentylresorcinol	non active	I
107	Homovanillyl alcohol	non active	I
108	1-(4-hydroxyphenyl)-1H-tetrazole-5-thiol	non active	I
109	Phenyl hydroquinone	non active	I
110	1-benzyl-4-hydroxypiperidine	non active	I
111	4-phenylpyridine	non active	I
112	2-(4-hydroxyphenyl)ethanol	non active	I
113	O-(2-hydroxyethyl) resorcinol	non active	I
114	1-benzyl-3-pyrrolidinol	non active	I
115	Toluene	non active	I
116	Chlorobenzene	non active	I
117	Benzyl benzoate	non active	I
118	Isoamyl benzoate	non active	I
119	Methyl benzoate	non active	I
120	Benzene	non active	I

Table 57. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Aromatic compounds
Type of Compounds	Bisphenols, benzophenones, flavonoids, biphenyls, phenols and other aromatic and bi-aromatic chemicals
Number of Compounds	120
Activity	EC_{50} for the α-human estrogen receptor (hERα)
Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	Semiempirical AM1 level, using MOPAC 6.0 [6]
Density Function	Fitted first-order Promolecular ASA (PASA), 3-21G basis set
MQSM Operator	Overlap MQSM
Molecular Alignment	Maximum similarity superposition algorithm
Reduction of Dimensions	Principal Components Analysis (PCA) [18]
Correlation Method	Stepwise Multiple Linear Regression [18]
Classification Method	Multidimensional Discriminant Analysis (MDA) [18]
Validation	Internal Leave-One-Out Cross-Validation (LOO-CV) [18] External test set

Results and discussion

For purposes of statistical analysis, active compounds were assigned a value of 1, and inactive compounds a value of 0. The accuracy of the predictions for active compounds was calculated by dividing the number of active compounds correctly assigned by the model by the total number of active compounds; conversely, the accuracy on the predictions for inactive compounds was calculated as the proportion of correctly predicted inactive compounds, out of the total number of inactive compounds. Finally, the overall accuracy of the model was also calculated, considering all the compounds.

Molecular quantum similarity indices (MQSI), which mainly encode information regarding steric or electrostatic distribution on the surface of the molecule, and indicator variables, indicating the presence or absence of explicit structural features were computed [15]. Besides, additional physico-chemical properties were also calculated. To account for hydrophobicity, the logarithm of the octanol-water partition coefficient was calculated using the KOWWIN software [124].

Table 58. Physico-Chemical Descriptors and Indicator variables.

Descriptor	Abbreviation
Logarithm of octanol-water partition coefficient	Log P
Molecular weight	MW
Number of atoms	At
Number of carbons	C
Number of hydrogen bond donor groups	HB_don
Number of hydrogen bond acceptor groups	HB_acc
Number of hydroxyl groups	OH
Number of hydroxyl groups in para-position	p-OH
Number of rings	R
Number of benzenes	Bz
Number of phenols	Ph

Overlap MQSMs, highly sensitive to exact atom superpositions, provide similarity terms with reliable values, due to the fact that all the compounds in the data set have at least one aromatic ring. The use of this type of this measure mainly accounting for steirc interactions is in agreement with previously published studies [125].

Table 59. Multilinear Discriminant Analysis Classification for the entire set, made of 117 compounds with available reported activity values. The optimal model has been marked in italic face.

No Descriptors	Selected Descriptors	%Corr.Class. (adjustement)	%Corr.Class. (cross-validation)
# Ph	0.838	0.838	1
# Ph, PC3,	0.829	0.829	2
# Ph, PC3, # C	0.880	0.872	3
# Ph, PC3, # C, # Rings	0.872	0.855	4

Table 60. Results for the optimal QSAR model for the entire set. (% CC= percentage of correct classifications; % CCCV= percentage of correct classifications for cross-validation.)

Adjustment			Cross-Validation		
% CC	% CC for estrogenic compounds	% CC for inactive compounds	% CC	% CC for estrogenic compounds	% CC for inactive compounds
88.0 103/117	90.1 (73/81)	83.3 (30/36)	87.2 (102/117)	88.9 (72/81)	83.3 (30/36)

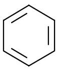
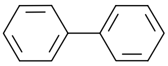
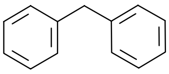
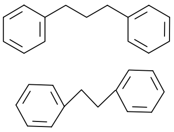
Table 61. Misclassified compounds.

Compound	Binary Activity	Predicted Activity	Misclassification
21	0	1	False positive
23	0	1	False positive
52	1	0	False negative
55	1	0	False negative
56	1	0	False negative
60	1	0	False negative
62	0	1	False positive
63	0	1	False positive
66	1	0	False negative
80	1	0	False negative
84	1	0	False negative
95	1	0	False negative
109	0	1	False positive
112	0	1	False positive

Table 62. Analysis of the misclassified compounds. ([a]intram HB stands for intramolecular Hydrogen Bonding).

False Positives		False Negatives	
N	Explanation	N	Explanation
21,23	OH form intram HB[a] with a carbonyl group	52,55, 56,60	absence of phenolic group, but O could could act as a weaker HB acceptor
62,63	OH form a weak intram HB[a] with a neighboring Cl	66	biased MQSM for the presence of a heavy atom (Br)
		80,84,95	lack of the appropriate hydrophobic area
		109,112	no clear explanation

Table 63. Selected optimal model for the fundamental structure-based classes.

Fund. Struc. Class	Number of molecules	N	Selected Descriptors	% CC	% CCCV	Misclassified Compounds	Predicted Activity
	37	3	p-OH, log P, PC6	97.3 (36/37)	91.9	104	1 False positive
	28	3	PC14, PC4, PC17	92.9 (26/28)	82.1	47, 57	2 False negatives
	28	2	p-OH, PC20	92.9 (26/28)	89.3	17, 86	2 False negatives
	23	2	Ph, PC1	100 (23/23)	100	-	-

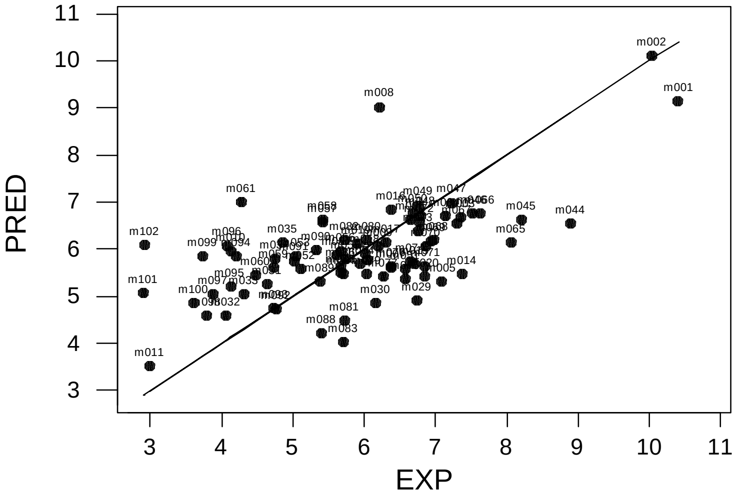


Figure 36. Predicted versus experimental activities.

The complete data set was split into a training and a test set, maintaining approximately the same number of molecules and a proportional ratio of active and inactive compounds for both of them. Test set selection was made according to two criteria. First, the two sets were chosen on the basis of the distribution of the three most significant descriptors, in order to ensure that training and evaluation sets contained chemicals representative of the diversity of structures. From knowledge of the spatial distribution of the compounds in the 3-D descriptor space, they were selected to provide a representative sample of structural diversity in both sets.

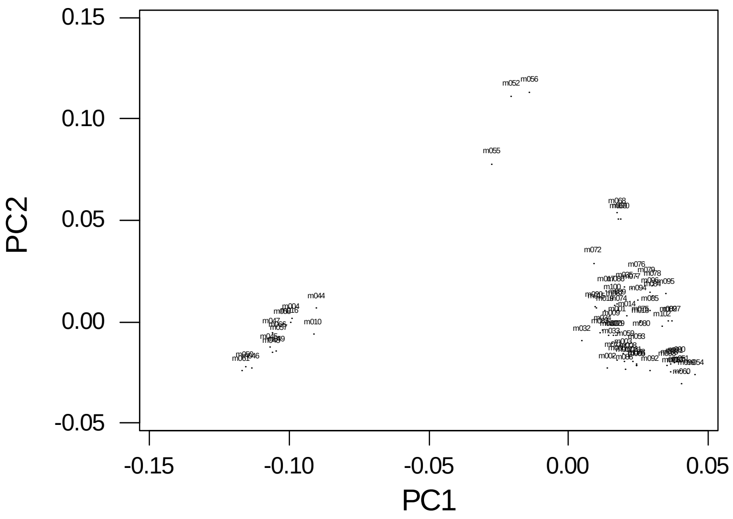


Figure 37. Second Principal Component (PC2) versus First Principal Component (PC1).

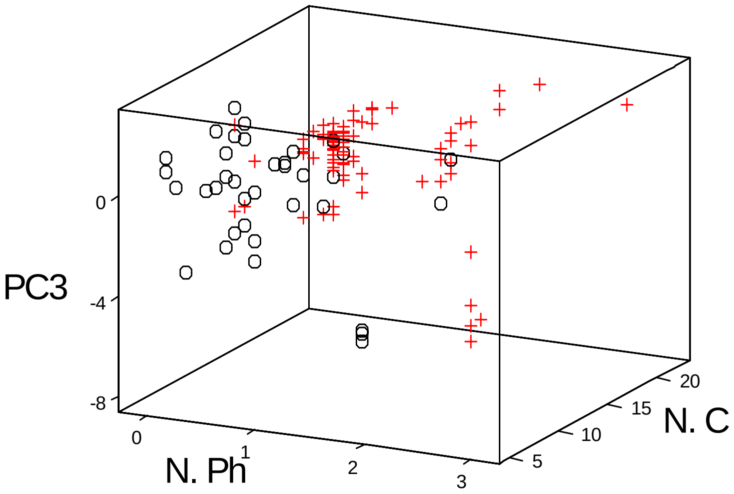


Figure 38. 3D descriptor space plot, where empty dots stand for the inactive compounds, while summation symbols stand for the active ones.

In a second approach to test set selection, the test set was extracted from the whole set randomly to ensure that the results were not conditioned for the distribution of the data. For the sake of comparison, training and test sets were swapped and modelling and validation reperformed.

Table 64. QSAR results for the training and test sets.

Selection Criteria	Training Test			Test Set		
	N.	% CC	% CCCV	N.	% CC	Misclassified Compounds
3D distribution of descriptors	59	77.9	76.3	58	87.9	23, 39, 40, 97, 98, 99
	58	91.4	84.5	59	83.1	21, 24, 56, 60, 62, 78, 80, 95, 96, 109
random	59	83.1	79.7	58	62.0	38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 69, 71, 73, 75, 83, 87, 91, 93, 99, 101
	58	89.7	87.9	59	79.7	21, 23, 39, 41, 55, 63, 68, 90, 96, 98, 102, 106

For the complete data set, the most important descriptors were some of the constitutional parameters accounting for structural features, such as the number of phenols and the number of carbons, together with the third principal component extracted from the matrix of molecular quantum similarity indices. The presence of a phenolic OH group resembling the 3-hydroxyl group of the estradiol molecule seems to be essential for effective binding to estrogen receptor. In addition, the number of carbons is indicative of the hydrophobic contribution from the ligand. Finally, the third principal component, accounting for the *4.3%* of the total variation of the similarity indices, was found to be the most predictive.

With the same tendency, within the class-based models classification it can be observed that most of the models include indicator variables, but the model corresponding to the bi-phenyl-like structure class. This may possibly be due to the exact interatomic alignment of the rigid biphenyl common pattern.

Comparison of results for the complete data and for separate classes indicates that slightly poorer results were obtained for the entire data set. For the whole set, results are biased to produce a greater percentage of false positives than false negatives. For drug discovery, false positives are of concern due to the unjustified cost of synthesizing a chemical with a low probability of being efficacious. Conversely, for regulatory purposes, the main goal is to minimize the rate of false negatives, in order to avoid any threat to public health; thus, results could be indicative of the utility of this model for health hazard assessment. As expected, cross-validated results for almost all the models showed a slight decrease in classification rate.

Concerning the results for the training and test tests, it can be observed that, in relation to the predictive ability of the test sets selected from knowledge of the 3D distribution of the descriptors was comparable to that for the complete data set. However, in the randomly chosen test sets, the correct classification rates were lower. The decrease in the predictivity of the models due to the test set selection method is to be expected. Analysis of the results for the different test sets indicates that more than the half of misclassified compounds in the two former test sets coincides with the misclassified ones in the latter cases.

The fact that almost the totality of compounds belonging to the first test set have been incorrectly assigned in the other test subsets, may be an indication of some intrinsic features in the remaining compounds hindering the structure-activity relationship; thus, some important information might be missing from these models and the selection of the test set. Conversely, for the misclassified compounds from the complete set, the classes based on fundamental structure and the test and training sets do not coincide. A reason for this behaviour could be the different type of information encoded in the different descriptors used to build the models.

Concerning the predictivity of the models, three compounds (**36**, **37** and **103**) do not possess an experimentally measured estrogenic binding affinity. When the complete set of molecules was used to predict the binding ability of these ligands, the model suggested all three compounds to be unequivocally active. This is in agreement with results obtained in the literature [126-128].

Computer-based methods provide the possibility to screen for potential estrogens, and predict their activity. Therefore, quantum similarity in conjunction with the use of structural descriptors is a valuable tool for QSAR and computer-aided drug design. The assignment of experimental data into discrete categories is useful in the use of high throughput screening (HTS) to identify lead compounds, especially in noncongeneric libraries where there is no common structure. The use of this qualitative approach allows for the correlation between chemical structures and discrete activities to be obtained, at least for preliminary compound selection. The success of the different proposed approaches confirms that, whereas hydrophobicity and the presence of some functional groups play a central role in determining toxic potency, the electronic effects derived from QST are able to discriminate between active and inactive estrogen ligands.

Appended Contribution

► Gallegos Saliner, A.; Amat, L.; Carbó-Dorca, R.; Schultz, T.W.; Cronin, M.T.D. Molecular Quantum Similarity Analysis of Estrogenic Activity. *J. Chem. Inf. Comput. Sci., 43*, **2003**, 1166-1176**.**

3.3 Antituberculotic Activity

Introduction

The current search for new antimycobacterial agents is very urgent as tuberculosis has become a major emerging opportunistic infection. The most common pulmonary tuberculosis is caused by infection by inhalation of the bacteria Mycobacterium tuberculosis and affects mainly the respiratory system, but also includes other vital organs.

In the past decades, tuberculosis cases began to increase even in the industrialized world. In addition, also the infections caused by atypical mycobacterial strains, e.g., Mycobacterium avium complex, show a rising occurrence among children, elderly, and HIV-infected patients [129-131], thus becoming a serious health problem.

Nowadays, although the mortality within the infected has significantly decreased, tuberculosis continues to be a devastating disease worldwide and is believed that approximately one-third of the world's population harbours Mycobacterium tuberculosis and it is at risk for developing the disease [132]. Indeed, it is estimated that about 8 million new cases of tuberculosis and 3 million deaths from this disease occur annually around the world [133].

Since the discovery of the first effective drug in the 1940s [134-135], no new drugs appeared on the market for 30 years. However, the developing resistance to conventional antituberculotics [136] has stimulated the research of new compounds. To such an extent, several QSAR studies [137-140] are devoted to the discovery of first line drugs.

In the course of the research into potential antimycobacterial and antifungal agents attention have been turned to benzanilides, thiobenzanilides, and related compounds. These groups of compounds are characterized by a wide spectrum of biological activities depending on the type of substitution [141-143]. It is known that the replacement of oxo group by thioxo group leads, in general, to the increase in the antimycobacterial activity [144].

3.3.1 Fragment-based Study of benzoxazines

In particular, this example case is focused in the study of benzoxazines, a very prospective group of new antimycobacterial compounds [145], which have been synthetized and tested for antimycobacterial activity.

QSAR models were carried out using fragment self-similarity measures, in order to provide a theoretical rationale for the observed increase of antimycobacterial activity induced by the replacement of the oxo group in 3-aryl-2H-1,3-benzoxazine-2,4(3H)-diones by sulphur. Especially, the antimycobacterial activity in six series of substituted 3-phenyl-2H-1,3-benzoxazine-2,4(3H)-dithiones and 3-phenyl-4-thioxo-2H-1,3-benzoxazine-2,4(3H)-diones was examined.

Table 65. Molecular sets of antituberculotics. Number of molecules and parent structures for each set. The number preceding each substituent indicates its position in the phenyl ring.

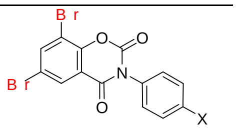
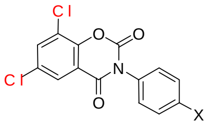
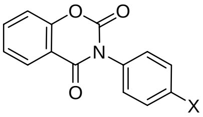
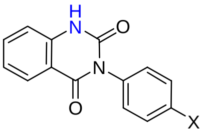
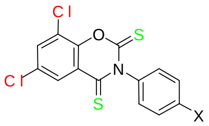
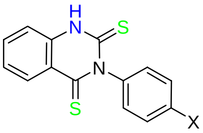
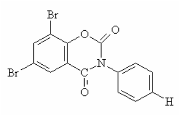
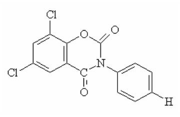
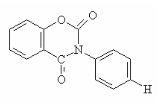
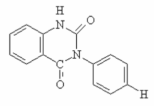
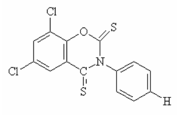
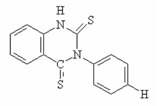
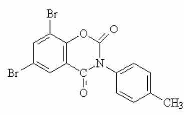
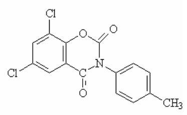
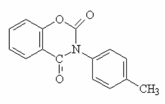
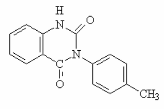
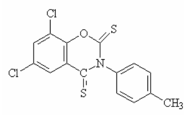
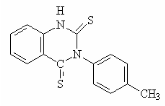
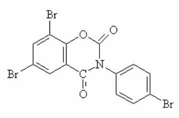
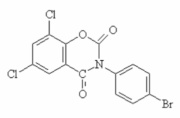
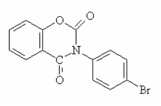
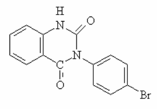
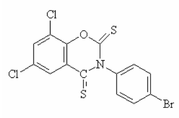
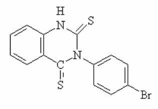
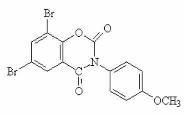
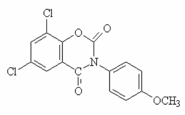
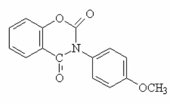
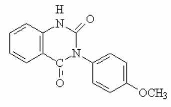
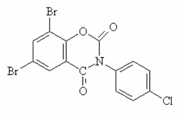
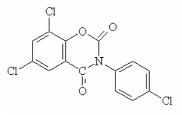
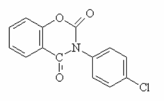
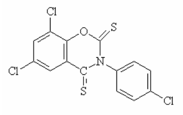
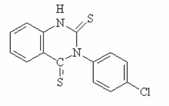
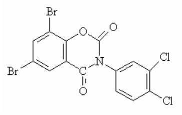
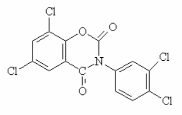
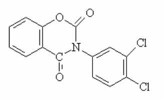
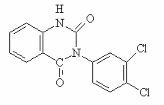
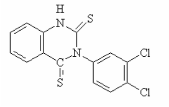
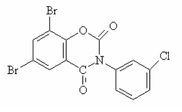
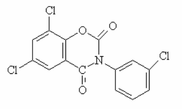
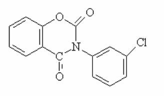
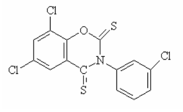
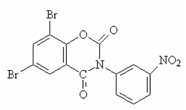
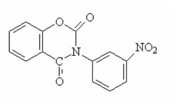
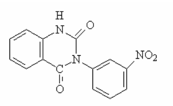
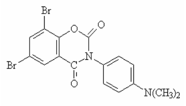
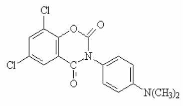
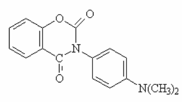
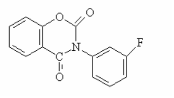
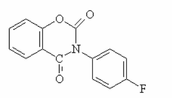
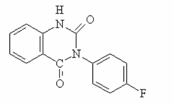
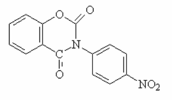
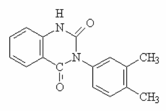
SET	Number of Molecules.	DERIVATIVES	PARENT STRUCTURE
M	9	6,8-dichloro-3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-dione [146]	
N	8	6,8-dibromo-3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-dione [146]	
S	11	3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-dione [147]	
U	8	3-phenylquinazoline-2,4(1*H*,3*H*)-dione [147]	
Y	5	6,8-dichloro-3-phenyl-2*H*-1,3-benzoxazine-2,4(3*H*)-dithione [148]	
Z	5	3-phenylquinazoline-2,4(1*H*,3*H*)-dithione [148]	

Table 66. Classification into subclasses and numbering for molecular structures of the entire antituberculotic set.

SUBST.	M	N	S	U	Y	Z
4H	M01 	N01 	S01 	U01 	Y01 	Z01 
$4CH_3$	M02 	N02 	S02 	U02 	Y03 	Z03 
4Br	M03 	N03 	S03 	U03 	Y02 	Z02 
$4OCH_3$	M04 	N04 	S04 	U04 		
4Cl	M05 	N05 	S05 		Y04 	Z04 
3Cl, 4Cl	M06 	N06 	S06 	U05 		Z05 
3Cl	M07 	N07 	S07 		Y05 	

SUBST.	**M**	**N**	**S**	**U**	**Y**	**Z**
$3NO_2$	M08 		S10 	U07 		
$4N(CH_3)_2$	M09 	N08 	S12 			
3F			S08 			
4F			S09 	U06 		
$4NO_2$			S11 			
$3CH_{3,}$ **$4\ CH_3$**				U08 		

Antimycobacterial activities of the sets have been evaluated in vitro [146-147], using Mycobacterium tuberculosis CNCTC My 331/88, i.e. different standard strains of mycobacteria, obtained from the Czech National Collection of Type Cultures (CNCTC), of the Institute of Public Health of Prague. The reported activity was expressed as the Minimum Inhibitory Concentration (MIC), i.e. the lowest concentration of a substance at which the inhibition of the growth occurs measured in µmol/l [147]. In order to have a narrower range of values, activity values for the tuberculous strain were logarithmically transformed.

Table 67. Distinctive substitution and antituberculotic activity (My 331/88) for each compound. Those activities marked with an asterisk have not been precisely measured.

Compound	Substitution	Log (M.tub)	Compound	Substitution	log (*M.tub*)
M01	4H	1,8	S07	3Cl	1,491
M02	4CH$_3$	1,5	S08	4F	2,097
M03	4Br	1,5	S09	3NO$_2$	1,204
M04	4OCH$_3$	1,8	S10	4NO$_2$	1,204
M05	4Cl	1,2*	S11	4N(CH$_3$)$_2$	2,097*
M06	3Cl, 4 Cl	0,9	U01	4H	2,699
M07	3Cl	0,9	U02	4CH$_3$	2,699
M08	3NO$_2$	1,2	U03	4Br	2,398
M09	4N(CH$_3$)$_2$	2,1*	U04	4OCH$_3$	2,097*
N01	4H	1,8	U05	3Cl, 4Cl	1,792*
N02	4CH$_3$	1,2	U06	4F	2,097*
N03	4Br	1,2	U07	3NO$_2$	2,398
N04	4OCH$_3$	2,1	U08	3CH$_3$, 4CH$_3$	2,398
N05	4Cl	1,2	Y01	4H	0,903
N06	3Cl, 4Cl	0,9	Y02	4Br	1,204
N07	3Cl	0,9	Y03	4CH$_3$	0,602
N08	4N(CH$_3$)$_2$	1,8*	Y04	4Cl	0,903
S01	4H	2,097	Y05	3Cl	0,602
S02	4CH$_3$	1,792	Z01	4H	1,505
S03	4Br	1,491	Z02	4Br	1,204
S04	4OCH$_3$	1,792	Z03	4CH$_3$	1,204
S05	4Cl	1,204	Z04	4Cl	1,505
S06	3Cl, 4Cl	0,903	Z05	3Cl,4Cl	1,204

It has to be noted that the activity for the compound **S08** was not available.

Table 68. Summary of the molecular data set and the statistical protocol.

Molecular Data Set	
Compounds	Benzoxazines
Type of Compounds	3-phenyl-2H-*1,3*-benzoxazine-2,4(3H)-dithiones 3-phenyl-4-thioxo-2H-*1,3*-benzoxazine-2,4(3H)-*di*ones
Number of Compounds	46
Activity	Log (MIC), MIC: Minimum Inhibitory Concentration
Computational Details	
Molecular Modelling	WebLAb Viewer Pro modelling
Geometry Optimization	PC Spartan software package [7]
Density Function	Fitted first-order Promolecular ASA (PASA), 3-21G basis set
MQSM Operator	Overlap Operator, QS-SM applied to fragments
Molecular Alignment	Not Needed
Reduction of Dimensions	Principal Component Analysis (PCA)
Selection of Variables	Most Predictive Variables Method (MPVM)
Correlation Method	Multiple Linear Regression (MLR)
Validation	Internal Leave-One-Out Cross-Validation (LOO-CV)

Results and discussion

For the development of QSAR, the original empirical parameters used as descriptors in QSAR equations were replaced by the corresponding theoretical counterparts based on appropriate similarity and/or self-similarity measures.

$$activity = b_0 Z_{AA}^{full} + \sum_{X = fragments} b_X Z_{AA}^{X} + a$$

$$activity = \sum_{X = fragments} b_X Z_{AA}^{X} + a$$

where Z_{AA}^{full} stands for QS-SM for the whole molecule, and Z_{AA}^{X} stands for a fragment QS-SM. In order to rationalize the observed biological activity in the studied series of compounds, the broad class of all possible single, two and three-parameter multilinear QSAR models was scrutinized.

Table 69. Full set of QSAR models considered. Fr$_i$ represents the *i*-th fragment, and full stands for the similarity measure corresponding to the whole molecule.

One-parameter Model	Two-parameter model	Two-parameter Model	Three-parameter model
Full			
fr1	fr1 ; full	fr1 ; fr2	fr1 ; fr2 ; full
fr2	fr2 ; full	fr1 ; fr4	fr1 ; fr4 ; full
fr3	fr3 ; full	fr1 ; fr5	fr1 ; fr5 ; full
fr4	fr4 ; full	fr1 ; fr6	fr1 ; fr6 ; full
fr5	fr5 ; full	fr2 ; fr3	fr2 ; fr3 ; full
fr6	fr6 ; full	fr3 ; fr5	fr3 ; fr5 ; full
fr7	fr7 ; full		

The statistical importance of the aforementioned models was evaluated using the statistical *P* analytical importance criterion [149] and the related confidence level of a correlation, *CL*.

This fragment-based similarity approach can be used even in the situation where the fragment responsible for the observed activity is not known beforehand. In the present case, the systematic scrutiny of theoretical QSAR models associated with different molecular fragments was used for the detection and localization of the fragment most likely to be responsible for the observed activity.

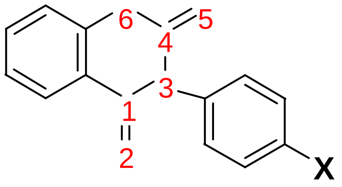
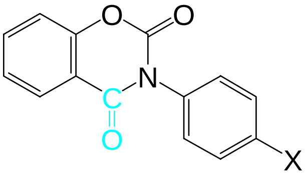
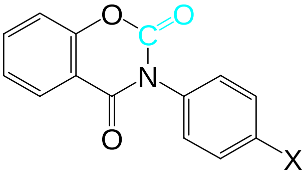
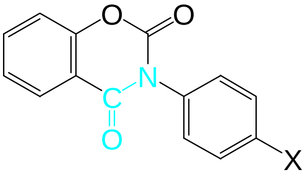
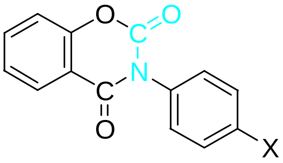
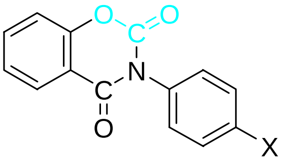
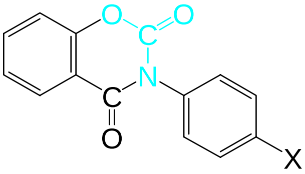
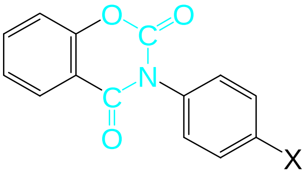


Figure 39. Numbering of atoms for the definition of molecular fragments considered as potential pharmacophores in the studied series of compounds.

Table 70. Definition of the molecular fragments. * O2, O5 and O6 can be correspondingly replaced by S2, S5 and N6.

Fragment	Fragment Number	Number of Atoms	FRAGMENT
$C^1=O^2$	1	2	
$C^4=O^5$	2	2	
$N^3-C^1=O^2$	3	3	
$N^3-C^4=O^5$	4	3	
$O^6-C^4=O^5$	5	3	
$N^3-C^4=O^5-(O^6)$	6	4	
$C^1=O^2-N^3-C^4=O^5-(O^6)$	7	6	

Results and discussion

Table 71. Selected QSAR equations and statistical significance of best QSAR models for individual sets of molecules.

Set	**N**[a]	**Fragment**[b]	**r^2**	**r^2cv**	**P**[c]	**% CL**[d]	**Slope**	**Intercept**
M	9	fr1	0,755	0,720	0,0023	99,77	-1069.578	119918.184
N	8	fr5	0,511	0,430	0,0463	95,37	-333.623	64292.314
N	8	fr1	0,510[†]	0,428	0,0467	95,33	-948.510	106344.478
S	11	fr5	0,558	0,509	0,0083	99,17	-232.364	44779.629
S	11	fr1	0,553[†]	0,504	0,0087	99,13	-732.153	82087.273
U	8	fr4	0,150	0,008	0,3438	65,62	260.272	-42728.963
Y	5	full	0,634	0,512	0,1070	89,30	0,000	0,547
Z	5	fr3	0,277	0,036	0,362	63,80	-662,325	588694,097

[a]Number of molecules in the set. [b]Label of the fragment. [c]Previously defined probability. [d]Confidence Level. [†] Second best correlation model

Table 72. Summary of best QSAR models for the data sets formed by joining several series of molecules.

Set	N[a]	Fr.[b]	r^2	r^2cv	P[c]	% CL[d]	Slope	Intercept
M/N	17	fr1	0,622	0,597	1,692E-04	99,98	-1007.783	112989.913
	14ǂ	fr1	0,531	0,491	3,131E-03	99,69	-1194.523	133926.466
M/N/S	28	fr1	0,524	0,506	1,338E-05	100,00	-684.817	76780.271
	24ǂ	fr1	0,422	0,395	5,967E-04	99,94	-651.294	73021.824
S/U	19	fr1	0,618	0,595	6,609E-05	99,99	-226.356	25379.508
	15ǂ	fr1	0,768	0,750	1,834E-05	100,00	-288.370	32332.312
U/Z	13[*]	fr1	0,795	0,776	4,300E-05	100,00	-0.0014	2.4766
	13[*]	fr5	0,795	0,776	4,300E-05	100,00	-0.0014	2.5483
	10ǂ[*]	fr1	0,943	0,935	3,067E-06	100,00	-0.0016	2.7031
	10ǂ[*]	fr5	0,943	0,935	3,067E-06	100,00	-0.0016	2.7889
Y/Z	10	fr1	0,613	0,564	7,396E-03	99,26	-77,709	65023,499
S/Z	16	fr3 fr5	0,566	0,499	4,405E-03	99,56	164,868 -171,624	6004,685
M/N/S/ U/Y/Z	39ǂ	fr1 fr5	0,593	0,569	9,900E-08	100,00	0.0292 -0.0303	4.0303

[a]Number of molecules in the set. [b]Label of the fragment. [c]Previously defined probability. [d]Confidence Level. ǂCompounds with not exactly measured activities omitted. [*]Comparable models could be obtained using practically any of the fragments.

The universality and flexibility of the similarity approach is demonstrated by the formulation of analogous theoretical QSAR models for wider data sets formed by joining several series of compounds.

From the whole set of analyzed QSAR models it can be deduced that the most statistically important correlations were in almost all cases obtained using single-parameter equations. In addition, a more careful inspection revealed that descriptors the most often repeating in successful models were always associated with the fragments *fr1* and *fr5*. This result is very interesting since these fragments just involve oxo and thioxo groups, whose role in determining the antimycobacterial activity of the studied series of molecules was the main goal of the scrutiny in this study.

For the series **S**, for which the traditional approach yielded satisfactory correlations using two-parameter correlation equations [149], comparable or even better accuracy could be obtained using theoretical QSAR models based on single-parameter equations employing fragment similarity measures associated with fragments *fr1* or *fr5*;

$$\log(\mathrm{MIC})^{-1} = a\sigma + b\pi + c$$

$$\begin{aligned}&\log(\mathrm{MIC})^{-1} = 0.826\sigma + 0.273\pi - 1.873\\ &\mathrm{r} = 0.865 \quad \mathrm{s} = 0.234 \quad \mathrm{F} = 10.38 \quad \mathrm{n} = 10\\ &\mathrm{P} = 0.008 \quad \%\mathrm{CL} = 99.20\end{aligned}$$

$$^{\mathrm{o}}\log(\mathrm{MIC})^{-1} = a\sigma + b\log P + c$$

$$\begin{aligned}&\log(\mathrm{MIC})^{-1} = 0.826\sigma + 0.313\log P - 3.084\\ &\mathrm{r} = 0.844 \quad \mathrm{s} = 0.25 \quad \mathrm{F} = 8.64 \quad \mathrm{n} = 10\\ &\mathrm{P} = 0.013 \quad \%\mathrm{CL} = 98.72\end{aligned}$$

The sets **U**, **Y** and **Z** exhibit a slightly more complex situation. For these cases, the best QSAR models were obtained using the fragment similarity measure associated to other fragments but *fr1* and *fr5*. However, the statistical importance of these correlations was very low so that it is difficult to speak of reasonable correlation. This is attributed to the relatively small number of molecules in sets **Z** and **Y** and, also, to the fact that the antimycobacterial activity of some of the studied molecules is so low that they are practically inactive [149].

Besides, empirical QSAR models were also reported for some of the extended data sets [148-146]. The models achieved in this study are comparable or even better to those previously reported. Thus, for example, the following empirical QSAR correlations were reported for the joined series **M** and **N** [146], but in contrast to the three-parameter empirical model,

$$\begin{aligned}&\log(\mathrm{MIC}) = -0.675\sigma - 0.293\pi - 0.057I + 1.680\\ &\mathrm{r} = 0.856 \quad \mathrm{s} = 0.234 \quad \mathrm{F} = 9.14 \quad \mathrm{n} = 14\\ &\mathrm{P} = 0.003 \quad \%\mathrm{CL} = 99.68\end{aligned}$$

the proposed theoretical approach leaded to a comparable statistical significance with a single-parameter equation, using the self-similarity measure associated with the fragment *fr1* as the corresponding descriptor.

$$\begin{aligned}&\log(MIC)^{-1} = \text{-}1194.523\mathrm{Z}_{\mathrm{AA}}^{\mathrm{C}^1=O^2} + 133926.466\\ &\mathrm{n} = 14 \quad \mathrm{r} = 0.728 \quad \mathrm{P} = 0.003 \quad \%\mathrm{CL} = 99.69\end{aligned}$$

Similarly, the reported QSAR model for the joint series **Y** and **Z** [148] required the use of seven indicator variables, while only a single parameter equation based on the self-similarity measures associated with the fragment *fr1* was again needed to obtain the same quality correlation using the similarity approach.

$$\begin{aligned}&\log(MIC)^{-1} = \text{-}77.709\mathrm{Z}_{\mathrm{AA}}^{\mathrm{C}^1=O^2} + 65023.499\\ &\mathrm{n} = 10 \quad \mathrm{r} = 0.783 \quad \mathrm{P} = 0.007 \quad \%\mathrm{CL} = 99.26\end{aligned}$$

Another example of the broad applicability of similarity approach concerns the joint set of molecules involving the series **S** and **U** for which a single-parameter theoretical QSAR model could again be formulated although no analogous QSAR model based on traditional approach was reported so far.

$$\begin{aligned}&\log(MIC)^{-1} = -226.356Z_{AA}^{C_1=O_2} + 25379.508\\ &\mathrm{n} = 19 \quad \mathrm{r} = 0.786 \quad \mathrm{P} = 0.000066 \quad \%\mathrm{CL} = 99.99\end{aligned}$$

This joined set is interesting because it combines the series **U** that involves the set of the practically inactive 3-phenylquinazolines with the set of active benzoxazines. The reasons for this successful correlation are clearly visible from the representation of experimental activiy versus self-similarity measures for *fr1*. While the scatter of the data within the inactive series **U** is relatively significant so that no reasonable correlation with *fr1* exists, the inclusion of the active series **S** improves the situation dramatically. This is due to the big difference in the activity of the molecules in sets **S** and **U**, which dominates the successful correlation of the combined sets **S**/**U**.

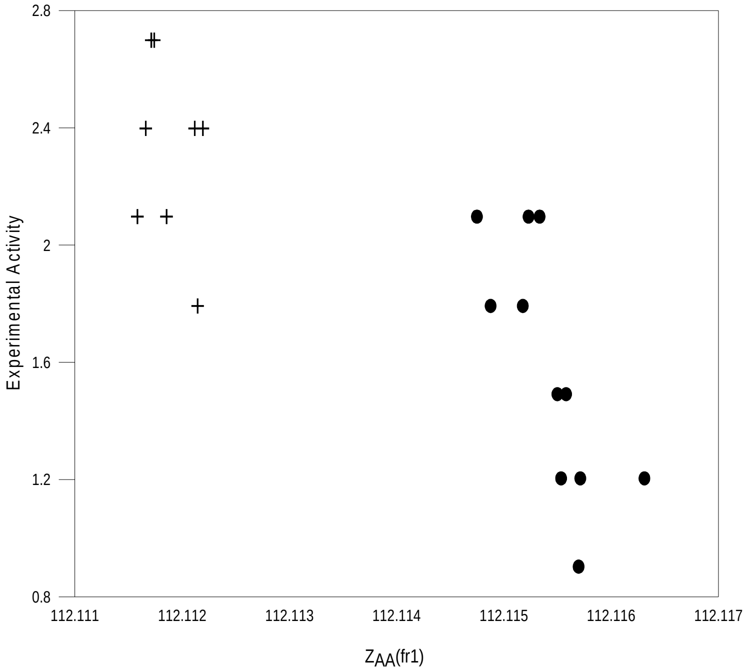


Figure 40. Experimental activity versus self-similarity measures for *fr1*, for the joint set **S/U**, where de circles symbolize the compounds of set **S**, and the crosses the ones of set **U**.

A similar advantage of the theoretical approach is also evident from the fact that comparable theoretical QSAR models can be formulated for the more extensive series of molecules involving all the studied sets, for which no analogous empirical equation has been reported using the traditional approach.

Thus, for example, the following two-parameter QSAR model was found for the extensive set of molecules involving the joint series **M, N, S, U, Y** and **Z.**

$$\log(MIC)^{-1} = 0.0292 Z_{AA}^{C_1=O_2} - 0.0303 Z_{AA}^{O_6-C_4=O_5} + 4.0303$$

$$\mathrm{n} = 39 \quad \mathrm{r} = 0.770 \quad \mathrm{P} = 0.00000009 \quad \%\mathrm{CL} = 100$$

Summarizing the above results, it can be concluded that the reported similarity approach is in complete harmony with previous experimental studies. Thus, for example, the replacement of the oxo group in the position **2** by the corresponding thioxo group was reported as the main factor responsible for the observed increase of antimycobacterial activity [148]; hence, it is interesting that it is just the fragment involving these groups (*fr1*), which was found to yield the best theoretical QSAR correlations with the experimental activity. A similar albeit weaker activating effect was also reported to accompany the replacement of the oxo group by thioxo in the position **4** and the importance of this particular group (*fr5*) is again clearly revealed by the fact that the corresponding similarity measure was detected as the second most successful molecular descriptor in the reported theoretical QSAR models. On the other hand, the replacement of *O* by *NH* in the series of benzoxazines and quinazolines is clearly accompanied by the drop of the activity which is also well reproduced by the calculated descriptors.

While traditional QSAR models were able to describe the activity only within each of the six sets of studied molecules individually, the present approach is much more general and a single universal QSAR model describing the activity of all the *39* studied molecules in all the studied series together has been successfully built. In particular, the replacement of the oxo group by the thioxo group in position **4** on the benzoxazine ring of the antitubercular 3-(phenyl)-2H-benzoxazine-2,4 (3H)-diones increases the activity, as well as the similar replacement in position **2**.

Appended contribution

► Gallegos, A.; Carbó-Dorca, R.; Ponec, R., Waisser, K. Similarity approach to QSAR. Application to antimycobacterial benzoxazines. *Int. J. Pharm., 269,* **2004**, 51-60.

REFERENCES

1. Workshop “Regulatory Acceptance of QSARs for Human Health and Environment Endpoints” Setubal, Portugal, **2002**.
2. WebLab ViewerPro version 3.2, Molecular Simulations Inc., **1998**. Free demonstration versions can be downloaded from http://www.msi.com.
3. Hyperchem 6.01, Hypercube Inc., Gainesville, FL, **1999**.
4. Dewar, M.J.S.; Zoebisch, E.G.; Healy, E.F.; Stewart, J.J.P. AM1: A New General Purpose Quantum Chemical Molecular Model. *J. Am. Chem. Soc., 107*, **1985**, 3902-3909.
5. Ampac-6.55, Semichem, 7128 Summit, Schawnee, KS 66216DA, **1998**.
6. Mopac 6.0, A General Molecular Orbitals Package: Steward, J. J. P., QCPE, **1990**.
7. PC Spartan Pro, version 1.4, Wavefunction Inc, Irvine, **1997**.
8. Amat, L.; Carbó-Dorca, R. Quantum Similarity Measures under Atomic Shell Approximation: first order density fitting using Elementary Jacobi Rotations. *J. Comput. Chem., 18,* **1997**, 2023-2039.
9. Available at http: //iqc.udg.es/cat/similarity/ASA/basisset.html [last accessed 28 May 2004].
10. Constants, P.; Amat, L; Carbó-Dorca, R. Toward a global maximization of the molecular similarity function: superposition of two molecules. *J. Comput. Chem., 18*, **1997**, 826-846.
11. Gironés, X.; Robert, D.; Carbó-Dorca, R. TGSA: a molecular superposition program based on Topo-Geometrical Considerations. *J. Comput. Chem., 22*, **2001**, 255-263.
12. Amat, L.; Constans, P.; Besalú, E.; Carbó-Dorca, R. MOLSIMIL 97, Institut de Química Computacional, Universitat de Girona, **1997**.
13. Amat, L.; Carbó-Dorca, R. MOLSIMIL 2000. Institute of Computational Chemistry, University of Girona, Girona, **2000**.
14. Gallegos, A.; Gironés, X.; Carbó-Dorca, R. TOPO. Software developed to calculate topological indices. Institute of Computational Chemistry, University of Girona, Girona, **2000**.
15. TSAR v3.2; Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Sandford-on-Thames, Oxford, OX4 4GA, UK.
16. Amat, L.; Robert, D.; Besalú, E. TQSAR-SIM. Institute of Computational Chemistry, University of Girona, Girona, **1997**.
17. Gironés, X.; Carbó-Dorca, R. TQSAR-PLS. Institute of Computational Chemistry, University of Girona, Girona, **2000**.
18. Minitab 13; Minitab Inc., **2002**.
19. Grimmer, G.; Naujac, K.W.; Dettbarn, G.; Brune, H.; Deutsch-Wenzel, R.; Misfeld, J. *Polynucleic Aromatic Hydrocarbons: Physics, Biology, and Chemistry* 6th International Symposium. Cooke, M.; Dennis, A.J.; Fisher, G.L. (Eds.) Bartelle Press: Columbus, **1981**.
20. Takada, H.; Onda, T.**;** Ogura, N. *Environ. Sci. Technol., 24*, **1990**, 1179.
21. Taylor, P.; Dellinger, B.**;** Lee, C.C. *Environ. Sci. Technol*., 24, **1990**, 316.
22. Freeman, D.J.**;** Cattell, F.C.R. *Environ. Sci. Technol*., 24, **1990**, 1581.
23. Grimmer, G. (Ed.) Environmental Carcinogens, Polycyclic Aromatic Hydrocarbons, Chemistry, Occurrence, Biochemistry, Carcinogenicity. CRC Press: Boca Raton, **1983**.
24. Pott, P. Surgical observations relative to the cancer of the scrotum. **1775**. Reprinted in Natl. Cancer Inst. Monogr., 10, 1973, 7.
25. United States Environmental Protection Agency (EPA). Integrated Risk Information System (IRIS). Environmental Criteria and Assessment Office, Office of Health and Environmental Assessment, Cincinnati OH, **1994**.
26. Agency for Toxic Substances and Disease Registry (ATSDR). Toxicological Profile for Polycyclic Aromatic Hydrocarbons. Acenaphthene, Acenaphthylene, Anthracene, Benzo(a)anthracene, Benzo(a)pyrene, Benzo(b)fluoranthene, Benzo(g,i,h)perylene, Benzo(k)fluoranthene, Chrysene, Dibenzo(a,h)anthracene, Fluoranthene, Fluorene, Indeno(1,2,3-c,d)pyrene, Phenanthrene, Pyrene. Prepared by Clement International Corporation, under Contract No. 205-88-0608. ATSDR/TP-90-20, **1990**.
27. International Agency for Research on Cancer (IARC). IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. *Polynuclear Aromatic Compounds. Part 1. Chemical, Environmental and Experimental Data*; *Vol. 32*. World Health Organization: Lyon, **1983**.
28. Govers, H.A.J. Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology. Karcher, W.**;** Devillers, J. (Eds.) Kluwer: Dordrecht, **1990**.
29. Hansch, C**;** Fujita, T. *J. Am. Chem. Soc., 86,* **1964**, 1616.

30. Nordén, U.E.**;** Wold, S. *Acta Chem. Scand., B32*, **1978**, 602.
31. Klopman, G. *J. Am. Chem. Soc., 106,* **1984,** 7315.
32. Lall, R.S. *Match., 15,* **1984**, 251.
33. Villemin, B.; Cherqaoui, D.**;** Mesbah, A. Predicting Carcinogenicity of Polycyclic Aromatic Hydrocarbons form Back-Propagation Neural Network. *J. Chem. Inf. Comput. Sci., 34,* **1994**, 1288-1293.
34. Isu, Y.; Nagashima, U.; Aoyama, T.**;** Haruo, H. Development of Neural Network Simulator for Structure-Activity Correlation of Molecules (NECO). Prediction of Endo/Exo Substitution of Norbornane Derivatives and of Carcinogenic Activity of PAHs from 13C-NMR Shifts. *J. Chem. Inf. Comput. Sci., 36,* **1996**, 286-293.
35. Vendrame, R., Braga, R.S.; Takahata, Y.**;** Galvão, D.S. Structure-Activity Relationship Studies of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons using Calculated Molecular Descriptors with Principal Component Analysis and Neural Network Methods. *J. Chem. Inf. Comput. Sci., 39,* **1999**, 1094-1104.
36. Roy, T.A.; Krueger, A.J.; Mackerer, C.R.; Neil, W.; Arroyo, A.M.; Yang, J.J. SAR models for estimating the percutaneous absorption of polynuclear aromatic hydrocarbons. *SAR QSAR Environ. Res., 9,* **1998,** 171-185.
37. Dipple, A. *Polynuclear Aromatic Carcinogens, Chemical Carcinogens*. Searle, C.E. (Ed.) American Chemical Society: Washington, **1976**.
38. Dipple, A.; Moschel, R.C.**;** Bigger, C.A.H. *Polynuclear Aromatic Carcinogens, Chemical Carcinogens*. American Chemical Society: Washington, **1984**, 2nd Ed.
39. Cavalieri, E.L.; Rogan, E.G.; Roth, R.W.; Saugier, R.K.**;** Hakan, A. The Relationship between Ionization Potential and Horseradish peroxidase/hydrogen peroxide-catalyzed binding of Aromatic Hydrocarbons to DNA. *Chem. Biol. Interact., 47,* **1983**, 87-109.
40. Iball, J. *Am. J. Cancer, 35,* **1939**, 188.
41. Badger, G.M. *Br. J. Cancer, 2,* **1948**, 309.
42. Barone, P.M.V.B.; Camilo, A., Jr**;** Galvão, D.S. Theoretical Approach to Identify Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons. *Phys. Rev. Lett., 77,* **1996**, 1186-1189.
43. Braga, R.S.; Barone, P.M.V.B.**;** Galvao, D.S. Identifying carcinogenic activity of methylated polycyclic aromatic hydrocarbons (PAHs). *J. Mol. Struct. (THEOCHEM), 464,* **1999**, 257-266.
44. Streitwieser, A. *Molecular Orbital Theory*. Wiley: New York, **1961**.
45. Vendrame, R.; Braga, R.S.; Takahata, Y.**;** Galvão, D.S.**;** Galvão, D.S. Structure-Activity Relationship Studies of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons using Calculated Molecular Descriptors with Principal Component Analysis and Neural Network Methods. *J. Chem. Inf. Comput. Sci., 39,* **1999**, 1094-1104.
46. Yang, J.J.; Roy, T.A.; Neil, W.**;** Krueger, A.J. Percutaneous and oral absorption of chlorinated paraffins in the rat. *Toxicol. Ind. Hlth., 3,* **1987**, 405-412.
47. Yang, J.J.; Roy, T.A**;** Mackerer, C.R. Percutaneous absorption of benzo(a)pyrene in the rat: comparison of in vivo and in vitro results. *Toxicol. Ind. Hlth., 2,* **1986**, 409-416.
48. Yang, J.J.; Roy, T.A**;** Mackerer, C.R. Percutaneous absorption of anthracene in the rat: comparison of in vivo and in vitro results. *Toxicol. Ind. Hlth., 3,* **1986**, 79-84.
49. Schmidt, O.Z. *Physik. Chem., B42,* **1939**, 83.
50. Pullman, A.**;** Pullman, B. *Rev. Sci., 84,* **1946**, 145.
51. Pullman, A. **;** Pullman, B. *Ad. Cancer Res.*, 3 (1955) 117.
52. Gayoso, J.**;** Kimri, S. Sur Une Tentative d'Unification des Théories Quantiques de la Cancérisation par les Polyacènes: I. Théorie des Régions M, L, et B. *Int. J. Quantum Chem., 38,* **1990**, 461-486.
53. Gayoso, J.**;** Kimri, S. *Int. J. Quantum Chem., 38,* **1990**, 487.
54. Jerina, D.M.; Lehr, R.E.; Schaefer, R.M.; Yagi, H.; Karle, J.M.; Thakker, D.R.; Wood, A.W.**;** Conney, A.H. *Origins of Human Cancer*. Hiatt, H.; Watson, J.D**.;** Winstin, I. (Eds.) Cold Spring Harbor: New York, **1977**.
55. Gute, B.D.; Grundwald, G.D.**;** Basak, S.C. Prediction of the Dermal Penetration of Polycyclic Aromatic Hydrocarbons (PAHs): A Hierarchical QSAR Approach. *SAR QSAR Environ. Res., 10,* **1999**, 1-15.
56. Kubinyi, H. (Ed.) QSAR: Hansch Analysis and Related Approaches. VCH: Weinheim, **1993**.
57. Stuper, A.J.; Brugger, W.E.**;** Jurs, P.C. Computer Assisted Studies of Chemical Structure and Biological Function. Wiley: New York, **1979**.
58. Potts, R.O.**;** Guy, R.H. *Pharm. Res., 9,* **1992**, 663.
59. Benigni, R.; Andreoli, C.**;** Giuliani, A. *Environ. Mol. Mutagen., 24,* **1994**, 208.

60. Levy, M.R. *Biology of Tetrahymena* Elliott, A.M. (Ed.) Dowden, Hutchison and Ross: Stroudsburg, **1973,** 227-258.
61. Holz, G.G. Jr. *Biology of Tetrahymena* Elliott, A.M. (Ed.) Dowden, Hutchison and Ross: Stroudsburg, **1973,** 89-98.
62. Cameron, I.L. *Biology of Tetrahymena* Elliott, A.M. (Ed.) Dowden, Hutchison and Ross: Stroudsburg, **1973,** 199-226.
63. Levy, M.R. *Biology of Tetrahymena* Elliott, A.M. (Ed.) Dowden, Hutchison and Ross: Stroudsburg, **1973,** 227-258.
64. Schultz, T.W.; Lin, D.T.; Wilke, T.S.**;** Arnold, M. Quantitative structure-activity relationships for the Tertahymena Pyriformis population growth endpoint: a mechanisms of action approach. *Practical Applications of quantitative structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*. Karcher, W.; Devillers, J. (Eds.) ECSC, EECº, EAEC: Brussels and Luxembourg, **1990,** 241-262.
65. TDR News (News from the WHO Division of Control of Tropical Diseases), *46*, **1994**, 5.
66. The Malaria Foundation International website, at http://www.malaria.org.
67. Bruce-Chwatt, L.J. (Ed.) *Chemotherapy of Malaria*. W.H.O.: Geneva, **1981**.
68. Peters, W. Chemotherapy and Drug Resistance in Malaria. Academic Press: London, **1987**.
69. Oaks, S.C.; Mitchell, V.S.; Pearson, G.W.; Carpenter, C.C. *MALARIA Obstacles and Opportunities*. National Academic Press: Washington, **1991.**
70. Oduola, A.M.J.; Omitowoju, G.O.; Gerena, L.: Kyle, D.E.; Milhous, M.K.; Sowunmi, A.; Salako, L.A. Reversal of Mefloquine Resistance with Penfluridol in Isolates of Plasmodiu Falciparum from South-West Nigeria. *Trans. R. Soc. Trop. Med. Hyg*., *87*, **1993**, 81-83.
71. Klayman, D.L. Qinghaosu (Artemisinin): an antimalaric drug from China. *Science, 228,* **1985**, 1049-1055.
72. Luo, X.-D.; Shen, C.-C. The chemistry, pharmacology, and clinical applications of qinghaosu (artemisinin) and its derivatives. *Med. Res. Rev., 7,* **1987**, 29-52.
73. Klayman, D.L. *Science*, *228*, **1985**, 1049.
74. Butler, A.R.**;** Wu, Y.L. *Chem. Soc. Rev., 21,* **1992,** 85.
75. Shen, C.C.**;** Zhuang, L.G. *Med. Res. Rev*., 4, **1984**, 47.
76. Zhou, W.-S.; Xu, X.-X. Total Synthesis of the Antimalarial Sesquiterpene Peroxide Qinghaosu and Yingzhaosu A. *Acc. Chem. Res*., *27,* **1994**, 211-216.
77. Zaman, S.S.; Sharma, R.P. Some Aspects of the Chemistry and Biological Activity of Artemisinin and Related Antimalarials. *Heterocycles, 32,* **1991,** 1593-1638.
78. Jung, M. Current Developments in the Chemistry of Artemisinin and Related Compounds. *Curr. Med. Chem*., *1,* **1994**, 35-49.
79. Avery, M.A.; Gao, F.; Mehrotra, S.; Chong, W.K.; Jennings-White, C. *The Organic and Medicinal Chemistry of Artemisinin and Analogs*. Research Trends Trivandrum: India, **1993**.
80. Posner, G.H.; Ploypradith, P.; Parker, M.H.; O'Dowd, H.; Woo, S.-H.; Northrop, J.; Krasavin, M.; Dolan, P.; Kensler, T.W.; Xie, S.; Shapiro, T.A. Antimalarial, Antiproliferative, and Antitumor Activities of Artemisinin-Derived, Chemically Robust, Trioxane Dimers. *J. Med. Chem*., *42,* **1999**, 4275-4280.
81. Grigorov, M.; Weber, J.; Tronchet, J.M.J.; Jefford, C.W.; Milhous, W.K.; Maric D. A QSAR Study of the Antimalarial Activity of Some Synthetic 1,2,4-Trioxanes. *J. Chem. Inf. Comput. Sci.*, *37,* **1997**, 124-130.
82. Jefford, C.W.; Velarde, J.A.; Bernadinelli, G.; Bray, D.H.; Warhust, D.C.**;** Milhous, W.K. *Helv. Chim. Acta, 76,* **1993,** 2775.
83. Jefford, C.W.; Wang, Y.; Bernadinelli, G. *Helv. Chim. Acta, 71,* **1988**, 2042.
84. Testa, B.; Kyburz, E.; Fuhrer, W.**;** Giger, R. (Eds.) *Perspectives in Medicinal Chemistry; Vol. 25*. VCH Publishers: Amsterdam, **1992**, 459-472.
85. Posner, G.H.; O´Dowd, H.; Ploypradith, P.; Cumming J.N.; Xie, S.**;** Shapiro, T.A. Antimalarial Cyclic Peroxy Ketals. *J. Med. Chem*., *41,* **1998**, 2164-2167.
86. Posner, G.H.; Parker, M.H.; Northrop, J.; Elias, J.S.; Ploypradith, P.; Xie, S.; Shapiro, T.A. Orally Active, Hydrolytically Stable, Semisynthetic, Antimalarial Trioxanes in the Artemisinin Family. *J. Med. Chem*., *42,* **1999**, 300-304.
87. Posner, G.H.; González, L.; Cumming, J.N.; Klinedinst, D.; Shapiro, T.A. Synthesis and Antimalarial Activity of Heteroatom-Containing Bicyclic Endoperoxides. *Tetrahedron, 53,* **1997**, 37-50.
88. Avery, M.A.; Mehrotra, S.; Bonk, J.D.; Vroman J.A.; Goins, D.K.; Miller, R. Structure-Activity Relationships of the Antimalarial Agent Artemisinin. 4. Effect of Substitution at C-3. *J. Med. Chem*., *39,* **1996**, 2900-2906.

89. Desjardins, R.E.; Canfield, C.J.; Haynes, D.E.; Chulay, J.D. Quantitative Assessment of Antimalarial Activity In Vitro by a Semiautomated Microdilution Technique. *Antimicrob. Agents Chemother.*, *16*, **1979,** 710-718.
90. Milhous, W.K.; Weatherly, N.F.; Bowdre, J.H.; Desjardins, R.E. In Vitro Activities of and Mechanisms of Resistance to Antifolate Antimalarial Drugs. *Antimicrob. Agents Chemother.*, **27, 1985**, 525-530.
91. Frisch, M.J., Trucks, G. W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Zakrzewski, V.G., Montgomery, Jr., J.A., Stratmann, R.E., Burant, J.C., Dapprich, S., Millam, J.M., Daniels, D., Kudin, K.N., Strain, M.C., Farkas, O., Tomasi, J., Barone, V., Cossi, M., Cammi, R., Mennucci, B., Pomelli, C., Adamo, C., Clifford, S., Ochterski, J., Petersson, G.A., Ayala, P.Y., Cui, Q., Morokuma, K., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Cioslowski, J., Ortiz, J.V., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P., Komaromi, I., Gomperts, R., Martin, R.L., Fox, D.J., Keith, M. T., Al-Laham, A., Peng, C.Y., Nanayakkara, A., Gonzalez, C., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Andres, J.L., Gonzalez, C., Head-Gordon, M., Replogle, E.S., and Pople, J.A. Gaussian 98, Revision A.6, Gaussian, Inc., Pittsburgh PA, **1998**.
92. Besalú, E. MLR-TBI. Software developed to calculate topological indices. Institute of Computational Chemistry, University of Girona, Girona, **1998**.
93. Avery, M.A.; Mehrotra, S.; Johnson, T.L.; Bonk, J.D.; Vroman, J.A.; Millar, R. Structure-Activity Relationships of the Antimalarial Agent Artemisinin. 5. Analogs of 10-Deoxoartemisinin Substituted at C-3 and C-9. *J. Med. Chem*., *39*, **1996**, 4149-4155.
94. Tsar, version 2.41, Oxford Molecular Group, Inc. CAChe Scientific: Beaverton.
95. Alzeer, J.; Chollet, J.; Heinze-Krauss, I.; Hubschwerlen, C.; Matile, H.; Ridley, R.G. Phenyl β-Methoxyacrylates: A New Antimalarial Pharmacophore. *J. Med. Chem*., *43*, **2000**, 560-568.
96. Cowden, W.B.; Halladay, P.K.; Cunningham, R.B.; Hunt, N.H.; Clark, I.A. Flavins as Potential Antimalarials. 2. 3-Methyl-10-(substituted-phenyl)flavins. *J. Med. Chem*., *34*, **1991,** 1818-1822.
97. Domínguez, J.N.; López, S.; Charris, J.; Iarruso, L.; Lobo, G.; Semenov, A.; Olson, J.E.; Rosenthal, P.J. Synthesis and Antimalarial Effects of Phenothiazine Inhibitors of a Plasmodium falciparum Cysteine Protease. *J. Med. Chem*., *40*, **1997**, 2726-2732.
98. Kavlock, R. J.; Daston, G. P.; DeRosa, C.; Fenner-Crisp, P.; Gray, L. E.; Kaattari, S. Lucier, G.; Luster, M.; Mac, M. J.; Maczka, C.; Miller, R.; Moore, J.; Rolland, R.; Scott, G.; Sheehan, D. M.; Sinks, T.; Tilson, H. A. Research needed for the risk assessment of health and environmental effects of endocrine disruptors, a report of the U. S. EPA-sponsored workshop. *Environ. Health Perspect., 104,* **1996**, 715-740.
99. U.S. Congress. The Food Quality Protection Act (FQPA) and the Safe Drinking Water Sct. (SDWA), **1996**.
100. Lai, K.M.; Scrimshaw, M.D.; Lester, J.N. The effects of natural and synthetic steroid estrogens in relation to their environmental occurrence. *Critical Reviews in toxicology, 32,* **2002**, 113-132.
101. Kelce, W.R.; Stone, C.R.; Laws, S.C.; Gray, L.E.; Kemppainen, J.A.**;** Wilson, E.W. Persistent DDT metabolite p, p’-DDE is a potent androgen receptor antagonist. *Nature, 375,* **1995**, 581-585.
102. Colburn, T.; Dumanoski, D.; Myers, J. P. *Our Stolen Future*. Pluma: New York, **1996**.
103. EDSTAC http://www.epa.gov/opptintr/opptendo/finalrpt.htm
104. Bradbury, S.; Menkenyan, O.A. Quantitative structure-activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity. An assessment of conformer flexibility. *Environ. Toxicol. Chem., 15,* **1996**, 1945-1954.
105. Tong, W.D.; Perkins, R.; Strelitz, R.; Collantes, E.R.; Keenan, S.; Welsh, W.J.; Branham, W.S.; Sheehan, D.M. Quantitative structure-activity relationships (QSARs) for estrogen binding to the estrogen receptor, predictions across species. *Environ. Health Perspect., 105,* **1997**, 1116-1124.
106. Tong, W.D.; Lowis, D.R.; Perkins, R.; Chen, Y.; Welsh, W.J.; Goddette, D.W.; Heritage, T.W.; Sheehan, D.M. Evaluation of Quantitative structure-activity relationships methods for large-escale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci., 38,* **1998**, 669-677.
107. Dodds, E.C.; Goldberg, L.; Lawson, W.; Robinson, R. Oestrogenic activity of alkylated stilboestrols. *Nature, 142,* **1938**, 34.
108. Dodge, J.A. Natural and anthropogenic environmental oestrogens, the scientific basis for risk assessment. Structure/activity relationships. *Pure & Appl. Chem., 70,* **1998**, 1725-1733.
109. Colburn, T. Environmental estrogens, health implications for humans and wildlife. *Environ. Health Perspect., 103,* **1995**, 135-136.

110. Sabih, J. Natural and anthropogenic environmental oestrogens, the scientific basis for risk assessment. Issues associated with the validation of in vitro and in vivo methods for assessing endocrine disrupting chemicals. *Pure & Appl. Chem., 70,* **1998**, 1735-1745.
111. Shi, L.M.; Fang, H.; Tong, W.D.; Wu, J.; Perkins, R.; Blair, R.M.; Branham, W.S.; Dial, S.L.; Moland, C.I.; Sheehan, D.M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci., 41,* **2001**, 186-195.
112. Nishihara, T.; Nishikawa, J.; Kanayama, T.; Dakeyama, F.; Saito, K.; Imagawa, M.; Takatori, S.; Kitagawa, Y.; Hori, S.; Utsumi, H. Estrogenic activities of 517 chemicals by yeast two-hybrid assay. *J. Health. Sci., 46,* **2000**, 282-298.
113. Cramer III, R.D.; Patterson, D.E.; Bunce, J.D. Comparative Molecular Field Analysis (CoMFA). 1. Effect on Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc., 110,* **1988**, 5959-5967.
114. Gantchev, T.G.; Ali, H.; van Lier, J.E. Quantitative structure-activity relationships/comparative molecular field analysis (QSAR/CoMFA) for receptor binding properties of halogenated estradiol derivatives. *J. Med. Chem., 37,* **1994**, 4164-4176.
115. Mekenyan, O.; Nikolova, N.; Karabunarliev, S.; Bradbury, S.P.; Ankley, G.T.; Hansen, B. New developments in hazard identification algorithm for hormone receptor ligands. *Quant. Struct.-Act. Relat., 18,* **1999**, 139-153.
116. Bradbury, S.; Kamenska, V.; Schmieder, P.; Ankley, G.; Mekenyan, O. A Computationally Based Identification Algorithm for Estrogen Receptor Ligands, Part 1. Predicting hERα Binding Affinity. *Toxicol. Sci., 58***2000**, 253-269.
117. Fang, H.; Tong, W.; Shi, L. M.; Blair, R.; Perkins, R.; Branham, W.; Hass, BS.; Xie, Q.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol., 14,* **2001**, 280-294.
118. Shi, L.; Tong, W.; Fang, H.; Xie.; Q.; Hong.; R.; Perkins, R.; Wu, J.; Tu, M.; Blair, R.M.; Branham, W.S.; Waller, C.; Walker, J.; Sheehan, D.M. An integrated "4-phase" approach for setting endocrine disruption screening priorities_phase I and II predictions of estrogen receptor binding affinity. *SAR QSAR Environ. Res., 13,* **2002**, 69-88.
119. Schultz, T.W., Sinks, G.D.; Cronin, M.T.D. Structure-Activity Relationships for Gene Activation Oestrogenicity. Evaluation of a Diverse Set of Aromatic Chemicals. *Environ. Toxicol., 17,* **2002**, 14-23.
120. Shi, L.M.; Fang, H.; Tong, W.D.; Wu, J.; Perkins, R.; Blair, R.M.; Branham, W.S.; Dial, S.L.; Moland, C.I.; Sheehan, D.M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci., 41,* **2001**, 186-195.
121. Routledge, E.J.; Sumpter, J.P. Estrogenic activity of surfactants and some of their degradation products assessed using a recombinant yeast. *Environ. Toxicol. Chem., 15,* **1996**, 241-248.
122. Schultz, T.W.; Sinks, G.D.; Cronin, M.T.D. Effect of substituent size and dimensionality on potency of phenolic xenoestrogens evaluated with a recombinant yeast assay. *Environ. Toxicol. Chem., 19,* **2000**, 2637-2642.
123. Gao, H; Williams, C.; Labute, P; Bajorath, J. Binary Quantitative-Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands. *J. Chem. Inf. Comput. Sci., 39,* **1999**, 164-168.
124. KOWWIN v1.66; Syracuse Research Corporation, **2000**.
125. Gao, H.; Katzenellenbongen, J.A.; Garg, R.; Hansch, C. Comparative QSAR Analysis of Estrogen Receptor Ligands. *Chem. Rev., 99,* **1999**, 723-744.
126. Klappa, P.; Freedman, R.B.; Langenbuch, M.; Lan, M.S.; Robinson, G.K.; Ruddock, L.W. The pancreas-specific protein disulphide-isomerase PDIp interacts with a hydroxyaryl group in ligands. *Biochem. J., 354,* **2001**, 553–559.
127. Bridgette M.; Collins, J.A.; McLachlan, S.F.A. The estrogenic and antiestrogenic activities of phytochemicals with the human estrogen receptor expressed in yeast. *Steroids, 62,* **1997**, 365-372.
128. Zand, R.S.; Jenkins, D.J.; Diamandis, E.P. Steroid hormone activity of flavonoids and related compounds. *Breast Cancer Res. Treat., 62,* **2000**, 35-49.
129. Bermudez, L.E., Young, L.S. New drug for therapy of mycobacterial infection. *Cur. Opin. Infec. Dis., 8,* **1995,** 428-437.
130. Dailloux M.; Laurain C.; Weber R.; Hartemann P. *Water Res., 33,* **1999**, 2219.
131. Slosarek, M.; Horova, B.; Rozsypal, H.; Stankova, M.; Bruckova, M. *Klin. Mikrobiol. Infekc. Lek., 9,* **1997**, 241.
132. Rouhi, A.M. Tuberculosis. A tough adversary. *Chem. Ind. News, 77,* **1999**, 52-69.
133. Martin, G.; Lazarus, A. *Postgrad. Med., 108,* **2000**, 42.

134. Schatz, A.; Waksman, A. Effect of streptomycin upon Mycobacterium tuberculosis and related organisms. *Proc. Soc. Exptl. Biol. & Med*., *67*, **1944,** 244-248

135. Schatz, A.; Bugle, E.; Waksman, S.A. Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. *Proc.Soc. Exptl. Biol. & Med., 55,* **1944,** 66-69

136. Riley L. W. *Clin. Infect. Dis., S442,* **1993**, 17.

137. Klimesova, V.; Palat, K.; Waisser, K.; Klimes, J. Combination of molecular modeling and quantitative structure-activity relationship analysis in the study of antimycobacterial activity of pyridine derivatives. *Int J Pharm., 207,* **2000**, 1-6.

138. Ragno, R.; Marshall, GR.; Di, Santo, R.; Costi, R.; Massa, S.; Rompei, R.; Artico, M. Antimycobacterial pyrroles: synthesis, anti-Mycobacterium tuberculosis activity and QSAR studies. *Bioorg Med Chem*., *8,* **2000**, 1423-32.

139. Rathbone, D.L.; Tims, K.J.; Attkins, N.; Cann, S.W.**;** Billington, D.C. QSAR Studies On A Large Set Of Antimycobacterial N1-Benzylidene-heteroarylcarboxamidrazones. *J. Pharm. Pharmacol, 52*, **2000**, 97.

140. Rugutt, J.K.; Rugutt, K.J. Relationships between molecular properties and antimycobacterial activities of steroids. *Nat Prod Lett., 16,* **2002**, 107-13.

141. Kubicova, L.; Waisser, K. *Cesk. Farm*., *41* **1992**, 208.

142. Waisser, K.; Kubicova, L. *Cesk. Farm*., *42,* **1993**, 218.

143. Waisser, K.; Kubicova L.; Dostal H. *Folia Pharm. Univ. Carol*., *23,* **1998**, 59.

144. Waisser, K.; Gregor J.; Kubicova L.; Klimesova V.; Kunes J.; Machacek M.; Kaustova J. *Eur. J. Med. Chem*., *35,* **2000**, 733.

145. Waisser, K.; Hladůvková, J.; Holý, P.; Macháček, M.; Karajannis, P.; Kubicová, L.; Klimešová, V.; Kuneš, J.; Kaustová, J. 2H-1,3-Benzoxazine-2,4(3H)-diones substituted in position 6 as antimycobacterial agents. *Chem. Pap*., *55*, **2001,** 323-334.

146. Waisser, K.; Hladůvkova, J.; Gregor, J.; Rada, T.; Kubicová, L.; Klimešová, V.; Kaustová, J. Relationships between the chemical structure of antimycobacterial substances and their activity against atypical strains. Part 14: 3-Aryl-6,8-dihalogeno-2H-1,3-benzoxazine-2,4(3H)-diones. *J. Arch. Pharm. Pharm. Med. Chem., 331*, **1998,** 3-6.

147. Waisser, K.; Macháček, M.; Dostál, H.; Gregor, J.; Kubicová, L.; Klimešová, V.; Kuneš, K.; Palát, K.; Hladůvkova, J.; Kaustová, J.; Möllmann, U. Relationships between the chemical structure of substances and their antimycobacterial activity against atypical strains. Part 18. 3-Phenyl-2H-1,3-benzoxazine-2,4(3H)-diones and isosteric 3-phenylquinazoline-2,4(1H, 3H)-diones. *Collect. Czech. Chem. Commun., 64*, **1999,** 1902-1924.

148. Waisser, K.; Gregor, J.; Dostál, H.; Kuneš, J.; Kubicová, L.; Klimešová, V.; Kaustová, J. Influence of the replacement of the oxo function with the thioxo group on the antimycobacterial activity of 3-aryl-6,8-dichloro-2H-1,3-benzoxazine-2,4(3H)-diones and 3-arylquinazoline-2,4(1H,3H)-diones. *Il Farmaco, 56*, **2001,** 803-807.

149. Pecka, J.; Ponec, R. Simple Analytical Method for Evaluation of Statistical Importance of Correlations in QSAR Studies. *J. Math. Chem., 27*, **2000,** 13-22.

Conclusions

Εν οίδα ότι ουδέν οίδα

Scio me nihil scire

Socrates

1 FINAL CONCLUSIONS

This memory deals with several aspects related to the calculation and implementation of Quantum Similarity Measures (QSM). The application of Quantum Similarity Indices (QSI) as molecular descriptors in the prediction of the functionality of chemical compounds illustrates the theoretical background for the relationships between molecular structure and a physicochemical property (Quantitative Structure-Property Relationships, QSPR), biological activity (Quantitative Structure-Activity Relationships, QSAR) or toxicity (Quantitative Toxicity-Property Relationships, QSTR). The use of Quantum Similarity Theory (QST) in the development of QSAR for application in Computer-Assisted Molecular Design (CAMD) or, more specifically, in Computer-Assisted Drug Design (CADD), leads to several general conclusions:

I) The application of Molecular Quantum Similarity (MQS) to QSAR yields not only acceptable but also satisfactory results in the construction of diverse structure-function correlation studies, such as QSPR, QSAR and QSTR. In general, structure-function correlation studies can reveal the environmental, chemical, biological, toxicological, or pharmacological information embedded in molecular structures. Thus, the application of Quantum Similarity (QS) in the QSAR framework allows the satisfactory description of functions, i.e. physical properties, biological activities, or molecular toxicities, associated to molecular sets. This approach has been implemented in different fields such as medicinal chemistry, environmental chemistry, and protein engineering, among others. Hence, QSAR provide valuable information about the biological behaviour of potential drugs, thus establishing a pattern for CADD, which may aid in rational-based medical research. Thus, QSAR enable the resolution of several problems such as the development of new therapies and the design of novel drugs to fight against drug-resistant infectious diseases like malaria and tuberculosis, or degenerative mutagenesis, like carcinogenicity and estrogenicity. This accurate methodology, developed with adequate mathematical and computational tools, leads to a faster, cheaper and more comprehensive synthesis of new products, avoiding or at least reducing experimental synthesis and testing with animals. In comparison with the high cost of in vitro and in vivo assays, such rational-based design strategies carry a low cost, which may aid to effectively reduce the cost of launching new products into the market.

II) The applicability of QS for the generation of descriptors in QSAR correlation studies yields valuable results for molecular systems of diverse interest using relatively simple statistical methodologies. The use of Molecular Quantum Similarity Measures (MQSM) in QSAR has several characteristic features of special interest for application in CAMD:

IIA) The substitution or at least complementation of traditional empirical parameters by theoretical descriptors based in Quantum Mechanical (QM) calculations provides a proved alternative way to develop QSAR models with good predictive capacities. Besides, similarity-based descriptors constitute not only a reliable source to derive QSAR but also a valuable complement to other descriptors and even other procedures. QSI, easily derived from well-defined QM principles, constitute a source of general consistent unbiased and homogeneous theoretical descriptors, useful to establish sound and reliable QSAR analyses. In summary, the application of MQS to QSAR provides unbiased Quantum Object (QO) descriptors, alternative to classical empirical molecular parameters

IIB) The building of QSAR models within the molecular similarity frame can be considered as limited and self-contained, provided that the calculation choices are constrained to a few variables such as the selection of the similarity operator, the normalization or scaling of descriptors, and the statistical methodology

IIC) The low computational cost results in affordable computational resources and time requirements

III) QSM constitute a natural source of chemical structure description for quantum objects, namely atoms and molecules. In particular, MQSM provide an accurate and complete degree of description of the information encoded in molecular structures. In the QSAR domain, several approximations based on QS can describe different representations of molecular structure:

IIIA) The first description considers the global electronic density of objects, including the spatial disposition of the compared molecules, that is, the three-dimensional structure in the space. This approach, which may require requires an alignment process to compare molecular structures, employs the whole Density Function (DF) as a source to generate MQSM, i.e. quantum chemical descriptors that parameterise molecular speciation and reactivity.

IIIB) A second approach, derived from the first one, is founded on the partition of the global density of molecules in fragments. For such purpose, Fragment Quantum Self-Similarity Measures (QS-SM) are used as descriptors to analyse the basic structural requirements for a given activity. QS-SM defined for fragments not only can model electronic properties due to the effect of substituents, but they can also identify a specific activity or property with the corresponding functional group or fragment located in the particular molecular region that exhibits the active or reactive chemical process. The detection of common molecular regions responsible for a biological response allows the obtaining of a pattern with the active regions, of high interest for drug design purposes. In special, fragment descriptors can be derived from the substitution of a common structure template in a set of congeneric compounds.

IIIC) Finally, the simple two-dimensional representation of structures by molecular graphs can be combined with the quantum similarity theory background. The inclusion of quantum mechanical parameters in the classical topological approach constitutes an alternative method for the calculation of molecular descriptors. The application of QST together with classical graph theory yields the obtaining of the so-called Topological Quantum Similarity Indices (TQSI). Classical topological matrices, based in distances and connectivity, have been substituted by topological quantum similarity matrices, which also take into account three-dimensional information. The satisfactory results obtained with this new procedure reveal the connection between molecular topology and the general theory of quantum similarity.

IV) The use of several approximations in the methodology allows the efficient application of QSM. The reduction of the computational cost permits the application of QSM to a high number of molecules or to large molecular systems of biological interest. To mention some approaches,

IVA) The first step to save computational costs can be a comparative analysis between different geometry optimization methods to choose the best compromise between accuracy and cost for the obtained descriptors

IVB) Fitted first-order electronic Density Functions (DF) for atoms can be computed with the Atomic Shell Approximation (ASA), at different levels of theory depending on the selection of the adjusted basis function set. Indeed, the density function for molecules can be accurately described by the use of the Promolecular ASA (PASA), implemented as an extension of ASA. The formulation of PASA density simply consists in the addition of the individual ASA density contributions for the atoms that conform the molecule. Despite its simplicity, the obtained QSM do not significantly differ from those derived from *ab initio* DF, but the amount of time and computational effort required to calculate the density are effectively reduced

IVC) The molecular superposition process is crucial in some molecular similarity studies founded in three-dimensional descriptors. The molecular pairwise alignment to compare two compounds can be rapidly performed using the Topo-Geometrical Superposition Algorithm (TGSA), as an alternative to the costly maximization of the molecular similarity integral that requires the repeatedly computation and optimization of QSM. TGSA, based in geometrical and topological concepts, compares the maximum common substructure of the compared molecules. One of the main advantages of this method consists of its comprehensive simplicity, and the coherence with chemical intuition. In addition, the absence of any kind of quantum mechanical or semi-empirical calculations results in a fast algorithm able to perform large amounts of alignments within reasonable time limits

V) A general protocol for the generation of predictive QSAR models based on the application of MQSM in QSAR has been presented. The protocol comprises molecular modelling, generation of descriptors, and statistical correlation and validation techniques. In particular, several different applications have been envisaged:

VA) Quantitative studies for the derivation of numerical QSAR correlations, expressed by mathematical equations. For such purposes, simple linear QSAR models have been developed by means of Multiple Linear Regression (MLR) techniques. The elaboration of simple linear models combined with simple selected statistical methods allows a more direct interpretation of results

VB) Semiquantitative studies for the qualitative classification of compounds into categorical classes using Linear Discriminant Analysis (LDA) techniques. The biological activity of molecular series can be estimated in a discrete manner, thus establishing a pattern for classification

VC) Besides, rigorous statistical validation techniques have been adopted for the QSAR models in order to assess that no chance correlations or over-parameterised models have been obtained. In addition to the internal Cross-Validation (CV) and the computation of random tests, real predictions have been attempted by using external test sets, when further data was available

VI) Different molecular activities have been tested for correlation. In most application examples, satisfactory correlations were obtained using a relative small number of molecular descriptors, and affordable computational requirements. Even if other methodologies might provide better results, it must be emphasized that the procedure used along this work, as well as the generation of molecular descriptors, has not been manipulated. Hence, the exposed QSAR protocol consists of a methodological pathway made of unbiased and universal MQSM descriptors able to characterize different molecular activities without introducing further information than those provided by quantum similarity based on electronic density functions, although additional refinements or statistical tools may be applied to the procedure in order to improve the results according to each molecular set under study. In summary, the application of MQSM to QSAR provides comparable or even better results to other highly predictive and widely established QSAR approaches. Additionally, it has been shown that most of the information characterized by the original descriptors is contained in molecular DF, the basis of MQSM, and can be extracted from MQSM using simple linear transformations. Thus, it can be concluded that, by combining internal and external validation techniques, real predictive QSAR models have been achieved.

2 FUTURE PERSPECTIVES

As future objectives, several options have been envisaged. In complex biological systems, such as receptor-ligand systems, although in some cases experimental data from molecular geometry may be available, in other cases they are not. Hence, the minimum energy conformation of the isolated molecule is assumed to elicit the biological activity. However, it must be remarked that this approximation might lead to suspicious results.

In the case of non-rigid molecules with rotatable bonds, the same assumption holds. Thus, the inclusion of flexibility in molecular alignment, allowing the rotation of torsion bonds and slight variations in angle and bond distances, has been already considered and implemented. Related to the urgent need to register the effect of bioactive conformations, the idea of developing a new Chiral MQSM that takes into account the molecular stereochemistry of biological systems has been projected.

Besides, the further deepening into the interpretation of QSAR models based in similarity has been always a desired aim in order to provide a stronger reliable structure-based insight into the knowledge of chemical problems.

3 LIST OF PUBLICATIONS

Authors **Gallegos, A**; Robert, D; Gironés, X.; Carbó-Dorca, R.

Title Structure-Toxicity Relationships of Polycyclic Aromatic Hydrocarbons using Molecular Quantum Similarity

Reference *J. Comput.-Aid. Mol. Des., 15(1),* **2001**, 67-80.

Abstract The establishment of quantitative structure-activity relationship (QSAR) models for the toxicity of polycylic aromatic hydrocarbons (PAHs) is described. Two properties: in vitro percutaneous absorption in rat skin and discrete levels of carcinogenic activity are examined using molecular quantum similarity measures (MQSM). The results show that MQSM produces comparable, or even better, results than other approaches using physicochemical, topological and quantum-chemical molecular descriptors. Furthermore, a careful analysis puts into evidence that most of the information characterized by the original descriptors is in fact contained in the molecular density functions, the basis of MQSM. The present paper, together with several other ones reported by our laboratory, prove that MQSM might be appropriate theoretical tools for QSAR and computer-aided drug design, comparable to other highly predictive QSAR methodologies.

Authors Gironés, X.; **Gallegos, A.**; Carbó-Dorca, R.

Title Modeling Antimalarial Activity: Application of Kinetic Energy Density Quantum Similarity Measures as Descriptors in QSAR

Reference *J. Chem. Inf. Comput. Sci., 40,* **2000**, 1400-1407.

Abstract In this work, the application of the recently described Kinetic Energy Density Function in the evaluation of the antimalarial activity is demonstrated. First, this new type of Density Function is briefly presented from its theoretical foundations, and its inclusion in the Molecular Quantum Similarity framework is discussed. The application of Kinetic Energy-based Quantum Similarity Measures to QSAR is presented with 2 molecular sets composed by artemisinin derivatives, in which the 50% inhibition of synthesis and reduction of hidrofolate (IC50) in different Plasmodium Falciparum clones are analyzed. Satisfactory correlations are obtained for all antimalarial activities in all molecular sets. Molecular Quantum Similarity analysis provides a consistent, unbiased and homogeneous set of molecular descriptors and is a feasible alternative to the use of physicochemical descriptors.

Authors Gironés, X.; **Gallegos, A.**; Carbó-Dorca, R.

Title Antimalarial Activity of Synthetic 1,2,4-Trioxanes and Cyclic Peroxy Ketals, a Quantum Similarity Study

Reference *J. Comput.-Aid. Mol. Des., 15(12),* **2001**, 1053-1063.

Abstract In this work, the antimalarial activity of two series of synthetic 1, 2, 4-trioxanes and a set of cyclic peroxy ketals are tested for correlation search by means of Molecular Quantum Similarity Measures (MQSM). QSAR models, dealing with different biological responses (IC50, IC90 and ED90) of the parasite Plasmodium Falciparum, are constructed using MQSM as molecular descriptors and are satisfactorily correlated.

Authors Besalú, E.; **Gallegos, A.**; Carbó-Dorca, R.

Title Topological Quantum Similarity Indices and Their Use in QSAR: Application to Several Families of Antimalarial Compounds

Reference *MATCH-Commun. Math. CO, 44*, **2001**, 41-64.

Abstract New 3D molecular topological indices are described. From the underlying theoretical foundation, it is revealed the connection between molecular topology and the general theory of Quantum Similarity. Finally, results concerning the establishment of QSAR models, related to five antimalarial molecular families, are presented.

Authors **Gallegos Saliner, A.**; Amat, L.; Carbó-Dorca, R.; Schultz, T.W.; Cronin, M.T.D.

Title Molecular Quantum Similarity Analysis of Estrogenic Activity

Reference *J. Chem. Inf. Comput. Sci., 43*, **2003**, 1166-1176.

Abstract The main objective of this study was to evaluate the capability of 120 aromatic chemicals to bind to the human alpha estrogen receptor (hER?) by the use of quantum similarity methods. The experimental data were segregated into two categories, i.e. those compounds with and without estrogenicity activity (active and inactive). To identify potential ligands, semi-quantitative structure-activity relationships were developed for the complete set correlating the presence or lack of binding affinity to the estrogen receptor with structural features of the molecules. The structure-activity relationships were based upon molecular similarity indices, which implicitly contain information related to changes in the electron distributions of the molecules, along with indicator variables, accounting for several structural features. In addition, the whole set was split into several chemical classes for modeling purposes. Models were validated by dividing the complete set into several training and test sets to allow for external predictions to be made.

Authors **Gallegos, A.**; Carbó-Dorca, R.; Ponec, R., Waisser, K.

Title Similarity approach to QSAR. Application to antimycobacterial benzoxazines

Reference *Int. J. Pharm., 269,* **2004**, 51-60.

Abstract The antimycobacterial activity in 8 series of substituted 3-phenyl-2H-benzoxazine-2, 4(3H)-dithiones and 3-(phenyl)-4-thioxo-2H-benzoxazine-2, 4 (3H)-diones has been studied using quantum molecular similarity approach. The approach is based on the use of fragment self-similarity measures as new universal molecular descriptors applicable for the design of novel theoretical QSAR models. Using this approach it was possible to show that while traditional QSAR models were able to describe the activity only within each of 8 sets of studied molecules individually, the proposed approach is much more general and single universal QSAR model was proposed which describes the activity of all 38 studied molecules in all 8 studied series together. The replacement of oxo group for thioxo group in position 4 on benzoxazine ring of antitubercular 3-(phenyl)--2H-benzoxazine-2, 4 (3H)-diones increases the activity more, as the similar replacement in position 2.

Authors **Gallegos, A.**; Gironés, X.; Carbó-Dorca, R

Title Topological Quantum Similarity Measures: applications in QSAR.

Reference In *Proceedings of the 5th Girona Seminar on Molecular Similarity*.
Sen, K. (Ed.) Nova Press. *In press.*

Abstract A novel method for computing new descriptors to construct Quantitative Structure-Toxicity Relationships is presented. First, a brief review on the classical graph theory is presented and, then, the link with molecular similarity is drawn. In the applications section, molecular topological indices are calculated using the interatomic Molecular Quantum Similarity Measure with a Coulomb weight operator. The use of similarity matrices instead of classical topological ones has been adopted according to the connection between molecular topology and the general theory of Quantum Similarity. Afterwards, the molecular descriptors, which include the structural information necessary to properly describe the system, are employed to derive numerical correlation with toxicities. The QSAR model is built using a multilineal regression technique.

Finally, some application examples are presented, including polycyclic aromatic hydrocarbons and aquatic toxicants, demonstrating the applicability of the exposed methodology.

Authors **Gallegos, A**.; Carbó-Dorca, R.; Lodier, F.; Cancès, E.; Savin, A

Title Maximal Probability Domains in Linear Molecules

Reference In *Proceedings of the 6th Girona Seminar on Molecular Similarity*. Kluwer. *Submitted.*

Abstract Regions of space are defined to maximize the probability to find n electrons in it. Their chemical significance, and their relationship to the electron localization function (ELF) is explored by analyzing the results for a few linear molecules: LiH, BH, N_2, CO, CS, C_2H_2, and C_4H_2.

Authors **Gallegos, A.**; Gironés, X.

Title Topological Quantum Similarity Indices based on Fitted Densities: Theoretical Background and QSPR Application.

Reference *In preparation.*

4 CONTRIBUTIONS TO CONFERENCES

4.1 Posters

Authors **Gallegos, A.**; Amat, L.; Carbó-Dorca, R.

Title Molecular Electronic Density Fitting Using Elementary Jacobi Rotations under Atomic Shell Approximation (ASA)

Conference *Workshop on Electron Densities and Electron Distributions*
University of Basque Country. Donostia, *April* **2000**

Authors **Gallegos, A.**; Carbó-Dorca, R.

Title Molecular Quantum Similarity Measures: Applications in Quantitative Structure-Activity Relationships

Reference *European Summerschool in Quantum Chemistry (ESQC00)*
Riolo Terme, Italy, *September* **2000**

Authors **Gallegos, A.**; Gironés, X.; Carbó-Dorca, R.

Title Topological Quantum Similarity Indices: Novel QSAR descriptors

Reference *Summerschool in Molecular Physics & Quantum Chemistry*
Oxford, *September* **2001**

Authors **Gallegos, A.**; Gironés, X.

Title Topological Quantum Similarity Indices based on Fitted Densities: Theoretical Background and QSPR Application

Reference *The 11th International Workshop on Quantitative Structure-Activity Relationships in the Human Health and Environmental Sciences (QSAR 2004)*
John Moores Liverpool University. Liverpool, *May* **2004**

4.2 Oral communications

Authors **Gallegos, A.**; Robert, D.; Carbó-Dorca, R.
Title Aplicació de Mesures de Semblança Molecular Quàntica en l'estudi de Relacions Estructura-Toxicitat en Hidrocarburs Aromàtics Policíclics
Reference *Internal Workshop of Institute of Computational Chemistry*
Girona, *March* **2000**

Authors **Besalú, E.**; Gallegos, A.; Carbó-Dorca, R.
Title Validació sistemàtica de la utilitat dels índexs topològics quàntics en estudis QSAR
Reference *Internal Workshop of Institute of Computational Chemistry*
Girona, *March* **2000**

Authors **Gallegos, A.**; Robert, D.; Carbó-Dorca, R.
Title Application of Molecular Quantum Similarity Measures in Structure-Toxicity Relationships in Polycyclic Aromatic Hydrocarbons
Reference *Final Meeting of COMET European Project*
Institute of Computational Chemistry. Girona, *December* **2000**

Authors **Gallegos, A.**; Gironés, X.; Carbó-Dorca, R.
Title Topological Quantum Similarity Measures: Applications in QSAR
Reference *V Girona Seminar on Molecular Similarity (V GSMS)*
Institute of Computational Chemistry. Girona, *July* **2001**

Authors **Gallegos, A.**; Besalú, E.
Title COMBINATOR: aplicació a l'estudi de tuberculostàtics
Reference *Internal Workshop of Institute of Computational Chemistry*
Girona, *April*, **2002**

Authors	**Gallegos, A.**; Carbó-Dorca, R.; Cronin, M.
Title	Molecular Quantum Similarity Analysis of Estrogenic Activity
Reference	*VI Girona Seminar on Molecular Similarity (VI GSMS)*
	Institute of Computational Chemistry. Girona, *July* **2003**

Authors	**Savin, A**.; Gallegos-Saliner, A.; Carbó-Dorca, R.; Cancès, E.; Lodier, F.
Title	How electrons guard the space
Reference	*VI Girona Seminar on Molecular Similarity (VI GSMS)*
	Institute of Computational Chemistry. Girona, *July* **2003**

Annex: Contributions

Ana Gallegos, et al. “Structure-toxicity relationships of polycyclic aromatic hydrocarbons using molecular quantum similarity”. *Journal of Computer-Aided molecular design*. Vol. 15, no. 1 (January 2001) : p. 67-80

doi:10.1023/A:1011150003086

http://www.kluweronline.com/article.asp?PIPS=275534

Abstract

The establishment of quantitative structure-activity relationship (QSAR) models for the toxicity of polycyclic aromatic hydrocarbons (PAHs) is described. Two properties, *in vitro* percutaneous absorption in rat skin and discrete levels of carcinogenic activity, are examined using molecular quantum similarity measures (MQSM). The results show that MQSM produces comparable, or even better, results than other approaches using physicochemical, topological and quantum-chemical molecular descriptors. Furthermore, a careful analysis puts into evidence that most of the information characterized by the original descriptors is in fact contained in the molecular density functions, the basis of MQSM. The present paper, together with several other reported by our laboratory, proves that MQSM might be appropriate theoretical tools for QSAR and computer-aided drug design, comparable to other highly predictive QSAR methodologies.

Keywords

atomic shell approximation (ASA) density functions, carcinogenic activity, molecular quantum similarity measures (MQSM), percutaneous absorption, polycyclic aromatic hydrocarbons (PAHs), quantitative structure-activity relationships (QSAR)

Article ID: 275534

Ana Gallegos Salinier, et al.“Molecular quantum similarity analysis of estrogenic activity”.
Journal of Chemical Information and Computer Sciences. Vol. 43, issue 4 (2004) : p. 1166-1176

http://pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ci034014a

http://pubs.acs.org/cgi-bin/article.cgi/jcisd8/2003/43/i04/pdf/ci034014a.pdf

http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2003/43/i04/abs/ci034014a.html

Received January 24, 2003 Web Release Date**:** June 27, 2003

Abstract:

The main objective of this study was to evaluate the capability of 120 aromatic chemicals to bind to the human alpha estrogen receptor (hERα) by the use of quantum similarity methods. The experimental data were segregated into two categories, i.e., those compounds with and without estrogenicity activity (active and inactive). To identify potential ligands, semiquantitative structure-activity relationships were developed for the complete set correlating the presence or lack of binding affinity to the estrogen receptor with structural features of the molecules. The structure-activity relationships were based upon molecular similarity indices, which implicitly contain information related to changes in the electron distributions of the molecules, along with indicator variables, accounting for several structural features. In addition, the whole set was split into several chemical classes for modeling purposes. Models were validated by dividing the complete set into several training and test sets to allow for external predictions to be made.

Ana Gallegos, et al. “Similarity approach to QSAR. Application to antimycobacterial benzoxazines”. *International Journal of pharmaceutics*. Vol. 269, issue 269 (January 2004) : p. 51-60

doi:10.1016/j.ijpharm.2003.08.013

Received 24 March 2003; revised 22 July 2003; accepted 30 August 2003. ; Available online 13 November 2003.

Abstract

The antimycobacterial activity in six series of substituted 3-phenyl-2*H*-benzoxazine-2,4(3*H*)-dithiones and 3-(phenyl)-4-thioxo-2*H*-benzoxazine-2,4(3*H*)-diones has been studied using a quantum molecular similarity approach. The approach is based on the use of fragment self-similarity measures as new universal molecular descriptors applicable for the design of novel theoretical QSAR models. Using this approach it was possible to show that while traditional QSAR models were able to describe the activity only within each of the six sets of studied molecules individually, the proposed approach is much more general and a single universal QSAR model describing the activity of all the 39 studied molecules in all the studied series together was built. The replacement of the oxo group by the thioxo group in position 4 on the benzoxazine ring of the antitubercular 3-(phenyl)-2*H*-benzoxazine-2,4(3*H*)-diones increases the activity, as well as the similar replacement in position 2.

Author Keywords: Quantum molecular similarity; Theoretical QSAR models; Mycobacterium; Tuberculostatics; Benzoxazine

ACRONYMS

A non-exhaustive compilation of the acronyms used throughout this thesis is listed below. It must be noted that the notation is the same, independently of the number, i.e. singular or plural, of the term.

3D-QSAR	Three-dimensional QSAR
ADME	Absorption-Distribution-Metabolism-Excretion
ALS	Adaptive Least Squares
AM1	Austin Model 1
ANN	Artificial Neural Networks
AO	Atomic Orbital
AQSM	Atomic Quantum Similarity Measures
ASA	Atomic Shell Approximation
CA	Cluster Analysis
CADD	Computer-Aided Drug Design
CAMD	Computer-Aided Molecular Design
CCA	Canonical Correlation Analysis
CCD	Central Composite Design
CFA	Correspondence Factor Analysis
CNDO	Complete Neglect of Differential Overlap
CoMFA	Comparative Molecular Field Analysis
CoRePa	Common Reactive Pattern approach
CR	Continuum Regression
CS	Classical Scaling
CSA	Cluster Significance Analysis
CV	Cross-Validation
DF	Density Function
DFT	Density Functional Theory
DME	Density Matrix Element
DOD	D-Optimal Design
EC50	Effective Concentration for 50% of maximal effect
ECVAM	European Centre for the Validation of Alternative Methods
ED50	Effective Dose for 50% of maximal effect
EIM	Electronic Index Methodology
EJR	Elementary Jacobi Rotations

ELF	Electron Localization Function
EPA	Environmental Protection Agency
ER	Estrogen Receptor
FA	Factor Analysis
FD	Factorial Design
FFD	Fractional Factorial Design
FW	Free Wilson Analysis
G/PLS	Genetic Partial Least Squares
GFA	Genetic Function Approximation
GTO	Gaussian-Type Orbital
hERα	α-human Estrogen Receptor
HIS	Hybrid Intelligent Systems
HOMO	Highest Occupied Molecular Orbital
HTS	High Throughput Screening
IC50	Inhibitory Concentration for 50% of maximal inhibition
IGC50	Inhibitory Growth Concentration for 50% of maximal growth inhibition
KE	Kinetic Energy
KE DF	Kinetic Energy Density Function
kNN	k-Nearest Neighbours
Kow	Octanol-water partition coefficient
LCAO	Linear Combination of Atomic Orbitals
LD50	Letal Dose at which 50% of species die
LDA	Linear Discriminant Analysis
LDOS	Local Density Of States
LFER	Linear Free Energy Relationships
LmO	Leave-many-Out
log P	Partition coefficient octanol/water
LOO	Leave-One-Out
LUMO	Lowest Unoccupied Molecular Orbital
MASA	Multicentre ASA
MCF-7	Human breast cancer cells
MDS	Multidimensional Scaling Techniques
MFA	Molecular Field Analysis
MIC	Minimum Inhibitory Concentration
MLR	Multiple Linear Regression
MM	Molecular Mechanics

MO	Molecular Orbital
MPVM	Most Predictive Variables Method
MQS	Molecular Quantum Similarity
MQSI	Molecular Quantum Similarity Indices
MQSM	Molecular Quantum Similarity Measure
MQS-SM	Molecular Quantum Self-Similarity Measure
MSA	Molecular Shape Analysis
MTI	Schultz index
NIPALS	Nonlinear Iterative Partial Least Squares
NLM	Non-Linear Mapping
NMR	Nuclear Magnetic Resonance
NMR	Nuclear Magnetic Resonance
NSS	Nested Summation Symbols
OLS	Ordinary Least Squares
PARC	Pattern Recognition
PASA	Promolecular ASA
PC	Principal component
PC1	First Principal Component
PCA	Principal Component Analysis
PCR	Principal Component Regression
PLS	Partial Least Squares
PRESS	Predictive Residual Error Sum of Squares
Q2	Cross-validated explained variance
QM	Quantum Mechanics
QO	Quantum Object
QPLS	Quadratic PLS
QS	Quantum Similarity
QSAR	Quantitative Structure-Activity Relationships
QSI	Quantum Similarity Index
QSM	Quantum Similarity Measure
QSPR	Quantitative Structure-Property Relationships
QS-SM	Quantum Self-Similarity Measures
QST	Quantum Similarity Theory
QSTR	Quantitative Structure-Toxicity Relationships
RMS	Root Mean Square
RSA	Receptor Surface Analysis

SAR	Structure-Activity Relationships
SD	Standard Deviation
SDEC	Standard Deviation of Errors of Calculation
SDEP	Standard Deviation of Errors of Prediction
SDEP	Standard Deviation of Errors of Prediction
SIMCA	Soft Independent Modelling Class Analogy method
SM	Similarity Matrix
SMD	Statistical Molecular Design
SSPE	Sum of Squares Prediction Errors
STO	Slater-Type Orbital
SVD	Single Value Decomposition
TGSA	Topo-Geometrical Superposition Algorithm
TI	Topological Index
TM	Topological Matrix
TQSI	Topological Quantum Similarity Index

Ana Gallegos Saliner

June, 2004

Look for the bare necessities
The simple bare necessities
Forget about your worries and your strife
I mean the bare necessities
Old Mother Nature's recipes
That brings the bare necessities of life
Wherever I wander, wherever I roam
I couldn't be fonder of my big home
The bees are buzzin' in the tree
To make some honey just for me
When you look under the rocks and plants
And take a glance at the fancy ants
Then maybe try a few
The bare necessities of life will come to you
They'll come to you!

Look for the bare necessities
The simple bare necessities
Forget about your worries and your strife
I mean the bare necessities
That's why a bear can rest at ease
With just the bare necessities of life
Now when you pick a pawpaw
Or a prickly pear
And you prick a raw paw
Next time beware
Don't pick the prickly pear by the paw
When you pick a pear
Try to use the claw
But you don't need to use the claw
When you pick a pear of the big pawpaw
Have I given you a clue ?
The bare necessities of life will come to you
They'll come to you!
So just try and relax, yeah cool it
Fall apart in my backyard
'Cause let me tell you something little britches
If you act like that bee acts, uh uh
You're working too hard
And don't spend your time lookin' around
For something you want that can't be found
When you find out you can live without it
And go along not thinkin' about it
I'll tell you something true
The bare necessities of life will come to you

Baloo's Song

Terry Gilkyson