Roberto Todeschini and Viviana Consonni

Molecular Descriptors for Chemoinformatics

Volume I: Alphabetical Listing

Second, Revised and Enlarged Edition

WILEY-VCH Verlag GmbH & Co. KGaA

Roberto Todeschini and
Viviana Consonni

Molecular Descriptors for Chemoinformatics

Methods and Principles in Medicinal Chemistry

Previous Volumes of this Series:

Roberto Todeschini and Viviana Consonni

Molecular Descriptors for Chemoinformatics

Volume I: Alphabetical Listing

Second, Revised and Enlarged Edition

WILEY-VCH Verlag GmbH & Co. KGaA

Series Editors

Prof. Dr. Raimund Mannhold
Molecular Drug Research Group
Heinrich-Heine-Universität
Universitätsstrasse 1
40225 Düsseldorf
Germany
mannhold@uni-duesseldorf.de

Prof. Dr. Hugo Kubinyi
Donnersbergstrasse 9
67256 Weisenheim am Sand
Germany
kubinyi@t-online.de

Prof. Dr. Gerd Folkers
Collegium Helveticum
STW/ETH Zurich
8092 Zurich
Switzerland
folkers@collegium.ethz.ch

Volume Authors

Prof. Dr. Roberto Todeschini
Dept. of Environm. Sciences
University Milano-Bicocca
Piazza della Scienza 1
0126 Milano
Italy
roberto.todeschini@unimib.it

Dr. Viviana Consonni
Dept. Environm. Sciences
University Milano-Bicocca
Piazza della Scienza 1
20126 Milano
Italy
viviana.consonni@unimib.it

Cover Description
Background: Front of a tablet fragment of a middle Assyrian code.
Foreground: Drawing of phenylurea.

Library of Congress Card No.: applied for

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Bibliographic information published by the Deutsche Nationalbibliothek
Die Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available on the Internet at http://dnb.d-nb.de

Cover Design Grafik-Design Schulz, Fußgönheim
Typesetting Thomson Digital, Noida, India
Printing betz-druck GmbH, Darmstadt
Binding Litges & Dopf GmbH, Heppenheim

Printed in the Federal Republic of Germany
Printed on acid-free paper

ISBN 978-3-527-31852-0

This book is dedicated with love to
Alessia, Davide, Edoardo, Marco, Milo, Marilena, and Giovanni

A good scientist should have the imagination of a child, the determination of a boy, the rationality of a man, and the experience of an old man. The difficulty is to have all these qualities at the same time.

R.T.

Any alternative viewpoint with a different emphasis leads to an inequivalent description. There is only one reality but there are many points of view.
It would be very narrow-minded to use only one context: we have to learn to be able imagining points of view.

Hans Primas *in Chemistry, Quantum Mechanics and Reductionism*
(Springer-Verlag, 1981)

Contents

Volume I

Volume II

The Authors

Roberto Todeschini is full professor of chemometrics at the Department of Environmental Sciences of the University of Milano-Bicocca (Milano, Italy), where he constituted the Milano Chemometrics and QSAR Research Group. His main research activities concern chemometrics in all its aspects, QSAR, molecular descriptors, multicriteria decision making, and software development. President of the International Academy of Mathematical Chemistry, president of the Italian Chemometric Society, and "ad honorem" professor of the University of Azuay (Cuenca, Ecuador), he is author of more than 170 publications in international journals and of the books "The Data Analysis Handbook," by I.E. Frank and R. Todeschini, 1994, and "Handbook of Molecular Descriptors," by R. Todeschini and V. Consonni, 2000.

Viviana Consonni received her PhD in chemical sciences from the University of Milano in 2000 and is now full researcher of chemometrics and chemoinformatics at the Department of Environmental Sciences of the University of Milano-Bicocca (Milano, Italy). She is a member of the Milano Chemometrics and QSAR Research Group and has 10 years experience in multivariate analysis, QSAR, molecular descriptors, multicriteria decision making, and software development. She is author of more than 40 publications in peer-reviewed journals and of the book "Handbook of Molecular Descriptors," by R. Todeschini and V. Consonni, 2000. In 2006, she obtained the International Academy of Mathematical Chemistry Award for distinguished young researchers and, in June 2009, has been elected as youngest Member of the Academy.

Acknowledgments

The idea of producing the book *Molecular Descriptors for Chemoinformatics* was welcomed by several colleagues whom we warmly thank for their suggestions, revisions, bibliographic information, and moral support; we are particularly grateful to Alexander Balaban, Milan Randić, and several members of the International Academy of Mathematical Chemistry.

Particular thanks go also to Maurizio Bruschi, Ugo Cosentino, Mircea Diudea, and Marco Vighi for their help in revising some topics of the book. The Authors gratefully acknowledge the cooperation with and the support of the editorial staff of Wiley-VCH. In particular, we have to thank Nicola Oberbeckmann-Winter, Frank Weinreich, Carola Schmidt, Susanna Pohl, Claudia Nussbeck, Waltraud Wüst. The Authors also warmly thank Raymund Mannhold, editor of the series, for stimulations and timely pressure to complete this book on time.

Finally, since we have been fully absorbed in writing the book for a long time, we would like to heartily acknowledge Davide Ballabio, Andrea Mauri, Alberto Manganaro, and Manuela Pavan of our team not only for their help but also for their patience and assistance during this period.

Preface

In 2000, Roberto Todeschini and Viviana Consonni wrote the highly valuable *Handbook of Molecular Descriptors*, part of our series "Methods and Principles in Medicinal Chemistry." This volume achieved high acceptance among researchers in the field of drug discovery and design. Now, eight years later, the significant developments in the area of molecular descriptors necessitated a rather comprehensive revision.

All new descriptors, QSAR approaches and chemometric strategies proposed since 2000 have been included in this handbook. Several new topics such as biodescriptors, characteristic polynomial-based descriptors, property filters, scoring functions, and cell-based methods have been added. Other topics, such as substructure descriptors, autocorrelation descriptors, delocalization degree indices, weighted matrices, connectivity indices, and so on, have been completely rewritten.

Attention is also paid to recent methods dedicated to virtual screening of libraries of molecules, such as cell-based methods, property filters, and scoring functions.

Special attention has been paid to strategies for generating families of molecular descriptors based on generalization of classical molecular descriptors; dedicated entries are, for instance, Wiener-type indices, Randić-like indices, Balaban-like indices, connectivity-like indices, and variable descriptors.

Several entries have been joined together in larger entries allowing easier readability and comparability among the different molecular descriptors; for example, entries such as matrices of molecules, weighted matrices, substructure descriptors, and vertex degrees, which were enlarged in order to include a lot of definitions. Moreover, some didactical routes are introduced at the beginning of the book to indicate the main entries concerning a topic.

General entries concerning statistical indices, regression parameters, classification parameters, similarity/diversity were completely rewritten trying to give an exhaustive view of the functions used to characterize data, modeling, and similarity/diversity analysis. For example, more than 50 distance and similarity functions have been reported.

Numerical examples (more than 150) and several tables listing molecular descriptors for two benchmark data sets are added to help students and nonexpert readers to comprehend the algorithms better. Indeed, this new edition has been conceived not

only for experts and professional researchers but also for PhD students and young researchers who whish to enter the field of molecular descriptors and related areas, giving special attention to a didactical use of the book and suggesting some possible routes for didactical purposes.

Molecular descriptors implemented in the most common software for descriptor calculation are discussed and bibliographic references have been extended from 3300 to 6400.

The series editors would like to thank Roberto Todeschini and Viviana Consonni for their brilliant work on this second edition. We also want to express our gratitude to Nicola Oberbeckmann-Winter and Frank Weinreich of Wiley-VCH for their valuable contributions to this project.

May 2009

Raimund Mannhold, Düsseldorf
Hugo Kubinyi, Weisenheim am Sand
Gerd Folkers, Zürich

A Personal Foreword

The first idea to collect into a book all the knowledge about molecular descriptors dates back to September 1997, when we were at a meeting of physical chemistry in Taormina. In this beautiful landscape, we had time to think about the several different ways a molecule can be described and how these often derive from different theories developed in noncommunicating research fields.

At the beginning we collected and studied a lot of papers on the topic driven by childish hope to conclude the job in a few months; however, after a lot of days spent in libraries to search for papers and nights to read them, the initial enthusiasm left place to the awareness of the hugeness of information on this topic and the difficulty of organizing it in a systematic fashion. . . . Finally, two years and half later – working full time – we concluded the *Handbook of Molecular Descriptors*.

The book *Molecular Descriptors for Chemoinformatics* consequently derives from the success of our first book, from the need to update it for the huge number of new molecular descriptors produced from 2000 to 2008, and from the awareness to revise several parts of it and to organize differently the new work to made it also usable for didactical purposes.

Milan, May 2009

Roberto Todeschini
Viviana Consonni

Introduction

The effort being made today to organize Knowledge is a way of participating in the evolution of Knowledge itself. The significance of attempting such organization can be looked for in its ability not only to give information but also to create know-how. Knowledge organization provides not only a collection of facts, a store of information, but also a contribution to the growth of Knowledge, knowledge organization being itself one way of doing research. This is the true end of an encyclopedic guide. In effect, to think that the organization of Knowledge is separated from its production is completely arbitrary.

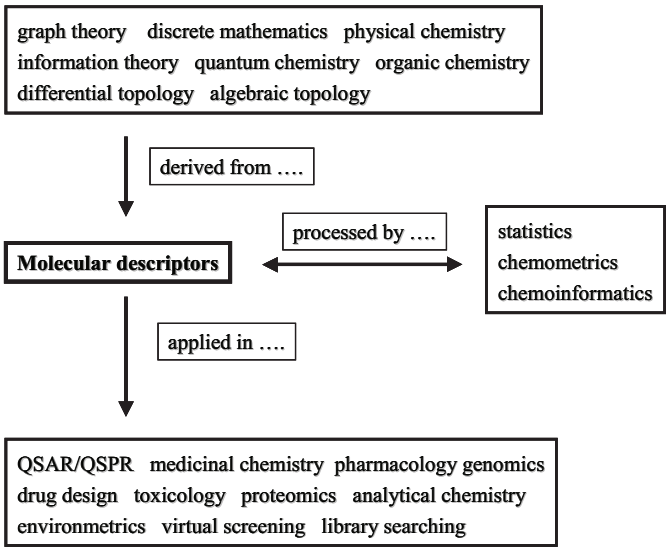
Knowledge should not be considered something given once and for all, based on some final basic theories, but as a *network of models* in progress. This network primarily consists of knots, that is, objects, facts, theories, statements, and models, and the links between the knots are relationships, comparisons, differences, and analogies: such a network is something more than a collection of facts, resulting in a powerful engine for analogical reasoning.

With these purposes in mind, the book *Molecular Descriptors for Chemoinformatics* has been conceived as an encyclopedic guide to molecular descriptors.

Molecular descriptors, tightly connected to the concept of molecular structure, play a fundamental role in scientific research, being the theoretical core of a complex network of knowledge.

Indeed, molecular descriptors are based on several different theories, such as quantum-chemistry, information theory, organic chemistry, graph theory, and so on, and are used to model several different properties of chemicals in scientific fields such as toxicology, analytical chemistry, physical chemistry, and medicinal, pharmaceutical, and environmental chemistry.

Moreover, to obtain reliable estimates of molecular properties, identify the structural features responsible for biological activity, and select candidate structures for new drugs, molecular descriptors are processed by several methods provided by statistics, chemometrics, and chemoinformatics. In particular, chemometrics for about 30 years has been developing classification and regression methods able to provide, although not always, reliable models for both reproducing the known experimental data and predicting the unknown data. The modeling process usually



has not only explanatory purposes but also predictive purposes. The interest in predictive models able to give effective reliable estimates has been largely growing in the last few years as they are more and more considered useful and safer tools for predicting data on chemicals.

It has been nearly 45 years since the QSAR modeling was brought first into the practice of agrochemistry and, successively, in drug design, toxicology, and industrial and environmental chemistry. Its growing importance in the years that followed may be attributed mainly to the rapid and extensive development in methodologies and computational techniques that have allowed to delineate and refine the many variables and approaches used to model molecular properties [Martin, 1979, 1998; Kubinyi, 1993a; Hansch and Leo, 1995; van de Waterbeemd, Testa *et al.*, 1997; Devillers, 1998; Kubinyi, Folkers *et al.*, 1998a, 1998b; Charton and Charton, 2002; Gasteiger, 2003b; Oprea, 2004].

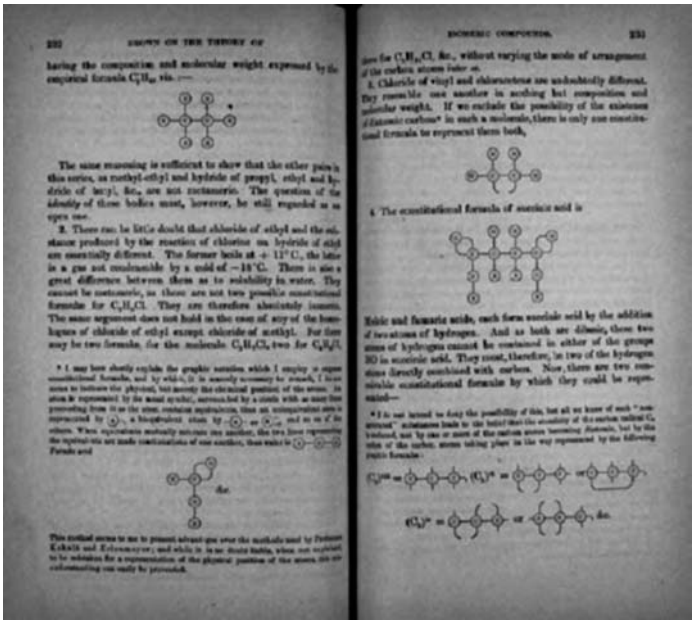
In recent years, "The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization" [Brown, 1998]. In fact, chemoinformatics encompasses the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and the use of chemical information [Gasteiger, 2003b; Oprea, 2003]; molecular descriptors play a fundamental role in all these processes being the basic tool to transform chemical information into a numerical code suitable for applying informatic procedures.

Molecular descriptors can be considered as the most important realization of the idea of Crum-Brown. His M.D. Thesis at the University of Edinburgh (1861), entitled "On the Theory of Chemical Combination", shows that he was a pioneer of mathematical chemistry science. In that, he developed a system of graphical

representation of compounds which is basically identical to that used today. His formulae were the first that showed clearly both valency and linking of atoms in organic compounds. Towards the conclusion of his M.D. thesis he wrote:

"It does not seem to me improbable that we may be able to form a mathematical theory of chemistry, applicable to all cases of composition and recomposition."

In 1864, he published an important study on the "Theory of isomeric compounds" in which, using his graphical formulae, he discussed various types of isomerism [Crum-Brown, 1864] guessing the link between mathematics and chemistry [Crum-Brown, 1867].



The molecular descriptor is the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into an useful number or the result of some standardized experiment [Todeschini and Consonni, 2000].

Attention is paid to the term "useful" with its double meaning: it means that the number can give more insight into the *interpretation* of the molecular properties *and/or* is able to take part in a model for the *prediction* of some interesting property of other molecules.

Why must we also accept "or"?

It should not be thought that molecular descriptors are good only if they show an evident link to some information about molecular structure, that is, they are easily interpretable from a structural/chemical point of view.

It often happens that interpretation of molecular descriptors could be weak, provisional, or completely lacking, but their predictive ability or usefulness in application to actual problems should be a strong motive for their use. On the other

hand, descriptors with poor predictive ability may be usefully retained in models when they are theoretically well founded and interpretable due to their ability to encode structural chemical information.

The incompletely realized comprehension of the chemical information provided by molecular descriptors cannot be systematically ascribed to weakness in the descriptors. Actually, our inability to reduce descriptor meaning to well-established chemical concepts is often because newly emergent concepts need new terms in the language and new, hierarchically connected levels for scientific explanation. Thus, what is often considered as scientific failure is sometimes the key to new, useful knowledge.

In any case, all the molecular descriptors must contain, to varying extents, chemical information, must satisfy some basic invariance properties and general requirements, and must be derived from well-established procedures, which enable molecular descriptors to be calculated for any set of molecules. It is obvious – almost trivial – that a single descriptor or a small number of numbers cannot wholly represent the molecular complexity or model all the physico-chemical responses and biological interactions. As a consequence, although we must get used to living with approximate models (*nothing is perfect*!), we have to keep in mind that "approximate" is not a synonym of "useless."

A molecular descriptor can be thought of as the mythological Dragon on the Babylon Ištar Gate (Pergamon Museum of Berlin), which actually is a mixing of several different animals, each corresponding to a different part of the Dragon body; likewise, a molecular descriptor has several different meanings which depend on one's point of view.

Several scientists think that only molecular descriptors derived from quantum-chemistry, which they consider the unique "true" chemical theory, or from simple experimental properties (e.g., partition coefficients or molar refractivity), thought of as the "experimental chemical evidence", can be legitimately used in QSAR/QSPR modeling. For several years, predictive ability, overfitting, and chance correlation have been not discussed for models derived from those "well-founded" molecular descriptors. On the contrary, a great criticism and skepticism arise against models based on descriptors derived from chemical graph theory, statistics applied to geometrical representations of molecules, and other innovative approaches. In most of the cases, these molecular descriptors give better results than the classical ones but their validity is often brought into question, although it is obvious that chance correlation can be obtained by any kind of descriptor, independent of their interpretability and scientific "nobility."

The historical development of molecular descriptors reflects some of the distinctive characteristics of the most creative scientists, that is, their capability of being at the same time engaged and/or detached, rational and/or quirky, and serious and/or not so serious. Science is a game and the best players appreciate not only the beauty of a discovery by a precise and logical reasoning but also the taste of making a guess, of proposing eccentric hypotheses, of being doubtful and uncertain when confronted by new and complex problems. Molecular descriptors constitute a research field where the most diverse strategies for scientific discovery can be applied.

Molecular descriptors will probably play an increasing role in science growth. The availability of large numbers of theoretical descriptors that provide diverse sources of chemical information would be useful to better understand relationships between molecular structure and experimental evidence, also taking advantage of more and more powerful methods, computational algorithms, and fast computers. However, as before, deductive reasoning and analogy, theoretical statements and hazardous hypotheses, and determination and perplexity still remain fundamental tools.

The field of molecular descriptors is strongly interdisciplinary and involves a huge number of different theories. For the definition of molecular descriptors, a knowledge of algebra, graph theory, information theory, computational chemistry, and theories of organic reactivity and physical chemistry is usually required, although at different levels. For the use of the molecular descriptors, a knowledge of statistics, chemometrics, chemoinformatics, and the principles of the QSAR/QSPR approaches is necessary in addition to the specific knowledge of the problem. Moreover, programming and sophisticated software and hardware are often inseparable fellow travelers of the researcher in this field.

This book tries to meet the great interest that the scientific community is showing about all the tools that chemoinformatics provides for a quick acquisition and mining of information on chemical compounds and evaluation of their effects on humans and environment in general. Besides the consolidated interest for the quantitative modeling of biological activity, physico-chemical properties, and environmental behavior of compounds, an increasing interest has been shown by the scientific community in recent years in the fields of combinatorial chemistry, high-throughput screening, similarity searching, and database mining, for which several approaches particularly suitable for informatic treatment have been proposed. Thus, several disciplines such as chemistry, pharmacology, environmental protection, drug design, toxicology, and quality control for health and safety, derive great advantages from these methodologies in their scientific and technological development.

The book, *Molecular Descriptors for Chemoinformatics*, collects the definitions, formulas, and short comments of most of the molecular descriptors known in chemical literature. The molecular descriptor definitions, about 3300, are organized in alphabetical order.

The importance of each definition is not related to its length. Only a few old descriptors, abandoned or demonstrated as wrong, were intentionally left out to avoid confusion. An effort was also made to collect appropriate bibliographic information under each definition. We are sorry if any relevant descriptor and/or work has been

missed out; although this has not been done deliberately, we take full responsibility for any omission.

Some molecular descriptors are grouped under a specific topic using a mixed taxonomy based on different points of view, in keeping with the leading idea of the book to promote learning by comparison. These book topics were mainly distinguished according to the *physico-chemical meaning* of molecular descriptors or the specific *mathematical tool* used for their calculation.

Some basic concepts and definitions of statistics, chemometrics, algebra, graph theory, similarity/diversity analysis, which are fundamental tools in the development and application of molecular descriptors, are also discussed in the book in some detail. More attention was paid to information content, multivariate correlation, model complexity, variable selection, applicability domain, and parameters for model quality estimation, as these are the characteristic components of modern QSAR/QSPR modeling.

The book contains nothing about the combinatorial algorithms for the generation and enumeration of chemical graphs, the basic principles of statistics, informatic code for descriptor calculation, or experimental techniques for measuring physicochemical, technological, and biological responses. Moreover, relevant chemometric methods such as Partial Least Squares regression (PLS) and other regression methods, classification methods, cluster analysis, and artificial neural networks are simply quoted, references are given, but no theoretical aspect is presented. Analogously, computational chemistry methods are quoted only as important tools for theoretical calculations, but no claim is made here to their detailed explanation.

Molecular descriptors on the Web

The authors together with the other members of the Milano Chemometrics and QSAR Research Group of the University of Milano-Bicocca (Milan, Italy) activated in 2007 a web site dedicated to molecular descriptors (http://www.moleculardescriptors.eu). This web site aims at promoting information exchange among all the scientists who propose new molecular descriptors and/or apply molecular descriptors in their research.

This web site collects different kinds of information related to molecular descriptors, thus helping researchers in their daily work. *Software*, *books*, *links*, *events*, *tutorials*, and *news* are organized in a systematic way to allow a quick and easy consultation. Moreover, this web site provides a *forum* on molecular descriptors, where experts can initiate discussions on different topics as well as collect lists of bibliographic references about descriptors or discuss their interpretations.

The authors would be grateful to all researchers who would like to send their observations and comments on the book contents, information about new descriptors, and bibliographic references. E-mail submissions can be made at info@moleculardescriptors.eu.

Bibliographic references

The reference list covers a period between 1741 and 2008, lists about 6400 references, for almost 7000 authors and 450 periodicals. Author names are given by the last name, followed by the initials of the first and middle names, if present.

In addition to the cited references, a thematic bibliography with almost 5,000 entries is available. Here, bibliographic references have been collected for some 70 topics of general interest. These references are additional references to those already quoted in the main text of the book. Topics are listed in alphabetic order, from "ADME properties" to "Wiener index". Selection of the topics was based on the most frequent keywords encountered in publications about molecular descriptors and related research fields.

The thematic bibliography is available as supplementary online material from the book homepape a wiley-vch.de. Please visit http://www.wiley-vch.de./publish/en/books/3-527-31852-6 for details.

Historical Perspective

The history of molecular descriptors is closely related to the history of what can be considered one of the most important scientific concepts of the last part of the nineteenth century and the whole twentieth century, that is, the concept of molecular structure.

The years between 1860 and 1880 were characterized by a strong debate on the concept of molecular structure, arising from the studies on substances showing optical isomerism and the studies of Kekulé (1861–1867) on the structure of benzene. The concept of the molecule thought of as a three-dimensional body was first proposed by Butlerov (1861–1865), Wislicenus (1869–1873), Van't Hoff (1874–1875), and Le Bel (1874). The publication in French of the revised edition of *La chimie dans l'éspace* by Van't Hoff in 1875 is considered a milestone in the three-dimensional conception of the chemical structures.

QSAR history started a century earlier than the history of molecular descriptors, being closely related to the development of the molecular structure theories.

QSAR modeling was born in toxicology field. Attempts to quantify relationships between chemical structure and acute toxic potency have been part of the toxicological literature for more than 100 years. In the defense of his thesis entitled "Action de l'alcohol amylique sur l'organisme" at the Faculty of Medicine, University of Strasbourg, France, on January 9, 1863, Cros noted a relationship existed between the toxicity of primary aliphatic alcohols and their water solubility. This relationship demonstrated the central axiom of structure–toxicity modeling, that is, the toxicity of substances is governed by their properties, which are determined in turn by their chemical structure. Therefore, there are inter-relationships among structure, properties, and toxicity.

Crum-Brown and Fraser (1868–1869) [Crum-Brown, 1864, 1867; Crum-Brown and Fraser, 1868] proposed the existence of a correlation between biological activity of different alkaloids and their molecular constitution. More specifically, the physiological action of a substance in a certain biological system (Φ) was defined as a function (f) of its chemical constitution (C):

$$\Phi = f(C).$$

Thus, an alteration in chemical constitution, ΔC, would be reflected by an effect on biological activity, $\Delta\Phi$. This equation can be considered the first general formulation of a quantitative structure–activity relationship.

The periodic table proposed by Mendelev [1870] gave relationships between atomic structure and properties; in the following years, the concept of an internal structure of atoms and molecules became more and more relevant and important studies were conducted such as those by G.N. Lewis [Lewis, 1916, 1923].

A hypothesis on the existence of correlations between molecular structure and physico-chemical properties was reported in the work of Körner [1874], which dealt with the synthesis of disubstituted benzenes and the discovery of *ortho*, *meta*, and *para* derivatives: the different colors of disubstituted benzenes were thought to be related to differences in molecular structure and the indicator variables for *ortho*, *meta*, and *para* substitution can be considered as the first three molecular descriptors [Körner, 1869, 1874].

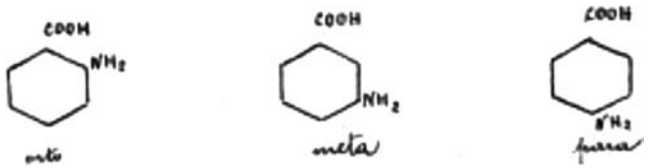


Figure from Chimica Organica by W. Korner, 1921, personal document of the Authors.

Ten years later, Mills [Mills, 1884] published in the *Philosophical Magazine* a study "On melting point and boiling point as related to composition."

The quantitative property–activity models, commonly referred to as those marking the beginning of systematic QSAR/QSPR studies [Richet, 1893], have come out from the search for relationships between the potency of local anesthetics and the oil/water partition coefficient [Meyer, 1899], between narcosis and chain length [Overton, 1901, 1991], and between narcosis and surface tension [Traube, 1904]. In particular, the concepts developed by Meyer and Overton are often referred to as the Meyer–Overton theory of narcotic action [Meyer, 1899; Overton, 1901].

The first theoretical QSAR/QSPR approaches date back to the end of 1940s and are those relating biological activities and physico-chemical properties to theoretical numerical indices derived from the molecular structure.

The → *Wiener index* [Wiener, 1947a, 1947b] and the → *Platt number* [Platt, 1947], proposed in 1947 to model the boiling point of hydrocarbons, were the first theoretical molecular descriptors based on the graph theory.

In the early 1960s, several other molecular descriptors were proposed, which marked the beginning of systematic studies on molecular descriptors, mainly based on the graph theory [Charton, 1964; Fujita, Iwasa *et al.*, 1964; Gordon and Scantlebury, 1964; Smolenskii, 1964; Spialter, 1964a; Hansch, Deutsch *et al.*, 1965; Reichardt, 1965; Hansch and Anderson, 1967; Balaban and Harary, 1968; Harary, 1969a; Kier, 1971; Cammarata, 1972; Gutman and Trinajstić, 1972; Hosoya, 1972c; Verloop, 1972].

The use of quantum-chemical descriptors in QSAR/QSPR modeling dates back to early 1970s [Kier, 1971], although they actually were conceived several years before to

encode information on relevant properties of molecules in the framework of quantum-chemistry. During 1930–1960, the pioneering studies that signaled the beginning of quantum-chemistry are those of Pauling [Pauling, 1932, 1939] and Coulson [Coulson, 1939] on the chemical bond, of Sanderson on electronegativity [Sanderson, 1952] and of Fukui [Fukui, Yonezawa *et al.*, 1954] and Mulliken on electronic distribution [Mulliken, 1955a].

Once the concept of molecular structure was definitively consolidated by the successes of quantum-chemistry theories and the approaches to the calculation of numerical indices encoding molecular structure information were accepted, all the constitutive elements for the take-off of QSAR strategies were available.

From the Hammett equation [Hammett, 1935, 1937], the seminal work of Hammett gave rise to the "$\sigma-\rho$" culture in the delineation of substituent effects on organic reactions, whose aim was to search for linear free energy relationships (LFER) [Hammett, 1938]: steric, electronic, and hydrophobic constants were derived for several substituents and used in an additive model to estimate the biological activity of congeneric series of compounds.

In the 1950s, the fundamental works of Taft in physical organic chemistry laid the foundation of relationships between physico-chemical properties and solute–solvent interaction energies (linear solvation energy relationships, LSER), based on steric, polar, and resonance parameters for substituent groups in congeneric compounds [Taft, 1952, 1953a, 1953b].

In the mid-1960s, led by the pioneering works of Hansch [Hansch, Maloney *et al.*, 1962; Hansch, Muir *et al.*, 1963; Fujita, Iwasa *et al.*, 1964], the QSAR/QSPR approach began to assume its modern look.

In 1962, Hansch, Maloney and Fujita [Hansch, Maloney *et al.*, 1962] published their study on the structure–activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity. Using the octanol/water system, a whole series of partition coefficients was measured and, thus, a new hydrophobic scale was introduced for describing the inclination of molecules to move through environments characterized by different degrees of hydrophilicity such as blood and cellular membranes. The delineation of Hansch models led to explosive development in QSAR analysis and related approaches [Hansch and Leo, 1995]. This approach known with the name of *Hansch analysis* became and it still is a basic tool for QSAR modeling.

In the same years, Free and Wilson [Free and Wilson, 1964] developed a model of additive substituent contributions to biological activities, giving a further push to the development of QSAR strategies. They proposed to model a biological response on the basis of the presence/absence of substituent groups on a common molecular skeleton [Free and Wilson, 1964; Kubinyi, 1988b]. This approach, called "de novo approach" when presented in 1964, was based on the assumption that each substituent gives an additive and constant effect to the biological activity regardless of the other substituents in the rest of the molecule.

At the end of 1960s, a lot of structure–property relationships were proposed based not only on substituent effects but also on indices describing the whole molecular structure. These theoretical indices were derived from a topological representation of

molecule, mainly applying the graph theory concepts, and then usually referred to as 2D-descriptors.

The fundamental works of Balaban [Balaban and Harary, 1971; Balaban, 1976a], Randić [Randić, 1974, 1975b], and Kier, Hall *et al.* [Kier, Hall *et al.*, 1975] led to further significant developments in QSAR approaches based on topological indices.

As a natural extension of the topological representation of a molecule, the geometrical aspects of molecular structures were taken into account since the mid-1980s, leading to the development of the 3D-QSAR, which exploits information on the molecular geometry. Geometrical descriptors were derived from the 3D spatial coordinates of a molecule and, among them, there were shadow indices [Rohrbaugh and Jurs, 1987a], charged partial surface area descriptors [Stanton and Jurs, 1990], WHIM descriptors [Todeschini, Lasagni *et al.*, 1994], gravitational indices [Katritzky, Mu *et al.*, 1996b], EVA descriptors [Ferguson, Heritage *et al.*, 1997], 3D-MoRSE descriptors [Schuur, Selzer *et al.*, 1996], EEVA descriptors [Tuppurainen, 1999a], and GETAWAY descriptors [Consonni, Todeschini *et al.*, 2002a].

At the end of 1980s, a new strategy for describing molecule characteristics was proposed, based on molecular interaction fields, which consist of interaction energies between a molecule and probes, at specified spatial points in 3D space. Different probes (such as a water molecule, methyl group, hydrogen, etc.) were used for evaluating the interaction energies in thousands of grid points where the molecule was embedded. As the final result of this approach, a scalar field (a lattice) of interaction energy values characterizing the molecule was obtained. The first formulation of a lattice model to compare molecules by aligning them in 3D space and extracting chemical information from molecular interaction fields was first proposed by Goodford [Goodford, 1985] in the GRID method and then by Cramer, Patterson, Bunce [Cramer III, Patterson *et al.*, 1988] in the Comparative Molecular Field Analysis (CoMFA).

Still based on molecular interaction fields, several other methods were successively proposed and, among them, there were Comparative Molecular Similarity Indices Analysis (CoMSIA) [Klebe, Abraham *et al.*, 1994], Compass method [Jain, Koile *et al.*, 1994], G-WHIM descriptors [Todeschini, Moro *et al.*, 1997], Voronoi field analysis [Chuman, Karasawa *et al.*, 1998], VolSurf descriptors [Cruciani, Pastor *et al.*, 2000], and GRIND descriptors [Pastor, Cruciani *et al.*, 2000].

Finally, the scientific community has been showing an increasing interest in recent years for virtual screening and design of chemical libraries, for which several similarity/diversity approaches, cell-based methods, and scoring functions have been proposed mainly based on *substructure descriptors* such as molecular fingerprints [Gasteiger, 2003b; Kubinyi, 2003b; Oprea, 2004].

QSAR/QSPR Modeling

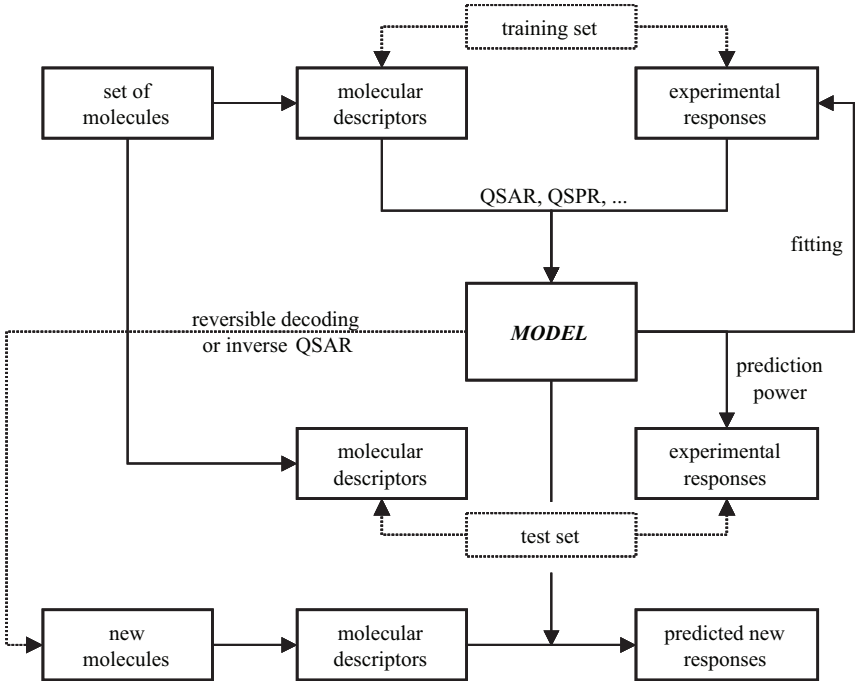
Quantitative Structure–Activity Relationships (QSARs) are the final result of the process that starts with a suitable description of molecular structures and ends with some inference, hypothesis, and prediction on the behavior of molecules in environmental, biological, and physico-chemical systems in analysis.

QSARs are based on the assumption that the structure of a molecule (for example, its geometric, steric, and electronic properties) must contain features responsible for its physical, chemical, and biological properties and on the ability to capture these features into one or more numerical descriptors. By QSAR models, the biological activity (or property, reactivity, etc.) of a new designed or untested chemical can be inferred from the molecular structure of similar compounds whose activities (properties, reactivities, etc.) have already been assessed.

The development of QSAR/QSPR models is a quite complex process.

Once the research goal has been clearly defined, which in most cases means defining the property to be modeled, that is, the end point, the decision to be made concerns how much general the final model should be. This entails the selection of the set of molecules the modeling procedure is applied to. For a long time, QSAR models were developed on sets of congeneric compounds, that is, molecules with a common parental structure and different substituent groups. Later, the interest in producing tools for quick molecular property estimations moved forward more general QSAR models suitable for diverse molecules belonging to different chemical classes, that is, not congeneric sets. The final decision in defining the molecule set mainly depends on the foreseen use of the model and availability of experimental data.

In this phase of the QSAR process, it is of primary concern to gain an exhaustive knowledge about the compounds in analysis with specific regard to the end point of interest. This obviously implies acquisition of reliable experimental data regarding the end point and possibly already existing models. Data of the chemicals can be produced experimentally or retrieved from literature. In both cases, accuracy should be carefully evaluated: the limiting factor in the development of QSAR/QSPR models is the availability of high-quality experimental data, since the accuracy of the property estimated by a model cannot exceed the degree of accuracy of the input data. Moreover, when data are collected from literature to avoid an additional variability



into the data due to different sources of information, data should be taken just from one source or from almost comparable sources.

Another important phase of the QSAR process is the definition of a reliable chemical space or, in other words, the selection of those structural features thought to be the most responsible for modeling the end point in analysis. This implies the selection of proper molecular descriptors but, in most cases, there is no *a priori* knowledge about which molecular descriptors are the best. Then, the tendency is to use a huge number of descriptors, which hopefully include the candidate variables for modeling and later apply a variable selection technique. Two basic strategies can be adopted: (a) the use of algorithms to select the optimal subset(s) of descriptors and (b) the use of chemometric methods (e.g., PCA or PLS) able to condense the large amount of available chemical information into a few principal variables. Before starting to generate quantitative models, relationships between structure and activity of molecules can be qualitatively evaluated by the aid of indices such as SAL index and SAR index, specifically conceived to measure the degree of roughness of the activity landscape in the selected chemical space [Maggiora, 2006]. If there are a number of cliffs, that is, discontinuities, then there are some options available: the chemical space can be changed by selecting a different set of molecular descriptors; nonlinear models can be used instead of the most common linear ones; more compounds need to be sampled in the most discontinuous regions.

Exploratory data analysis is a common preliminary step in all the QSAR/QSPR studies. In particular, Principal Component Analysis (PCA) and clustering methods