Advances in Mathematical Chemistry and Applications

Advances in Mathematical Chemistry and Applications

Volume 1 (Revised Edition)

Edited By

Subhash C. Basak

International Society of Mathematical Chemistry
1802 Stanford Avenue, Duluth
MN 55811 and UMD-NRRI
5013 Miller Trunk Highway
Duluth MN 55811
USA

Guillermo Restrepo

Laboratorio de Química Teórica
Universidad de Pamplona
km 1 vía Bucaramanga
Pamplona, Norte de Santander
Colombia

&

José L. Villaveces

Universidad de los Andes
Carrera 1 No 18A-12
Bogotá, D. C.
Colombia

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

ELSEVIER

Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK
225 Wyman Street, Waltham, MA 02451, USA

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-1-68108-198-4

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

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

Cover Art

The cover represents an Erlenmeyer flask made from symbols of chemistry and mathematics of almost 2000 years. The symbols are organized in a chronological order starting with the Platonic solids at the bottom. Two equations at the top, the Schrödinger equation and the Wiener index equation, represent a balance between continuous and discrete mathematics used in current mathematical chemistry. The mosquito at the mouth of the flask and the chiral mosquito repellent represent practical applications of mathematical chemistry. The cover was designed by Guillermo Restrepo and Subhash C. Basak.

FOREWORD

To the edifice of Mathematical Chemistry, a new brick is being added by the present book, edited by S. C. Basak, G. Restrepo and J. L. Villaveces. During the last three decades, Dr. Subhash C. Basak's (the "apostle to USA and India") persistent efforts have led to the organization of eleven international symposia centered on Mathematical Chemistry and held either at the University of Minnesota Duluth- Natural Resources Research Institute, or at various locations in India. The second editor, Dr. Guillermo Restrepo, is the "apostle to Latin America", who, in collaboration with Drs. Basak and Villaveces, organized two recent mathematical chemistry symposia in Colombia; he co-authored two chapters in this book: one in Vol. 2 deals with similarity in molecular structure reflected in similarity of chemical reactions and then in similarity of reaction networks; the other chapter in the present Vol. 1 presents a comparison between statistical methods for analyzing physical and chemical features determining how chemical elements combine into substances.

An important feature is the fact that from the 27 chapters of the two volumes, seven have been written by the scientists who initiated the research in the respective field. Thus, Professors A. Kerber and C. Rücker with several collaborators describe their latest version of the computer program MOLGEN 5.0 for molecular structure generation. Dr. A. Nandy reviews the beginnings and present status of graphical representations for DNA, RNA, and protein sequences – the very essence of life on our planet. Professor D. Bonchev's overall topological representation of molecular structure is the topic of an interesting chapter; the newly developed Bourgas indices, which are real numbers, offer a promise as discriminating molecular descriptors for measuring graph complexity and centrality. Molecular topology is also the topic of a chapter by J. Galvez and his collaborators, which provides a pedagogical approach to the development and use of topological indices for drug design. N. Trinajstić with two coworkers present for acyclic graphs the matrices and derived topological indices that result from summing or multiplying local graph invariants (vertices or edges). P. Willett and two coworkers review similarity-based virtual screening of molecules for bioactivity based on weighted two-dimensional fingerprint fragments. Last but not least, S. C. Basak's chapters discuss (1) the factors that have led to the rapid development of discrete mathematical applications in chemistry during the last few decades; one of these factors has been the development of hardware and software allowing the exploration of large chemical databases for understanding the structural basis of physical and biochemical properties, enabling computer-aided drug design to become an indispensable tool of the pharmaceutical industry; and (2) the molecular descriptors (especially topological indices) as tools for hierarchical QSAR modeling (topostructural, topochemical, geometrical/chiral, and 3D-descrriptors); in turn, quantum chemical computational methods – semiempirical followed by *ab-initio* – have their hierarchy, first ignoring and then taking into account the solvent.

Among topics dealing with biomedical applications, mention should be made of chapters describing: (i) computational methods (molecular docking and dynamics) for the molecular design of substances that inhibit sensing systems; (ii) pharmacophore models for repellants and biocides against insects or protozoa; (iii) factors influencing protein folding and how to control them; (iv) for the more restricted class of proteins that are metalloenzymes, critical evaluations of quantum-chemical methods for explaining the catalytic activity; (v) computer-aided drug design for antitubercular compounds based on structural descriptors; (vi) for an analogous purpose, various QSAR models exemplified by five toxicological studies using the program CAESAR; (vii) QSAR modeling of toxicity for marine algae; (viii) drug-likeness evaluated by comparison with known drug databases and databases for bioactive molecules that are not drugs.

Finally, the reader will also find interesting chapters on (i) topological ranking of fullerene stability; (ii) molecular descriptors with high discriminating ability, i. e. low degeneracy; (iii) the periodicity of di-, tri-, and tetra-atomic molecules; (iv) molecular taxonomy, extended to various types of elementary particles, not only atoms; (v) statistical methodology to be employed in QSAR/QSPR when the number of properties exceeds the number of structures;(vi) so-called comparability graphs for analyzing molecular graphs and network data; (vii) using point set topology for chemical and biochemical; applications; (viii) employing conceptual density functional theory for a deeper understanding of chemical reactivity.

One should congratulate the editors for having persuaded 68 scientists from 15 countries (Austria, China, Colombia, Croatia, Denmark, Germany, India, Iran, Italy, Malaysia, Slovenia, Spain, Turkey, United Kingdom, USA) to write the 27 chapters of these two volumes, and to coordinate their contributions.

Students, professors, and anyone interested in chemical or biomedical research based on discrete applied mathematics will profit from reading this book.

Alexandru T. Balaban
Emeritus Professor
Texas A&M University at Galveston
USA

PREFACE

The Universe is a grand book which cannot be read until one first learns to comprehend the language and become familiar with the characters in which it is composed. It is written in the language of mathematics.
Galileo Galilei

I dive down into the depth of the ocean of forms,
hoping to gain the perfect pearl of the formless.
Rabindranath Tagore

The perfection of chemistry might be secured and hastened by the training of the minds of chemists in the mathematical spirit [...]. Besides that mathematical study is the necessary foundation of all positive science, it has a special use in chemistry in disciplining the mind to a wise severity in the conduct of analysis: and daily observation shows the evil effects of its absence.
Auguste Comte

In this eBook we introduce our readers to one of the most comprehensive and thematically diverse treatise on the emerging discipline of *mathematical chemistry*, or, more accurately, *discrete mathematical chemistry*. Although mathematical representation and characterization of chemical objects were known for a long time, the incursion of discrete mathematics into chemistry had a tremendous growth spurt in the second half of the twentieth century and the trend is continuing even today in an unabated manner. We think such a growth has been fueled and sustained primarily by two factors: i) Novel applications of discrete mathematics to chemical and biological systems, and ii) Availability of high speed computers and associated software whereby *hypothesis driven* as well as *discovery oriented* research can be carried out within a reasonable time frame. This trend of research has led not only to the development of many novel concepts, but also to numerous useful applications to scientifically, socially and economically important areas such as drug discovery, protection of human as well as ecological health, chemoinformatics, bioinformatics, toxicoinformatics, and computational biology, to name just a few. This book is a clear depiction of those concepts and applications.

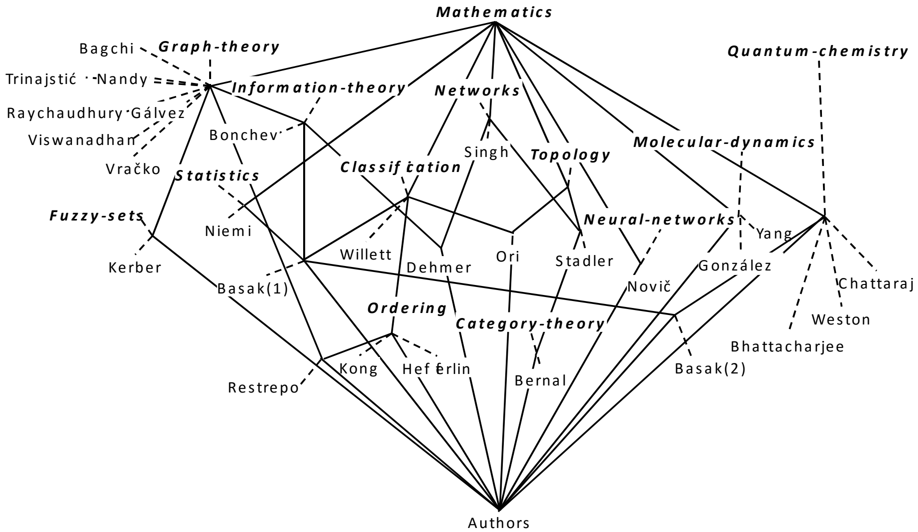
Another perspective of Mathematical Chemistry is the very mathematics-chemistry relationship, which also shows its growing community. If we look at journals devoted exclusively to the link between mathematics and particular natural sciences, then the *Journal of Mathematical Physics* has to be mentioned first, for it appeared for the first time in 1960; at that time, both chemistry and biology lagged behind physics. Fourteen years later, the *Journal of Mathematical Biology* was launched and one year later the first journal for mathematical chemistry showed up: *MATCH Communications in Mathematical and in Computer Chemistry*. Later, in 1987, the *Journal of Mathematical Chemistry* was initiated and just recently (2010) the *Iranian Journal of Mathematical Chemistry* published its first issue. The reasons for such a delay, in contrast with physics, have recently been a matter of discussion among philosophers of science, particularly of chemistry, where *HYLE - International Journal for Philosophy of Chemistry* has played a central role. The fact of being the last of the three sciences in launching a scientific journal devoted to its relationship with mathematics contrasts with the three journals specifically devoted to such a link. This suggests the growth of the community, where a single journal is not able to cope with the amount of novel results in the area of mathematical chemistry.

A seminal piece of research in modern mathematical chemistry was the path breaking work of Harry Wiener (1947), which stimulated a wealth of investigations on applications of discrete mathematics in chemistry, e.g., graph theory, matrix theory, and information theory. In the early 1980s, Professors R. Bruce King and Dennis H. Rouvray accelerated this process by the initiation of the *International Conference on Mathematical Chemistry* series at the University of Georgia at Athens, Georgia, USA. The conferences under the leadership of King and Rouvray were organized in North America and Europe during 1983-2005 and the conference proceedings attest to the high scientific standard of discourse in those events, where not only the mathematical theories aforementioned found a fertile land but also topology and group theory, to name but a few more. These conferences led to numerous discussions on the organization of the community of people delving into the wonders of the chemomathematical relationship. Hence, the *International Society of Mathematical Chemistry* (ISMC) was founded in 1987 with Professor Milan Randić being its President till 2003; thereafter Subhash C. Basak took over the presidency of ISMC from Professor Randić. An important meeting point of members of the community is the *MATH/CHEM/COMP* symposium, traditionally organized in Dubrovnik (Croatia) for more than 25 years. This yearly series seeks to foster the exchange of ideas among chemists, mathematicians, and computer scientists; there is no doubt of the importance this meeting has for the mathematical chemistry community, which owes much to Professor Ante Graovac, who recently passed away (2012), and who was always behind the organization of the MATH/CHEM/COMP meetings. As part of the growing process, mathematical chemists looked for another type of organization, this time an academy, namely the *International Academy of Mathematical Chemistry*[1] (IAMC), founded in 2005, which also organizes a yearly symposium. The organizational seed of the Balkans soon crossed the oceans and went to India and USA. In the 1990s, Subhash C. Basak and Dilip K. Sinha initiated the *Indo-US Workshop on Mathematical Chemistry* series[2], whose aim was to bring both senior scientists and young scholars to a single and homely forum to discuss the advancing frontiers of mathematical chemistry and allied sciences. In 2007, with the objective of bringing the young scholars in close contact with the internationally renowned experts for exclusive mentorship and training, the *Indo-US Lecture Series on Discrete Mathematical Chemistry*[3] was established. More recently, this enthusiasm for mathematical chemistry infected South America and led to the creation of the *Mathematical Chemistry Workshop of the Americas*,[4] involving countries of North and South Americas.

Besides the two aforementioned initial journals dedicated to Mathematical Chemistry, it is worth mentioning others that have facilitated the circulation of chemomathematical knowledge and its manifold applications: *Journal of Chemical Information and Computer Sciences* (which gave place to the *Journal of Chemical Information and Modeling* in 2005), *SAR & QSAR in Environmental Research*, *Croatica Chemica Acta*, *Journal of Molecular Graphics & Modelling*, *Journal of Molecular Structure – Theochem*, *Current Computer – Aided Drug Design*, *QSAR & Combinatorial Science*, and the *Journal of Computational Chemistry*, to name but a few. Besides specialized books on particular subjects of Mathematical Chemistry, several books have also been published collecting the chemomathematical knowledge of their times, where Dennis H. Rouvray has played a central role as editor; Professors Balaban, Bonchev, Kier, Hall, Trinajstic as well as King have also made outstanding contributions in their books.

[1] http://www.iamc-online.org/index.htm
[2] http://www.nrri.umn.edu/indousworkshop
[3] www.nrri.umn.edu/indouslecture
[4] http://sites.google.com/site/mathchemamericas/

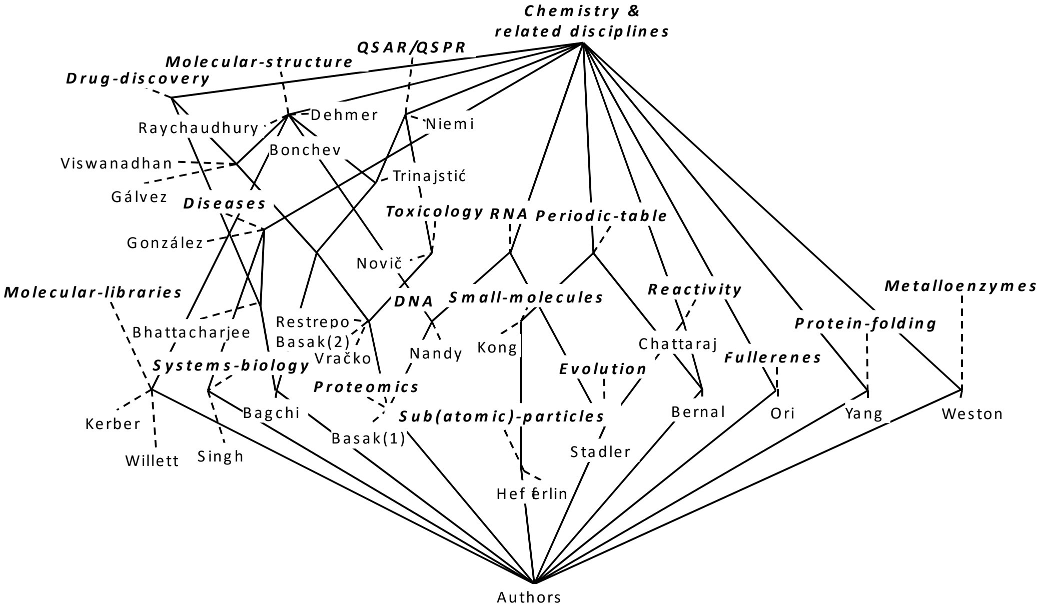
At such an astonishing historical moment of a scientific community that has been able to organize, create, and use the needed intellectual fermentation and communication channels to keep growing, this book comes into play. The 27 chapters of the current book are derived from multiple sources: i) Papers presented at the *Second Mathematical Chemistry Workshop of the Americas* in Bogota, Colombia, in 2011; ii) Papers presented at the *First Indo-US Lecture Series on Discrete Mathematical Chemistry* in Bangalore, India (2007), and iii) Invited chapters from distinguished researchers in Mathematical Chemistry and related areas. These chapters deal both with the development and history of the basic concepts as well as their applications. They also show the fruitful relationship between different branches of mathematics and chemistry and related disciplines. The branches of mathematics considered are graph, information and category theories, as well as statistics, fuzzy sets, network analysis, classification techniques, ordering, topology, neural networks and mathematical aspects of molecular dynamics and quantum chemistry. Due to the large number and scientific diversity of the chapters in the current book, an attempt to summarize the aforementioned branches of mathematics interacting with chemistry goes beyond our capabilities in the limited space of this preface. What we have done, instead, is to use mathematics to guide the reader through the multiple paths this book offers, which in the end constitutes a book of several (finite) sub-books. In this sense, this book, besides containing the current status of Mathematical Chemistry, is also a *Rayuela*[5], a combinatorial book like the one by the famous writer Julio Cortazar. We show in the following figure a map, based on order theory, to explore the book depending on the particular interests of our readers. Each node in the graph contains two kinds of information: a set of mathematical branches (italics) and a set of authors; for the sake of simplicity we have labeled each chapter by the surname of its corresponding author. Once a node is selected, all those nodes found in a downward path give information about the node selected, e.g. if we are interested in "Information theory", then the respective node shows that Bonchev's chapter as well as Basak(1)'s, and Dehmer's chapters are related to that mathematical branch. Likewise, if we are interested in "Topology", then the authors to read are Ori, Stadler, and Bernal; the latter two being also related to "Category theory".



[5] The book is known in English as *Hopscotch*.

As the book also shows the result of the incursion of the above mentioned branches of mathematics with chemistry and other disciplines, then we drew another diagram depicting how authors' chapters are related to several areas of knowledge. For example, we see that the chapters relating mathematics with RNA are those by Nandy, Basak(1), and Stadler. The areas chemical and related disciplines considered by the authors are: Drug discovery processes, molecular structure characterizations, QSAR/QSPR models related studies to tackle several diseases, models for toxicology, RNA studies, periodic tables, algorithms for exploring molecular libraries, systems biology, proteomics, DNA characterizations, studies of small molecules, biological evolution, chemical reactivity, protein folding studies, fullerenes' and metalloenzymes characterizations.

Both diagrams can also be read in the upward direction, where the information extracted is on the topics tackled by the authors. Hence, if we take, e.g. Viswanadhan, we see that his chapter is on Drug discovery and Molecular structure, combined with Graph theory.



Which kinds of readers do we expect for our book? The book is useful to the uninitiated with some grounding in chemistry, mathematics and biology, e.g., senior undergraduate students; graduate students / postdocs as well as senior researchers who wish to get started as new investigators in the field.

To conclude, we would like to take the opportunity of thanking all those who have assisted in any way with the realization of this project. Certainly included here are all the authors who have contributed with their important chapters; Professors Balaban and Kier, leading Mathematical Chemistry figures of our time, who kindly wrote their inspiring forewords for our book; Professor Esperanza Paredes, Dean of the Universidad de Pamplona (Colombia) and Professors René Meziat and Wolfram Baumann, heads of the Departments of Mathematics and Chemistry, respectively, of the Universidad de los Andes (Colombia), who gave us the financial support to have the *Second Mathematical Chemistry Workshop of the Americas* organized in Colombia, and Wilmer Leal, who formatted the chapters according to Bentham Science Publishers guidelines.

Subhash C. Basak, being already involved in the organization of a total thirteen events in the three mathematical chemistry workshop series mentioned above, has been immensely helped by numerous colleagues, friends, and collaborators (*called members of his virtual team*) the long list of whom cannot be mentioned here for brevity. Help extended to Basak by Dilip K. Sinha (former Vice-Chancellor of Visva Bharati University, Santiniketan, India), Michael J. Lalich (Former director, University of Minnesota Duluth- Natural Resources Research Institute, UMD-NRRI, USA), Ashok Kolaskar (former Vice Chancellor, University of Pune, India); Brian Gute, Denise Mills, Gerald Niemi, Donald Harriss, Vincent Magnuson, Gregory Grunwald from UMD/ NRRI; Kanika Basak, Sarat Basak, Moumita Basak, Nabamita Basak; Indira Ghosh, Uma Vuruputuri, Manish Bagchi, Ramanathan Natarajan, R. Balakrishnan, S. Parthasarathy, Subhendu Gupta, Ashesh Nandy, Vellarkad Viswanadhan, Chandan Raychaudhury, Marimuthu Ramalingam, P. Venuvanalingam, Tarun Jha, Mohanraj Subramanian, M. Srinivasan, from India; Gilman Veith, Milan Randić, Krishnan Balasubramanian, Moiz Mumtaz, Apurba Bhattacharjee, Kevin Geiss, Frank Witzmann, Chandrika Moudgal, George Vacek, from USA; Kannan Krishnan, Shahul Nilar (Canada); Marjan Vračko, Marjana Novič from Slovenia; Vladimir Palyulin, Nikolay Zefirov from Russia; Rainer Brüggemann (Germany) and Haruo Hosoya (Japan), are gratefully acknowledged.

We are thankful to the staff of Bentham Science Publishers, Ms. Asma Ahmed in particular, for the untiring efforts in all aspects of the publication of this book.

We sincerely hope that the two volumes of the eBook *Advances in Mathematical Chemistry* will not only apprise its readers of the advancing frontiers of mathematical chemistry along with its wide variety of applications, but also will stimulate further research in the field both by young scholars and senior researchers.

Subhash C. Basak
International Society of Mathematical Chemistry
University of Minnesota
USA

Guillermo Restrepo
Universidad de Pamplona
Colombia

&

José L. Villaveces
Universidad de los Andes
Colombia

CONTRIBUTORS

Shereena M. Arif
Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK; Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia

Subhash C. Basak
International Society of Mathematical Chemistry, 1802 Stanford Avenue, Duluth, MN 55811 and UMD-NRRI, 5013 Miller Trunk Highway, Duluth MN 55811, USA

Apurba Bhattacharjee
Department of Medicinal Chemistry, Division of Experimental Therapeutics, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, MD 20910-7500, USA

Danail Bonchev
Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284-2030, USA

Pratim K. Chattaraj
Department of Chemistry and Center for Theoretical Studies, Indian Institute of Technology, Kharagpur 721302, India

Matthias Dehmer
Institute of Bioinformatics and Translational Research, UMIT, A-6060, Hall in Tyrol, Austria

Jorge Gálvez
Molecular Connectivity and Drug Design Research Unit, Faculty of Pharmacy, Department of Physical Chemistry, University of Valencia Avd, V.A. Estellés, s/n 46100-Burjassot, Valencia, Spain

María Gálvez-Llompart
Molecular Connectivity and Drug Design Research Unit, Faculty of Pharmacy, Department of Physical Chemistry, University of Valencia Avd, V.A. Estellés, s/n 46100-Burjassot, Valencia, Spain

Ramón García-Domenech
Molecular Connectivity and Drug Design Research Unit, Faculty of Pharmacy, Department of Physical Chemistry, University of Valencia Avd, V.A. Estellés, s/n 46100-Burjassot, Valencia, Spain

Ralf Gugisch
Department of Mathematics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

Ray Hefferlin
Physics Department, Southern Adventist University, Collegedale, Tennessee 37315, USA

John D. Holliday
Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK

Adalbert Kerber
Department of Mathematics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

Axel Kohnert
Department of Mathematics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

Reinhard Laue
Department of Mathematics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

Bono Lučić
The Ruđer Bošković Institute, P.O.Box 180, HR-10 002 Zagreb, Croatia

Subhabrata Majumdar
School of Statistics, University of Minnesota Twin Cities, 224 Church Street SE, Minneapolis, MN 55455, USA

Markus Meringer
Department of Atmospheric Processors, German Aerospace Center (DLR), Oberpfaffenhofen, Münchner Straße 20, 82234 Wessling, Germany

Lakshminarasimhan Rajagopalan
Department of Computational Chemistry, Jubilant Biosys Limited, Bangalore 560 022, India

Hariharan Rajesh
Department of Computational Chemistry, Jubilant Biosys Limited, Bangalore 560 022, India; Shanmugha Arts, Science, Technology, and Research Academy, Thanjavur 613 402, TN, India

Guillermo Restrepo
Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia

Debesh R. Roy
Department of Chemistry and Center for Theoretical Studies, Indian Institute of Technology, Kharagpur 721302, India; Department of Applied Physics, S. V. National Institute of Technology, Surat 395007, India

Christoph Rücker
Institute of Sustainable and Environmental Chemistry, Leuphana University Lüneburg, Scharnhorststraße 1, 21335 Lüneburg, Germany

Lavanya Sivakumar
Institute of Bioinformatics and Translational Research, UMIT, A-6060, Hall in Tyrol, Austria

Ivan Sović
The Ruđer Bošković Institute, P.O.Box 180, HR-10 002 Zagreb, Croatia

Nenad Trinajstić
The Ruđer Bošković Institute, P.O.Box 180, HR-10 002 Zagreb, Croatia

Vellarkad N. Viswanadhan
Department of Computational Chemistry, Jubilant Biosys Limited, Bangalore 560 022, India

Marjan Vračko
Kemijski inštitut/National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

Alfred Wassermann
Department of Mathematics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany

Peter Willett
Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK

ACKNOWLEDGEMENTS

The editors of the eBook "Advances in Mathematical Chemistry and Applications" and Bentham Science Publishers would like to gratefully acknowledge the dedicated work of the following distinguished scientists who reviewed the chapters of this book and gave valuable suggestions as well as constructive criticisms to the authors and editors that substantially enhanced the quality of this volume.

REVIEWERS OF EBOOK CHAPTERS

Subhash C. Basak

International Society of Mathematical Chemistry, 1802 Stanford Avenue, Duluth, MN 55811 and UMD-NRRI, 5013 Miller Trunk Highway, Duluth MN 55811, USA

Rajarshi Guha

National Center for Advancing Translational Science, NIH, 208 Saddle Hill Road, Manchester CT 06040, USA

Haruo Hosoya

Ochanomizu University, Bunkyo-ku, Tokyo 112-8610, Japan

Lemont B. Kier

Life Sciences, Virginia Commonwealth University, 821 West Franklin Street, Richmond, VA 23284, USA

Douglas J. Klein

Department of Marine Sciences, Texas A&M University, Galveston TX, USA

Fanao Kong

The Institute of Chemistry, Chinese Academy of Sciences, Beijing, 100190, China

Wilmer Leal

Fundación Instituto de Inmunología de Colombia FIDIC, Bogotá, Colombia, Universidad del Rosario, Bogotá, Colombia; Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia

Yenamandra S. Prabhakar

Medicinal and Process Chemistry Division, Central Drug Research Institute, CSIR, Lucknow-226001, India

Milan Randić

Kemijski inštitut/National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

Guillermo Restrepo

Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia

Roberto Todeschini

Department of Environmental Sciences, University of Milano-Bicocca, P.za della Scienza, 1 - 20126 Milano, Italy

Marjan Vračko

Kemijski inštitut/National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

CHAPTER 1

Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications

Subhash C. Basak*

International Society of Mathematical Chemistry, 1802 Stanford Avenue, Duluth, MN 55811 and UMD-NRRI, 5013 Miller Trunk Highway, Duluth MN 55811, USA

Abstract: Mathematical chemistry or more accurately discrete mathematical chemistry had a tremendous growth spurt in the second half of the twentieth century and the same trend is continuing now. This growth was fueled primarily by two major factors: 1) Novel applications of discrete mathematical concepts to chemical and biological systems, and 2) Availability of high speed computers and associated software whereby *hypothesis driven* as well as *discovery oriented* research on large data sets could be carried out in a timely manner. This led to the development of not only a plethora of new concepts, but also to various useful applications to such important areas as drug discovery, protection of human as well as ecological health, and chemoinformatics. Following the completion of the Human Genome Project in 2003, discrete mathematical methods were applied to the "omics" data to develop descriptors relevant to bioinformatics, toxicoinformatics, and computational biology. This chapter will discuss the major milestones in the development of concepts of mathematical chemistry, mathematical proteomics as well as their important applications in chemobioinformatics with special reference to the contributions of Basak and coworkers.

Keywords: Graph theory, molecular graphs, networks, graph invariant, weighted pseudograph, graph theoretic matrices, adjacency matrix, distance matrix, topological indices, information theoretic indices, Wiener index, Hosoya index, Balaban index, connectivity indices, valence connectivity indices, E-state indices, mathematical chemodescriptor, quantum chemical descriptors, hierarchical quantitative structure-activity relationship (HiQSAR), differential QSAR (DiffQSAR), partial least square (PLS), principal components regression (PCR), ridge regression (RR), naïve q^2, true q^2, proper cross validation, leave one out (LOO) method, quantitative molecular similarity analysis (QMSA), tailored similarity, combinatorial chemistry, clustering, analog selection, mode of action

***Corresponding author Subhash C. Basak:** International Society of Mathematical Chemistry, 1802 Stanford Avenue, Duluth, MN 55811, USA; Tel: 1-218-727-1335; Fax: 1-218-720-4238; E-mail: sbasak@nrri.umn.edu

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50001-1

(MOA), new drug discovery, environmental protection, prediction of property/bioactivity, predictive toxicology, mutagenicity, 2-D gel electrophoresis, biodescriptor, mathematical proteomics, spectrum like descriptors, information theoretic proteomics descriptors, DNA sequence descriptor.

INTRODUCTION

> *"No human inquiry can be a science unless it pursues its path through mathematical exposition and demonstration"*
>
> *Leonardo da Vinci*

A contemporary trend in quantitative structure-activity relationships (QSARs), new drug discovery, and computational toxicology is the prediction of properties of chemicals from their structural descriptors [1-10]. This is conveniently expressed by the following equation:

$$P = f(S) \tag{1}$$

where P is any physical, biological, medicinal or toxicological property of a chemical and S symbolizes the subset of its structural features related to the property under investigation.

A perusal of recent published literature would show that various classes of calculated properties, *viz*., topological, geometrical, quantum chemical, substructural, are used routinely in predicting properties of interest. This field, quantitative structure-activity relationship (QSAR), had its modest beginning at the second half of the nineteenth century. In 1968, Crum-Brown and Fraser [11] reported that the structure of quaternary compounds was responsible for their "physiological activity." Later, in 1993, Richet [12] observed that the toxicological activity of diverse organic chemicals was inversely related to their water solubility.

In 1964, Hansch and Fujita [13] formulated the linear free energy related (LFER) method of QSAR combining hydrophobicity of molecules with their electronic [14] and steric [15] parameters derived from physical organic chemistry into a multiparameter correlation approach. The linear solvation energy related (LSER) technique [16] also is in line with the LFER methodology.

The common theme among the Richet's rule and the LFER as well as LSER methods is that these are property-property relationships (PPRs), *i.e.*, physicochemical properties of chemicals are used to predict their physical or biological properties. Such PPR or PAR (property-activity relationship) methods worked well in estimating toxicity and biological activity of chemicals which are congeneric.

But both in drug design and prediction of toxicity of chemicals, however, we have to deal with structurally diverse (non-congeneric) chemicals. In many cases, physicochemical properties of most of the chemicals under investigation are not available. Currently, there are some methods for the calculation of hydrophobicity (logP, octanol/ water) of molecules from their structure. But that is not the case for many other properties. For example, the current list of industrial chemicals of the United States Environmental Protection Agency (USEPA), the well-known Toxic substances Control Act (TSCA) Inventory, has over 85,000 substances [17]. The majority of these chemicals have no physicochemical or useful toxicity data [18].

Also, modern drug discovery protocols that use high throughput screening (HTS) and combinatorial chemistry require fast screening of large and structurally diverse chemical databases which do not have many experimental property data. PPR techniques like the LFER and LSER methods have limited usefulness in such cases. A practical approach of addressing this quagmire is to use properties that can be derived algorithmically from molecular structure only. Graph theoretical invariants, substructures as well as geometrical (3-D) and quantum chemical indices fall in this category of properties [2-10]. One problem with quantum chemical indices is that for large sets of chemicals such descriptors could be very resource intensive [19]. Descriptors derived from molecular topology have been widely used in numerous QSAR studies [2-7, 9, 10].

MOLECULAR STRUCTURE

> *"Ostensibly there is color, ostensibly sweetness, ostensibly bitterness,*
> *but actually only atoms and the void."*

Galen of Pergamon (AD 129–c. 200/c. 216 modern-day Bergama,Turkey),
in Nature and the Greeks, Erwin Schrodinger, 1954

In structural chemistry, the term “molecular structure” does not always represent the same reality; on the contrary, it probably symbolizes a set of disjoint and non-equivalent concepts [20]. The representation of a molecule by a particular method creates a “model object” which encodes the relationship among its constituents [21, 22]. The Greeks represented different fundamental forms of matter by mathematical objects like regular polyhedra: fire by tetrahedron, earth by the cube, air by the octahedron, and water by icosahedron [23]. Different methods of abstraction from the same reality lead to the formulation of the various model objects by different practitioners in the field. This is most probably the reason behind what chemists call the “*molecular structure conundrum*” [21].

Graph Theoretical Representation of Molecules

Molecular structure is symbolized by graphs of different types, *viz*., simple graph, multigraph, pseudograph, *etc.* In a graph G = [V, E], V symbolizes the set of points and E is the binary relation on the set V [24].

In molecular graph models of chemical structure, the points represent the atoms and the bonds among the constituent atoms are symbolized by the binary relation. The points may represent either all atoms in the molecule or only the non-hydrogen atoms, in the latter case the graph being called hydrogen-suppressed graph. Basak *et al.* [25] pointed out that many different types of molecules can be represented by weighted pseudographs. In mathematical chemistry and chemical graph theory, the common practice is to use hydrogen-suppressed graphs. But hydrogen-filled graphs are preferred to the hydrogen-suppressed ones when the hydrogen atoms play an important role in the chemistry of the molecule.

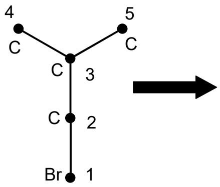
Characterization of Molecular Graphs

When molecular graphs represent molecules, graph invariants can be used to characterize them [26]. Invariants can be conveniently calculated from various graph theoretic matrices, *e.g.*, distance matrix, adjacency matrix, *etc.* Fig. (**1**) gives the labeled hydrogen suppressed graph of isopentane, its distance matrix, and shows how the Wiener index [27], W, can be calculated from the distance matrix. Hosoya [28] coined the term "topological index" for invariants of molecular graphs and showed that the topological index, W, can be calculated from the distance matrix *D*(*G*) of a molecular graph *G* as the sum of entries of the upper triangular submatrix.

Wiener Index, W

$$W = 1/2\sum_{ij} d_{ij}$$

where d_{ij} is the distance between vertices v_i and v_j in G



$$\begin{array}{c|ccccc|c} & 1 & 2 & 3 & 4 & 5 & \text{Row Sum} \\ \hline 1 & 0 & 1 & 2 & 3 & 3 & 9 \\ 2 & 1 & 0 & 1 & 2 & 2 & 6 \\ 3 & 2 & 1 & 0 & 1 & 1 & 5 \\ 4 & 3 & 2 & 1 & 0 & 2 & 8 \\ 5 & 3 & 2 & 1 & 2 & 0 & 8 \\ & & & & & & 36 \end{array}$$

$$\mathbf{W = 36 / 2}$$
$$\mathbf{= 18}$$

Figure 1: Hydrogen-suppressed graph and calculation of Wiener index for isopentane.

Graph invariants like the Wiener index (Fig. (**1**)) quantify different aspects of molecular structure like shape, size, branching, *etc.* Currently available computer software, *e.g.*, Dragon [29], MolconnZ [30], POLLY [31], APProbe [32], are capable of computing many Topological Indices (TIs) including connectivity indices [33, 34], electrotopological state indices [2], Triplet indices [35], neighborhood based information theoretic indices [36], and information theoretic indices developed by Bonchev and collaborators [37, 38].

For a detailed exposition of the use of experimental methods of property determination in the laboratory *vis-à-vis* theoretical approaches to property estimation using descriptor based QSARs, see refs. [21, 39].

STATISTICAL METHODS FOR QSAR MODEL DEVELOPMENT

As indicated above, many descriptors can be calculated today using software, but the number of data points to be modeled is often much smaller than the number of molecular descriptors. In such rank deficient situations, one has to use robust methods of statistical model building [40]. For scientifically correct approaches to model building, cross validation, and descriptor thinning, see refs. [40-42]. Some good examples of QSAR using both chemodescriptors and biodescriptors are available in refs. [43, 44].

Basak *et al.* carried out hierarchical QSAR [45-52] of various data sets of physical, biomedical, and toxicological properties using topostructural, topochemical, geometrical, and quantum chemical descriptors (Table **1**). Results show that the addiction of quantum chemical indices makes very little or no difference in model quality after the use of TIs in model building.

Table 1: Results of HiQSAR using topostructural, topochemical, 3-D and quantum chemical indices

Description of Data Set and Property/ Activity	**Model Quality Enhancement by the inclusion of Quantum Chemical Descriptors**	**Refs.**
Acute toxicity of benzene derivatives	None	[45]
Dermal penetration of polycyclic aromatic hydrocarbons (PAHs)	None	[46]
Mutagenicity of amines (heteroaromatic and aromatic)	None	[47]
Mutagenicity/non-mutagenicity of 508 diverse compounds	None	[48]
Cellular toxicity of halocarbons	minimal	[49]
Mosquito repellency of aminoamides	None	[50]
Blood: air and tissue: air partition coefficients for rat and human	None	[51]
Aryl hydrocarbon receptor binding affinity of dibenzofurans	None	[52]

Calculated topological indices and substructures of molecular graphs can be powerful tools for new drug discovery as shown in Fig. (**2**).

For large virtual or real chemical libraries TIs can be used to create descriptor spaces quite fast and to cluster large data sets for the management of explosive data situation [53]. In the area of computational toxicology, calculated descriptors can be used in predicting property and toxicity endpoints related to the hazard assessment as well as prediction of modes of action (MOAs) of pollutants from their descriptors [10, 19, 21, 34, 36, 44-49, 51, 54].

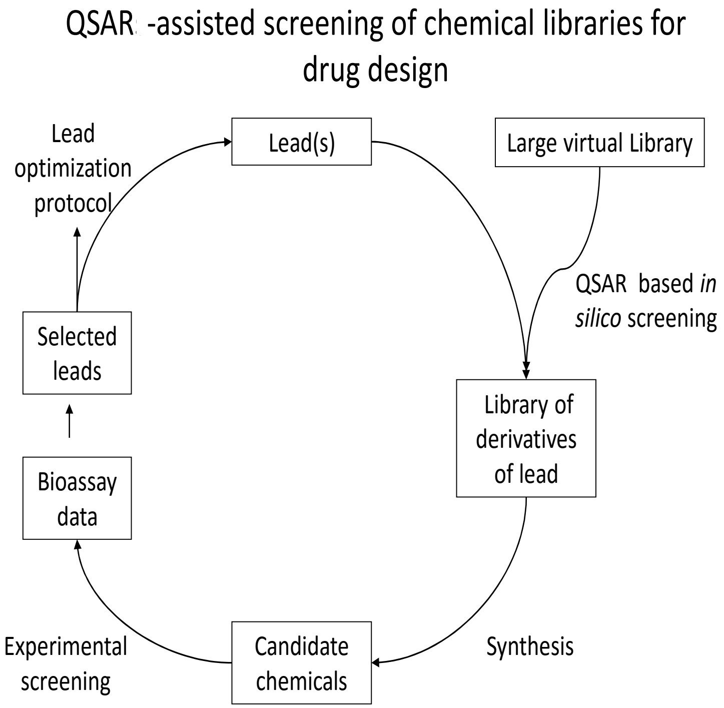

QSAR -assisted screening of chemical libraries for drug design
Lead(s)
Large virtual Library
Lead optimization protocol
QSAR based *in silico* screening
Selected leads
Library of derivatives of lead
Bioassay data
Experimental screening
Candidate chemicals
Synthesis

Figure 2: QSAR-assisted new drug discovery protocol.

DIFFERENTIAL QSAR TO CHARACTERIZE MOLECULAR BASIS OF DRUG RESISTANCE

In the development of drug resistance, *e.g.*, the emergence of drug resistant malaria parasites [55], the target undergoes alterations because of exposure to the drug. QSAR developed for a group of 58 cycloguanil derivatives using calculated mathematical descriptors showed that only a couple of influential descriptors were common between the models for the dihydrofolate reductases (DHFRs) from sensitive and resistant *Plasmodium falciparum* [56]. Differential QSAR of this type can help in understanding the mode of alterations in ligand-target interactions involved in resistance development. Basak *et al.* [57] also used this novel method in the analysis of bioassay data for five different varieties of DHFRs, one from the wild and four from resistant mutant varieties of the malaria parasite.

SIMILARITY: BIRDS (AND CHEMICALS!) OF A FEATHER FLOCK TOGETHER

"HAMLET: Do you see yonder cloud that's almost in the shape of a camel?

POLONIUS: By th' mass and 'tis like a camel indeed.

HAMLET: Methinks it is like a weasel.

POLONIUS: It is backed like a weasel.

HAMLET: Or like a Whale

POLONIUS: Very like a Whale

William Shakespeare

Toxicologists and pharmaceutical chemists use the concept of similarity widely [58]. When a promising drug candidate is discovered, the drug designer wants to know whether its analogs have similar biological properties. One method is to search databases like the Chemical Abstract Service (CAS) database containing approximately 71 million substances [59]. Such similarity search can be done efficiently using similarity methods based on molecular descriptors which can be computed fast.

Most industrial chemicals in USEPA's TSCA Inventory do not have experimental property/toxicity data necessary for their risk assessment [18]. USEPA uses QSAR derived from specific class of chemicals or the properties of selected analogs to carry out hazard assessment. One first looks for QSAR for the class which contains the chemical under investigation. If this does not work out, molecular similarity is used based on the notion that similar structures usually have similar properties [39, 60-62]. For example, if two chemicals have the same pharmacophore, many of their physical and biological properties may be analogous. But there are also some exceptions to this notion; *e.g.* benzene is a known carcinogen whereas toluene, with a structure very similar to that of benzene, is non-carcinogenic. Bioisosteric chemicals have similar biological targets although they have very little apparent similarity in structure [63]. That one can choose and derive mutually different molecular similarity methods starting from the same set of computed indices makes the work of similarity scientists a difficult one [61, 64].

It was noted by Basak *et al.* that similarity relation is a tolerance relation [65, 66]. In selecting analogs for estimating property, because of the nature of the tolerance relation, the structure of the analogs become progressively dissimilar as we go further and further from the query molecule. Therefore, the researcher must be on guard about the utility of the selected analogs using the k-nearest neighbor (KNN) approach. Please see Basak *et al.* [58] for further information on this topic. For the estimation of specific property of interest using quantitative molecular similarity analysis (QMSA) techniques more effectively, Basak *et al.* [58] developed the method called the tailored QMSA (t-QMSA). Most QMSA methods mentioned previously have been called arbitrary or user-defined QMSA by Basak *et al.* [58, 67-69]. The tailored similarity method, however, develops structure spaces for analog selection considering the property under investigation.

MATHEMATICAL DESCRIPTORS OF NUCLEIC ACID SEQUENCES

"If your chromosomes are XYY,

And you are a naughty, naughty guy,

Your crimes, the judge won't even try,

'Cause you have a legal reason why

He'll raise his hands and gently sigh!

"I guess for this you get a bye."

By Carl A. Dragstedt

In: Perspectives in Biology and Medicine, Vol. 14, # 1, autumn, 1970

In the post-genomic era, a lot of data for DNA, RNA, and protein sequences are being generated continuously. We need methods for the characterization of sequences so that one can relate such sequences (structures) to their biological function.

In line with the representation-mathematical characterization approach discussed earlier [21, 22] in the formulation of graph invariants for molecules, representation of sequences of bases in a DNA or RNA strand using graphical methods was initiated by various authors including Hamori and Ruskin [70], Gates [71], Nandy [72] and Leong and Morgenthaler [73]. For reviews on the topic, see Nandy *et al.* [74] and Randic *et al.* [75].

In the last few years, this field had a tremendous growth spurt in terms of the number of papers published on the topic. The present author, however, feels that a brief history of the development of this exciting field beginning in 1998 is necessary here. Dilip K. Sinha and Subhash C. Basak initiated the Indo-US Workshop on Mathematical Chemistry [76] in 1998 with the first event organized at the Visva Bharati University, Santiniketan, West Bengal, India. Raychaudhury and Nandy [77] presented a paper on mathematical characterization of DNA sequences using their graphical method [72]. This caught the attention of Basak who subsequently developed a research group on the mathematical characterization of DNA/RNA sequences using funding from the University of Minnesota Duluth-Natural Resources Research Institute (UMD-NRRI) and University of Minnesota. This led to the publication of the first couple of papers on DNA sequence invariants [78, 79]. The rest of the development of DNA/RNA sequence invariants and mathematical descriptors is self evident on the pages of numerous international journals.

DESCRIPTORS FROM MATHEMATICAL PROTEOMICS

Contemporary sequencing, microarray, proteomics, and related techniques generate a lot of information on sequences as well as cellular transcription, translation, and post-translational modification processes.

The two-dimensional gel electrophoresis (2-DE) technique gives us information on the abundance, mass, and charge of proteins at a particular moment of time. It is a daunting task to manage such a huge amount of information. Basak and coworkers [80-83] developed various mathematical proteomics approaches for the quantitative characterization of proteomics maps generated by the 2-DE methods.

COMBINED USE OF CHEMODESCRIPTORS AND BIODESCRIPTORS FOR BIOACTIVITY PREDICTION

Beneficial or deleterious property (P) or biological response (BR) generated by chemicals, is the consequence of ligand-target interactions. This may be expressed by:

$$\text{P or BR} = f\,(S, B) \qquad (2)$$

where P/BR represent the observed property or bioactivity and B symbolizes the biochemical part of the target system which is perturbed by the chemical to produce the effect. The factor S solely determines BR when the nature of B is practically the same from chemical to chemical. Under such circumstances, Eq. 2 approximates to:

$$\text{BR} = f\,(S) \ldots . \qquad (3)$$

Please note that Eq. 3 is identical with Eq. 1 above, the accepted paradigm of the field of QSAR

But often chemical-biological interactions are not as simple as depicted by Eq 3 and so in many cases the chemical structure of the ligand alone cannot predict the biological action of molecules effectively. This is more evident in complex biological responses such as chemical carcinogenesis where chemical structure alone has been found to be grossly inadequate for a reasonable prediction of bioactivity of chemicals. In such cases, the experts have recommended the use of some *biological criteria along with structural criteria* in developing estimation methods for cancer risk. Arcos [85], for example, suggested the use of specific biological data, *e.g.*, degranulation of endoplasmic reticulum, peroxisome proliferation, unscheduled DNA synthesis, anti-spermatogenic activity, *etc.* as biological indicators of carcinogenesis.

Basak and coworkers [8, 9, 43, 83, 84] reasoned that in the structural-functional pair of criteria for the prediction of bioactivity, calculated chemodescriptors can represent the structural aspects and proteomics based biodescriptors mentioned above can be looked upon as functional criteria. In order to carry out this line of research, we needed substantial grant funding and collaboration of experimental proteomics research groups. Fortunately we got outstanding collaboration from Dr. Kevin Geiss of Wright Patterson Air Force Base and Dr. Frank Witzmann of Indiana University in the experimental areas and were supported in our mathematical proteomics research by the following US Air Force grants awarded to Subhash C. Basak as the principal investigator:

1) Integration of biodescriptors and chemodescriptors for predictive toxicology: A mathematical/computational approach, US Air Force Office of Scientific Research, 11/2000 – 10/2001.

2) Use of biodescriptors and chemodescriptors in predictive toxicology: A mathematical/computational approach, US Air Force Office of Scientific Research, 3/2002 – 2/2005.

3) Predicting chemical toxicity from proteomics and computational chemistry: An integrated approach, US Air Force Office of Scientific Research, 7/2005 – 1/2008.

In our earlier studies [49, 86], we reported HiQSAR of a set of 55 halocarbons on cell level toxicity. Because this group of chemicals is very important as synthetic chemistry agents and also as environmental pollutants, a subset of fourteen halomethanes, haloethanes, and ethylenes were chosen for the proteomics study. Six cell level toxicity data, *viz*., mitochondrial function (MTT), membrane integrity (LDH), total cellular thiols (SH), lipid peroxidation (LP), reactive oxygen species (ROS), and catalase activity (CAT), were determined for these chemicals. Proteomics analysis of the exposed cells was carried out by Dr. Frank Witzmann at the Indiana University and the data were provided to Dr. Subhash C. Basak's group at the University of Minnesota. For QSAR modeling, we calculated the chemodescriptors by software mentioned above and included quantum chemical indices. For biodescriptors, we used map information content [81], spectrum like descriptors [82] and subsets of spots (SOS) deemed important by the toxicologists. The details of ridge regression results using the above chemo-

and biodescriptors are not given here for brevity. To give a brief summary, in terms of cross validated R^2 using ridge regression, the results were: 1) MTT had best predictive model using a combination of chemodescriptors and map information content [81] calculated from halocarbon exposed proteomics patterns; 2) LDH was best predicted by chemodescriptors and spectrum like descriptors [82]; 3) TI and SOS gave best results for SH toxicity data; 4) Chemodescriptors plus MIC gave best results for LP; 5) ROS was most effectively predicted by computed chemodescriptors with absolutely no improvements in model quality with the addition of all classes of biodescriptors; 6) CAT also was best predicted by the computed chemodescriptors with marginal improvement in the effectiveness of predictability with the addition of proteomics based quantitative biodescriptors and SOS selected by the toxicologist. So, it may be said that neither chemodescriptors nor biodescriptors alone had sufficient information capable of predicting all the six cell level toxicity data generated for the halocarbons. As more data is available over time, researchers can attempt such approaches on newer and larger data sets to compare the effectiveness of chemodescriptors *versus* biodescriptors in predictive pharmacology and toxicology.

CONCLUSION

> *"All generalizations are dangerous, even this one".*
>
> *Alexandre Dumas*

At this juncture, after reviewing results of a large number of QSAR/QSTR studies using chemodescriptors and biodescriptors, we may ask ourselves: Quo Vadimus? We have seen that calculated chemodescriptors are capable of predicting physicochemical, pharmacological, and toxicological properties as well as toxic modes of action of chemicals. Research using biodescriptors of different types also shows that such descriptors derived from proteomics maps have reasonable power in discriminating among structurally and mechanistically related toxicants. Can we, at this stage, opt for either chemo- or biodescriptors alone? The answer is *NO*, as evident from our experience with the six cellular toxicity endpoints of halocarbons. This shows that in the foreseeable future in predictive pharmacology and toxicology we will need an integrated QSAR (I-QSAR) approach consisting of both chemodescriptors and biodescriptors in order to obtain the best results, as shown in Fig. (**3**) below.

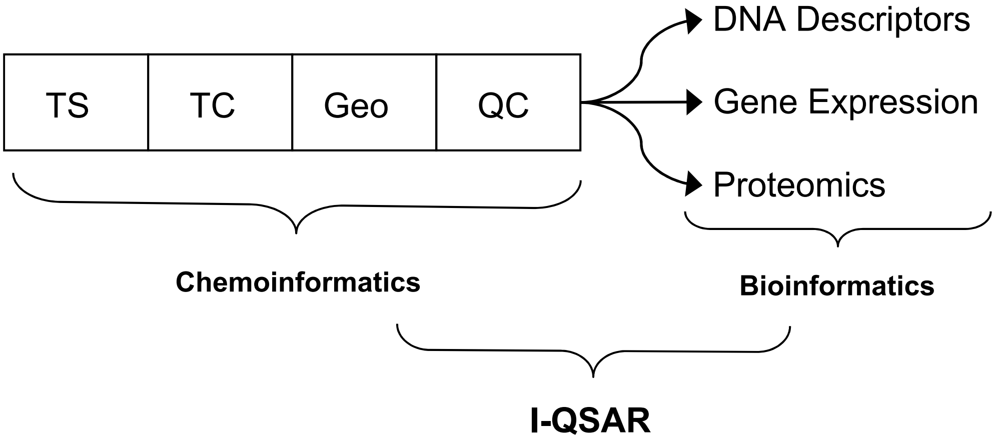


Figure 3: Integrated QSAR, combining chemodescriptors and biodescriptors.

This is analogous to the combination of structural and functional criteria proposed by Arcos [85] more than two decades ago for the assessment of chemical carcinogenesis; however, in the post-genomic era we can use more sophisticated genomics and proteomics data as the source of biodescriptors as opposed to the classical laboratory bioassay and test data.

Applications of discrete mathematical techniques to chemistry had a great growth spurt around the middle of the twentieth century. More recently, such methods have been applied for the characterization of the "omics" data also, as evident from the results reviewed here. We reproduce the editorial by Dilip K. Sinha and Subhash C. Basak to point out this expanding "chemobioinformatics continuum."

GUEST EDITORIAL

Fourth Indo-U.S. Workshop on Mathematical Chemistry,

January 8-12, 2005, Pune, Maharashtra, India

"The Fourth Indo-U.S. Workshop on Mathematical Chemistry with applications in drug design, risk assessment of chemicals, chemoinformatics, bioinformatics, computational biology, and toxicology was held on January 8-12, 2005, in Pune, Maharashtra, India, under the joint sponsorship of the Natural Resources Research Institute (NRRI) of the University of Minnesota, Duluth, USA, and the University of Pune. This issue of the Journal of Chemical Information and Modeling contains papers presented at the workshop. The concept of the Indo-U.S. workshop series was originally conceived by Subhash Basak, a senior scientist at NRRI, and received enthusiastic support from Dilip K. Sinha, a mathematician and educator from India. Together, Basak and Sinha have remained the chairpersons of the biennial, international Indo-U.S. Workshop series from the USA and India, respectively. The first event of the series was held in 1998 at Visva Bharati University, India, where Dilip Sinha was the vice chancellor at that time; the second and third workshops were organized by NRRI on the campus of the University of Minnesota, Duluth. The success of the Fourth Indo-U.S. Workshop, with the participation of over 125 participants from five continents, shows that the workshop series has established itself as one of the most important conferences in the field. The quality of the presented papers published in this volume after peer review demonstrates the high standard of scientific discourse taking place at the workshop.

> *"Discrete mathematical chemistry has made important advances in the past 25 years. This has been fueled primarily by two factors: (a) the formulation of new concepts and (b) easy access to high-speed computers. Methods developed in this field have found applications in pharmaceutical drug design and hazard assessment of environmental pollutants. Interestingly, discrete mathematical concepts, originally developed for the characterization of chemical systems, are being extended to deal with the explosion of data in "omics" science, namely, genomics, proteomics, and so forth. A few of the 17 papers from the Fourth Indo-U.S. Workshop presentations published in this issue of JCIM are outstanding examples of this expanding* ***chemo-bioinformatics continuum****."*
>
> *"By Dilip K. Sinha, Chairperson (India), Indo-U.S. Workshop on Mathematical Chemistry Series; Subhash C. Basak, Chairperson (USA), Indo-U.S. Workshop on Mathematical Chemistry Series and President, International Society of Mathematical Chemistry"*

Reprinted with permission from (*J. Chem. Inf. Model.* **2006,** *46,* 1-1, **1**). Copyright (2006), American Chemical Society.

Whereas different fields of science usually have their tight boundaries like silos, in today's knowledge based interdisciplinary research environment often such boundaries are being shattered for the benefit of all. One important application of discrete mathematical methods to science is network analysis [86]. Many basic philosophical issues related to the applications of discrete mathematics to chemical and biological systems have been discussed by Basak [87] in an article published in HYLE.

ACKNOWLEDGEMENTS

I am thankful to US Environmental Protection Agency, National Institute of Neurological Disorders and Stroke (NINDS/NIH), Minnesota Technology Inc, Air Force Office of Scientific Research (AFOSR, US Air Force), Center for Disease Control and Prevention (CDC), and US Department of Agriculture (USDA) for sustained support of my research program at UMD-NRRI for over three decades.

I am thankful to Alexandru T. Balaban, Krishnan Balasubramanian, Kanika Basak, Kevin Geiss, Gregory Grunwald, Xiaofeng Guo, Brian Gute, , Marissa Harle, Douglas Hawkins, Denise Mills, Ashesh Nandy, Ramanathan Natarajan, Terry Neumann, Gerald J. Niemi, Sonja Nikolic, , Marjana Novic, Milan Randic, Dilip K. Sinha, Nenad Trinajstic, Gilman D. Veith, José Luis Villaveces, Marjan Vracko, Frank Witzmann, Qianhong Zhu for collaboration in research and workshop organization activities.

I deeply appreciate the help that I received from the co-editor of this book, Guillermo Restrepo. Finally technical support from Wilmer Leal is highly appreciated.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

REFERENCES

[1] Hansch, C.; Leo, A. *Exploring QSARs: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society, Washington, DC **1995**.

[2] Kier, L.B.; Hall, L. *Molecular Structure Description:* The Electrotopological State; Academic Press: San Diego, CA, **1999**.

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

[4] Diudea, M.V., Ed. *QSPR / QSAR Studies by Molecular Descriptors*; Nova: Huntington, N.Y., **2001**.

[5] Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, **2000**.

[6] Balaban, A.T., Ed. *From Chemical Topology To Three-Dimensional Geometry*; Plenum Press: **1997**.

[7] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, **2009**, Vol. I and II. pp. 257.

[8[Hawkins, D. M.; Basak, S. C.; Kraker, J. J.; Geiss, K. T.; Witzmann, F. A., Combining chemodescriptors and biodescriptors in quantitative structure-activity relationship modeling, *J. Chem. Inf. Model*., **2006**, 46, 9-16.

[9] Basak, S. C., Role of Mathematical Chemodescriptors and Proteomics-Based Biodescriptors in Drug Discovery, *Drug Develop. Res*., **2010**, 72, 1-9.

[10] Basak, S. C.; Mills, D.; Hawkins, D. M., Predicting allergic contact dermatitis: A hierarchical structure-activity relationship (SAR) approach to chemical classification using topological and quantum chemical descriptors. *J. Comput. Aided Mol. Des*., **2008**, 22, 339-343.

[11] Crum-Brown A.; Fraser, T. R., On the connection between chemical constitution and physiological action. Part1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans R Soc Edinb*, **1868**, 25: 151-203.

[12] Richet M, C., Note sur le rapport entre la toxicite´ et les proprie´ te´ s physiques des corps. *CR Soc Biol (Paris)*, **1893,** 45: 775-776.

[13] Hansch, C.; Fujita, T. ρ- σ-∏ Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, **1964,** 86,1616-1626.

[14] Hammett, L. P., **1940**. Physical organic chemistry. New York: McGraw-Hill, pp. 404.

[15] Taft, R. W. Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters, *J. Am. Chem. Soc.,* **1952**, 74, 2729-2730.

[16] Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W. Linear Solvation Energy Relationships. 23. A Comprehensive Collection of the Solvatochromic Parameters, ∏*, *α,* and β, and Some Methods for Simplifying the Generalized Solvatochromic Equation *J. Org. Chem*. **1983,** 48, 2877- 2887.

[17] Cash, G. Personal Communication. Toxic Substances Control Act (TSCA) Inventory, United States Environmental Protection Agency (USEPA), Washington, D. C., **2013**.

[18] 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*., **1990**, 87,183-197.

[19] Kier, L.B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, **1976**, pp. 257.

[20] Primas, H., *Chemistry, Quantum Mechanics and Reductionism: Perspectives in Theoretical Chemistry*, Springer, Berlin, **1981**, pp. 451.

[21] Basak, S. C.; Niemi, G. J.; Veith, G. D., Predicting properties of molecules using graph invariants, *J. Math. Chem.*, **1991**, 7, 243-272.

[22] Bunge, M., *Method, Model and Matter*, Reidel Publishing Co., D. Dordrecht-Holland/Boston, **1973**.

[23] Trinajstic, N.,: **1997**: 'Mathematics and chemistry: The unlikely partners', in: D. H. Rouvray (ed), *Concepts in Chemistry: A contemporary challenge*, Research Studies Press Ltd., Taunton, Somerset, England, pp. 17-39.

[24] Harary, F. *Graph theory*, 2nd ed; Addison-Wesley: Reading, MA, **1969**.

[25] Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R., Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.* **1988**, 19, 17-44.

[26] Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, **1992**, pp. 352.

[27] Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17-20.

[28] Hosoya**,** H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn***.** **1971***,* *44*, 2332-2339.

[29] DRAGON - Software for the Calculation of Molecular Descriptors, Version 5.4, **2006;** Todeschini, R.; Consonni, V.; Mauri, A. *et al.*, Talete srl.; Milan, Italy.

[30] *MolConnZ*, Version 4.05, **2003**; Hall Ass. Consult.; Quincy, MA.

[31] Basak, S. C.; Harriss, D. K.; Magnuson, V. R. 1988. *POLLY v. 2.3:* **1988***;* Copyright of the University of Minnesota.

[32] Basak, S. C.; Grunwald, G. D., APProbe. **1993**; Copyright of the University of Minnesota.

[33] Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609-6615.

[34] Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure Activity Analysis*; Wiley: London, **1986**, pp.262.

[35] Filip, P. A.; Balaban,T. S.; Balaban, A. T. A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability. *J. Math. Chem.* **1987**, *1*, 61-83.

[36] Basak, S, C. 1999. Information theoretic indices of neighborhood complexity and their applications, In: J. Devillers and A.T. Balaban, editors. *Topological Indices and Related Descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers. The Netherlands, pp. 563-593.

[37] Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517-4533.

[38] Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research studies Press: Chichester, U.K.; **1983**.

[39] Johnson, M, Basak, S. C.;, Maggiora, G. A characterization of molecular similarity methods for property prediction. *Mathl. Comput. Modelling* ***1988,*** 11, 630-634.

[40] Hawkins, D. M.; Basak, S. C.; Shi, X. QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.*, **2001**, 41, 663-670.

[41] Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci., **2003**,* 43, 579-586.

[42] Kraker, J. J.; Hawkins, D. M.; Basak, S. C.; Natarajan, R.; Mills, D. Quantitative structure-activity relationship (QSAR) modeling of juvenile hormone activity: Comparison of validation procedures. *Chemometr. Intell. Lab. Syst.* **2007,** 87: 33-42.

[43] Basak, S. C.; Mills, D. Mathematical Chemistry and Chemoinformatics: A Holistic View Involving Optimism, Intractability, and Pragmatism. In: MATHEMATICAL METHODS AND MODELLING FOR STUDENTS OF CHEMISTRY AND BIOLOGY; Graovac, A.; Gutman, I.; Vukicevic, D., Eds., University of Split, and Institute Ruder Boskovic, Zagreb, **2009**, pp. 211-242.

[44] Basak, S. C.; Mills, D. Predicting vapor pressure of chemicals from structure: A comparison of graph theoretic *versus* quantum chemical descriptors. *SAR QSAR Environ. Res*. **2009,** 20, 119-132.

[45] Gute, B. D.; Basak, S. C. Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach, *SAR QSAR Environ. Res.* **1997**, 7, 117-131.

[46] Gute, B. D. Grunwald, G. D.; Basak, S. C. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach, *SAR QSAR Environ. Res.* **1999**, 10, 1-15.

[47] Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: A hierarchical QSAR approach, *J. Chem. Inf. Comput. Sci.* **2001**, 41, 671-678.

[48] Hawkins, D. M.; Basak, S. C.; Mills, D. QSAR for chemical mutagens from structure: ridge regression fitting and diagnostics, *Environ. Toxicol. Pharmacol.* **2004**, 16, 37-44.

[49] Gute, B. D.; Basak, S. C.; Balasubramanian, K.; Geiss, K.; Hawkins, D. M. Prediction of halocarbon toxicity from structure: A hierarchical QSAR approach, *Environ. Toxicol. Pharmacol.* **2004**, 16, 121-129.

[50] Basak, S. C.; Natarajan, R.; Mills, D. Structure-activity relationships for mosquito repellent aminoamides using the hierarchical QSAR method based on calculated molecular descriptors. *WSEAS Transactions on Information Science and Applications*, **2005,** 7, 958-963.

[51] Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. A. Prediction of tissue: air partition coefficients: A comparison of structure-based and property-based methods, *SAR QSAR Environ. Res*. **2002**, 13, 649-665.

[52] Basak, S. C.; Mills, D.; Mumtaz, M. M.; Balasubramanian, K. Use of topological indices in predicting aryl hydrocarbon (Ah) receptor binding potency of dibenzofurans: A hierarchical QSAR approach, *Indian. J. Chem.*, **2003,** 42A, 1385-1391.

[53] Basak, S. C.; Mills, D.; Gute, B. D.; Balaban, A. T.; Basak, K.; Grunwald, G. D. Use of Mathematical Structural Invariants in Analyzing Combinatorial Libraries: A Case Study with Psoralen Derivatives, *Curr. Comp. Aided Drug Design*, **2010**, *6*, 240-251.

[54] Basak, S. C.; Grunwald, G. D.; Host, G. E.; Niemi, G. J.; Bradbury, S. P. A comparative study of molecular similarity, statistical and neural network methods for predicting toxic modes of action of chemicals, *Environ. Toxicol. Chem.* **1998**, 17, 1056-1064.

[55] Gurwitz, D. Malaria Drugs: Clues From Malaria Resistance Genetics, *Drug Dev Res*. **2010**, 71, 1-3.

[56] Basak, S. C.; Mills, D. Quantitative structure-activity relationships for cycloguanil analogs as PfDHFR inhibitors using mathematical molecular descriptors, SAR QSAR Environ. Res. **2010**, 21, 215-229.

[57] Basak, S. C.; Mills, D.; Hawkins, D. M. Characterization of Dihydrofolate Reductases from Multiple Strains of *Plasmodium falciparum* using Mathematical Descriptors of their Inhibitors, *Chemistry and Biodiversity*, **2011,** 8, 440-453.

[58] Basak, S. C.; Gute, B. D.; Mills, D. Similarity methods in analog selection, property estimation and clustering of diverse chemicals, ARKIVOC, **2006,** 9, 157-210.

[59] CAS Registry Number and Substance Counts Home Page: http://www.cas.org/cgi-bin/regreport.pl (accessed June 6, **2013**).

[60] Goodman, A.G., Wall, T.W., Nies, A.S., Taylor, P., Eds. Goodman and Gilman's The Pharmacological Basis of Therapeutics; Pergamon Press: New York, 1990.

[61] Johnson, M.; Maggiora, G.M., Eds. *Concepts and Applications of Molecular Similarity;* John Wiley & Sons, Inc.: New York, **1990**.

[62] Willett, P.; Barnard, J.; Downs, G. Chemical similarity searching. *J. Chem. Inf. Comput. Sci* **1998,** *38*, 983-996.

[63] Thornber, C.W. Isosterism and molecular modification in drug design. *Chem. Soc. Rev.* **1979,** *8*, 563-580.

[64] Carbo-Dorca, R., Mezey, P.G. Eds. Advances in Molecular Similarity. Vol. 2; JAI Press: Stamford, Connecticut., USA, **1998**.

[65] Basak, S.C.; Grunwald, G.D. Tolerance space and molecular similarity. *SAR QSAR Environ. Res*. **1995,** *3*, 265-277.

[66] Schreider, J.A. *Equality, Resemblance, and Order;* Mir Publishers; Moscow, **1975.**

[67] Basak, S. C.; Gute, B. D.; Mills, D.; Hawkins, D. M. Quantitative molecular similarity methods in the property/toxicity estimation of chemicals: A comparison of arbitrary *versus* tailored similarity spaces. *J. Mol. Struct. (THEOCHEM),* **2003,** *622*, 127-145.

[68] Gute, B. D.; Basak, S. C.; Mills, D.; Hawkins, D. M. Tailored similarity spaces for the prediction of physicochemical properties. *Internet Electr. J. Mol. Design,* **2002,** *1*, 374-387.

[69] Basak, S. C.; Gute, B. D.; Mills, D. Quantitative molecular similarity analysis (QMSA) methods for property estimation: A comparison of property-based, arbitrary, and tailored similarity spaces. *SAR QSAR Environ. Res*., **2002,** 7-8, 727-742.

[70] Hamori, E.; Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.*, **1983**, *258*, 1318-1327.

[71] Gates, M.A. A Simple way to look at DNA. *J. Theor. Biol*., **1986**, *119*, 319-328.

[72] Nandy, A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. *Curr. Sci*., **1996**, *70*, 661-668.

[73] Leong, P.M.; Morgenthaler, S. Random walk and gap plots of DNA sequences. *Comput. Applic. Biosc*., **1995**, *11*, 503-507.

[74] Nandy, A.; Harle, M.; Basak, S.C. Mathematical descriptors of DNA sequences: development and applications. *Arkivoc,* **2006**, *9*, 211-238.

[75] Randić, M.; Zupan, J.; Balaban, A.T.; Vikic-Topic, D.; Plavsic, D. Graphical Representation of Proteins. *Chem. Rev.,* **2011**, *111,* 790-862.

[76] Indo-US Workshop Series on Mathematical Chemistry: http://www.nrri.umn.edu/indousworkshop/

[77] Raychaudhury, C.; Nandy, A. Indexation Schemes and Similarity Measures for Macromolecular Sequences. Paper presented at the Indo-US Workshop on Mathematical Chemistry, Visva Bharati University, Santiniketa, West Bengal, India, January 9-13, **1998**.

[78] Randić, M.; Vracko, M.; Nandy, A.; Basak, S.C. On 3-D representation of DNA primary sequences, *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 1235-1244.

[79] Guo, X.; Randić, M.; Basak, S.C. A novel 2-D graphical representation of DNA sequences of low degeneracy, *Chem. Phys. Lett.,* **2001**, *350,* 106-112.

[80] Randic, M.; Witzmann, F.; Vracko, M.; Basak, S. C. On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: Application to peroxisome proliferators, *Med. Chem. Res.,* **2001**, 10, 456-479.

[81] Basak, S. C.; Gute, B. D.; Witzmann, F. Information-theoretic biodescriptors for proteomics maps: Development and applications in predictive toxicology, *WSEAS Transactions on Information Science and Applications*, **2005,** 7, 996-1001.

[82] Vracko, M.; Basak, S. C.; Geiss, K.; Witzmann, F. Proteomics maps-toxicity relationship of halocarbons studied with similarity index and genetic algorithm, *J. Chem. Inf. Model.*, **2006,** 46, 130-136.

[83] Basak, S. C.; Gute, B. D. Mathematical descriptors of proteomics maps: Background and applications, C*urr. Opin. Drug Discov. Devel*., **2008**, 11, 320-326.

[84] Basak, S. C.; Gute, B. D.; Monteiro-Riviere, N.; Witzmann, F. Characterization of toxicoproteomics maps for chemical mixtures using information theoretic approach, In: Principles and Practice of Mixtures Toxicology, M. Mumtaz, Ed., Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, **2010**, pp. 215-232.

[85] Arcos, J. C. Structure-Activity Relationships: Criteria for Predicting the Carcinogenic Activity of Chemical Compounds, *Environ. Sci. Technol*., **198**7, 21, 743-745.

[86] Dehmer, M.; Basak, S. C., Eds, Statistical and Machine Learning Approaches for Network Analysis, Wiley, Hoboken, New Jersey, USA, **2012**.

[87] Basak, S. C. Philosophy Of Mathematical Chemistry: A personal perspective, *HYLE,* **2013**, *19,* 3-17.

CHAPTER 2

Ordering Thinking in Chemistry

Guillermo Restrepo*

Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Colombia

Abstract: We give some basic mathematical ideas of partially ordered sets (posets), which frame into the mathematical way of thinking illustrated in the Erlangen Programme by Felix Klein. The programme entails extracting relevant variables to study, symbolising them and relating them through a function. We show several examples where the mathematical way of thinking, restricted to partial orders, is found in chemistry. The examples are: Geoffroy's affinity table, benzene's structure, posetic predictive methods, multicriteria situations and derivation of concepts. Finally we question the ranking process by showing how it disregards its underlying, and not always recognised, posetic nature.

Keywords: Partial order, partially ordered sets, mathematical way of thinking, erlangen programme, posets, mathematical chemistry, ranking, affinity tables, benzene's structure, posetic predictive methods, multicriteria approaches, formal concept analysis, mutagenicity, estimation of properties, philosophy of chemistry.

INTRODUCTION

Ordering is important in daily life routines as it pervades decision making processes [1]. We are always interested in making the best decision based on different attributes of the possible decisions to come up with an ordering of the possibilities. It has been found [2] that in a conversation, where several people talk about a particular subject, the order in which they discuss is important, for listeners report as their own memory what the first speaker reported more than what a subsequent speaker reported.

Rudolf Arnheim, film theorist and perceptual psychologist, defined order as "the degree and kind of lawfulness governing the relations among the parts of an

***Corresponding author Guillermo Restrepo:** Laboratorio de Química Teórica, Universidad de Pamplona, Pamplona, Norte de Santander, Colombia; Tel: +57 568 5303; Fax: +57 570 3742;
E-mails: grestrepo@unipamplona.edu.co, guillermorestrepo@gmail.com

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50002-3

entity" [3]. According to Lorand [4], Arnheim definition suggests that order i) resides in *complex systems*, for these systems have distinct parts; ii) is *quantitative*, *i.e.* one can talk about degrees of order; iii) consists of *relations* amongst the parts of the systems and iv) is *lawful* as it involves a law or principle governing the relations amongst the parts. Now, let us find the mathematical ideas behind Arnheim's and Lorand's assertions. The idea of sets is behind complex systems, where its constitutive parts are set elements. If total orders are taken as a reference, then one can establish measures of nearness to that total order; in that sense one can quantify the degree of order of a set. By total order is meant a set where each element of the set can be associated to a unique natural number. A more formal definition of total order is given below. The relations mentioned by Lorand correspond to mathematical relations, particularly to order relations as we will show below. Finally, lawful can be led to mathematics by relating it with a principle of ordering that depends on the context, *i.e.* given a set of elements, they can be ordered in manifold ways depending on the attributes of the elements considered for the ordering. A more formal definition of order is given below.

Ordering has quite a lot of importance in different fields of knowledge and in particular applications, *e.g.* rankings. A ranking is a derived product of an ordering and its proliferation has led to find rankings of universities [5, 6], health systems [7], biodiverse ecosystems [8], scientific journals [9] and even scientists [10]. Ranking is present too in document retrieval processes [11, 12] and even in search engines exploring the World Wide Web such as the PageRank [13] of Google.

In economy, for example, ordering is found in the arrangement of countries based on their public level of satisfaction regarding different public services [14]; in studies on diversity, it has been proven that the different diversity measures underlie a general order [15]. One of the aims of observational studies [16] is to measure the effect of a cause, *e.g.* a medical treatment; which yields an ordering on a control set and on a treatment set.

In the following section we discuss in detail some orderings in chemistry, particularly those in which the authors have been active, and refer the reader to important reviews on the subject. Such a section has also comments on recent works on the mathematical way of thinking, which are explained and combined with order ideas.

MATHEMATICAL WAY OF THINKING

In a recent paper [17] we argue that mathematical chemistry is the realisation of the mathematical way of thinking as discussed by Weyl [18]. This approach follows the functional thinking after Felix Klein's [19] Erlangen Programme. The general idea is to look for variables, symbols and functions of the problem tackled; for the particular case of mathematical chemistry the three components of the thinking are looked for in events of chemical interest [20]. These components are related in the following way: i) by treating a chemical situation, the scientist looks for its characterisation through variables that are filtered and reduced to a small amount of relevant ones. The selected variables are then abstracted through their symbolisation, which finally leads to finding functions relating the selected variables.

An example of the mathematical way of thinking, followed in several of the chapters of the current book, is the prediction of substances' properties based on their molecular structure. Note the selection of the molecular structure as a relevant variable, for several other substances' features could be selected, *e.g.* experimental properties as the initial QSAR approaches by Hansch [21]. The molecular structure can then be characterised by several ways, *e.g.* using molecular descriptors [22] or fingerprints [23]. At this point, the problem of predicting substances' properties based on their molecular structure has led to the relevant variables: molecular descriptors and mutagenicity (if the property of interest is mutagenicity, for example). Then, one proceeds to the second step, which is the symbolisation of those variables, *e.g.* d_1, d_2, …, d_n for the n descriptors and m for mutagenicity. Finally, the functional thinking is completed when a function of the form $m = f\,(d_1, d_2, \ldots, d_n)$ is found. Some other examples of the chemical thinking are shown in reference [17].

As this particular chapter deals with order in chemistry, we now introduce some formal ideas of order theory.

Definition 1. A binary relation $\preccurlyeq$ on a non-empty set X is called a *partial order* if:

1. $x \in X \Longrightarrow x \preccurlyeq x$.
2. $x, y \in X, x \preccurlyeq y$ and $y \preccurlyeq x \Longrightarrow x = y$.
3. $x, y, z \in X, x \preccurlyeq y$ and $y \preccurlyeq z \Longrightarrow x \preccurlyeq z$.

Which makes $\preccurlyeq$ a relation fulfilling reflexivity, antisymmetry and transitivity. The set X along with $\preccurlyeq$ is called a *partially ordered set* (*poset*) and is denoted by $(X, \preccurlyeq)$.

Definition 2. Given $x, y \in X$, they are called *comparable* if either $x \preccurlyeq y$ or $y \preccurlyeq x$, otherwise they are *incomparable*.

Definition 3. For any $x, y \in X$, *y covers x* (*x* is *covered* by *y*) if $x \preccurlyeq y$ and there is no $z \in X$ for which $x \preccurlyeq z$ and $z \preccurlyeq y$. This is denoted by $x \preccurlyeq: y$.

Definition 4. Let $H = (X, C)$ be a directed graph of $(X, \preccurlyeq)$, where C is the set of directed edges containing the cover pairs in X. H is called the *Hasse diagram* of the poset $(X, \preccurlyeq)$ if it is drawn in the Euclidean plane whose horizontal/vertical coordinate system requires that the vertical coordinate of $x \in X$ be larger than the one of $y \in X$ if $y \preccurlyeq: x$.

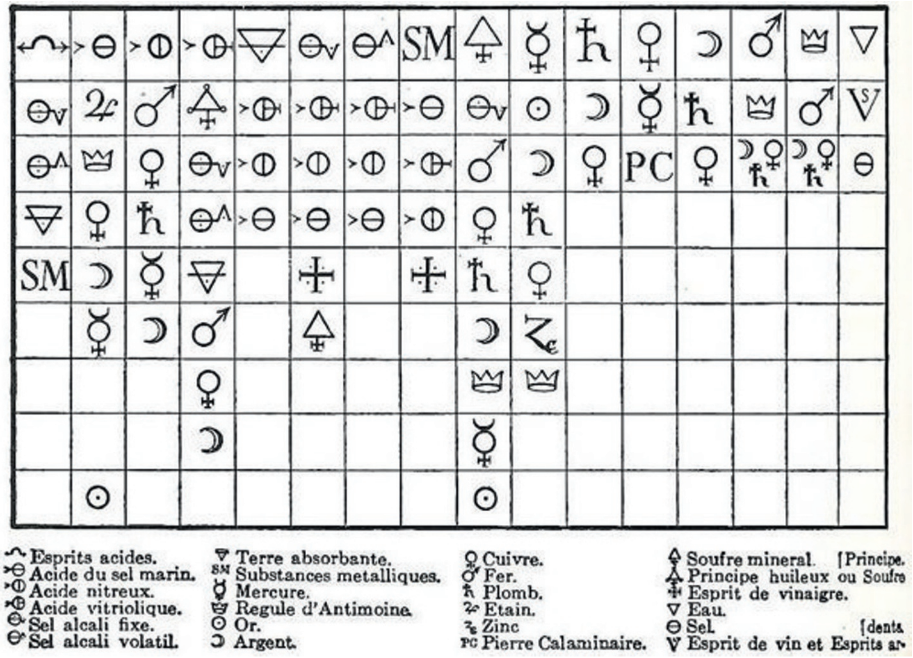
Some particular instances of order theory in chemistry are Geoffroy's affinity table (see below), Ruch's algebraic description of chirality [24-26]; Randić's approaches to molecular structure [27], Halfon *et al.* ordering of substances in environmental chemistry [28]; Brüggemann's further mathematical explorations of Halfon *et al.* approach [1, 29, 30] and Klein's approaches to estimate substances' properties [31-34]. An in-deep discussion on posets in chemistry is found in Ref. [35] and in the several papers of *MATCH Communications in Mathematical and in Computer Chemistry*, Volume 42 (2000).

ORDER THEORY IN THE MATHEMATICAL WAY OF THINKING IN CHEMISTRY

In the following discussion we show some examples of order theory meeting the mathematical way of thinking in chemistry.

Geoffroy's Affinity Table

Following Cartesian philosophy, in 17th century France, several scholars developed the vision of a rational, mathematized chemical knowledge, being Etienne-François Geoffroy (1672-1731) one of them. Geoffroy introduced the *Table des differents rapports observés entre differentes substances* (Fig. **(1)**). He wanted a table containing the different *rapports* (relations) of the "principal matters one is accustomed to work on in chemistry" [36]. The table soon became not only a collection of information but also a tool for predicting salts and their chemical reactions [17].



↼⌒⇀	>⊖	>⦶	>⊕	🜃	⊖v	⊖^	SM	🜍	☿	♄	♀	☽	♂	🜬	▽
⊖v	♃	♂	🜍	>⊕	>⊕	>⊕	>⊖	⊖v	☉	☽	☿	♄	🜬	♂	🜈
⊖^	🜬	♀	⊖v	>⦶	>⦶	>⦶	>⊕	♂	☽	♀	PC	♀	☽ ♀ ♄	☽ ♀ ♄	⊖
🜃	♀	♄	⊖^	>⊖	>⊖	>⊖	>⦶	♀	♄						
SM	☽	☿	🜃		🜊		🜊	♄	♀						
	☿	☽	♂		🜍			☽	ζ						
			♀					🜬	🜬						
			☽					☿							
	☉							☉							

↼⌒⇀ Esprits acides.
>⊖ Acide du sel marin.
>⦶ Acide nitreux.
>⊕ Acide vitriolique.
⊖v Sel alcali fixe.
⊖^ Sel alcali volatil.

🜃 Terre absorbante.
SM Substances metalliques.
☿ Mercure.
🜬 Regule d'Antimoine.
☉ Or.
☽ Argent.

♀ Cuivre.
♂ Fer.
♄ Plomb.
♃ Etain.
ζ Zinc
PC Pierre Calaminaire.

🜍 Soufre mineral. [Principe.
🜍 Principe huileux ou Soufre
🜊 Esprit de vinaigre.
▽ Eau.
⊖ Sel. [dents
🜈 Esprit de vin et Esprits ar-

Figure 1: Table of affinities by Geoffroy 1718.

The table is interpreted as follows: i) the top row has different substances employed in 17th-century chemistry; ii) below each of them, different substances are ordered according to their strength of affinity regarding the top substance. As an example, let us take the first column (left hand side), headed by acid spirits followed by fixed alkali salt, volatile alkali salt, absorbent earth and metallic substances. The column shows that fixed alkali salt reacts more favourably with acid spirits than the other substances down the column and that it displaces all substances below it from their existing combination with acid spirits. The column shows that volatile alkali salt displaces absorbent earth and metallic substances from their combinations with acid spirits but does not displace fixed alkali from its combination with acid spirits. As pointed out in reference [17], Geoffroy found order amongst substances based on their reactivity with reference to a substance (top of the column). The chemical context [20], a relational one, is highlighted in the table as in the second column headed by acid of marine salt, silver reacts more favourably with the acid in question than mercury. This behaviour contrasts with a different context, *e.g.* nitrous acid (third column), where the order is reversed and mercury reacts more favourably with nitrous acid than silver [36, p. 136].

Geoffroy's table meets the mathematical way of thinking as follows:

1. Variables: chemical substances customarily employed in the 17th century.

2. Symbols: those depicted in Fig. **(1)**.

3. Functions: order relationships shown in each column of Fig. **(1)**.

Fig. **(2)** shows an example of the function applied to the first column of Fig. **(1)**.

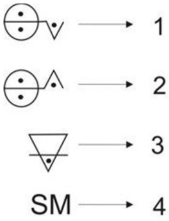


Figure 2: Order relation of the substances depicted in the left-hand-side column of Fig. (**1**) regarding their affinities towards acid spirits (top of the column); 1 indicates the greatest affinity, 4 the lowest one.

Benzene's Structure

In [32] Klein and Bytautas discuss how ordering of substances and their isomeric substituents were fundamental for finding the hexagonal symmetry of benzene molecules. The authors mention that Hässelbarth [37] proved that by counting the number of isomers at the different degrees of substitution, a wealth of the molecular symmetry can be known. Posets come into play as the way of relating substances by substitutions, for example. Klein and Bytautas claim that those posets, although not drawn explicitly in Kekulé's times (19th century), were relevant for the discussion on the molecular structure of benzene, *i.e.* a hexagon or a trigonal-prism structure. Geometrical aspects of molecules were normally disregarded in the 19th century and the inclination towards the hexagonal symmetry came finally with chemical reasons, *e.g.* chemical reactions by von Baeyer suggesting a better interpretation if the two substituted sites in the *ortho* isomer were geometrically adjacent or considering the number of naphthalene isomers [32]. Despite disregarding geometrical aspects of molecular structure that give place to symmetry, Klein and Bytautas show that symmetry considerations when treating the hexagon and the trigonal-prism lead to different numbers of isomers, which would have clearly differentiated both structures in favour of the hexagonal symmetry (Fig. **(3)**). However, 19th-century chemists were more akin to connectivity in atoms, more oriented towards graphs, than to geometrical assembles.

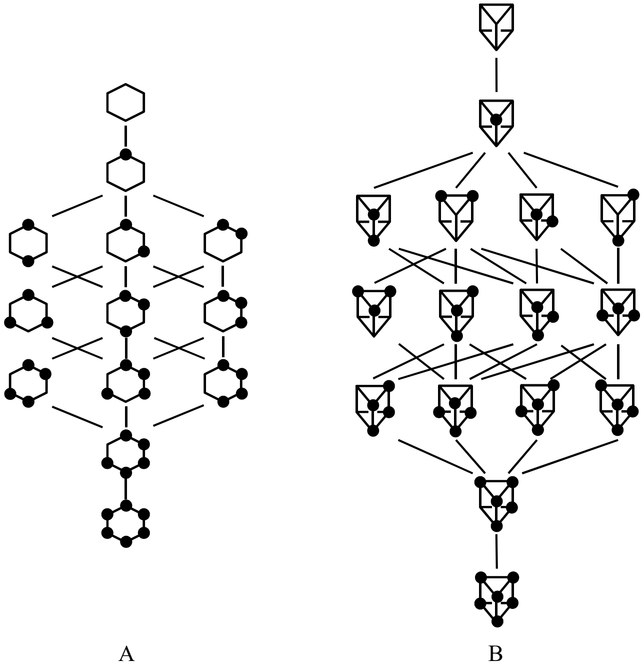


Figure 3: Posets for substituted benzenes with A) hexagonal and B) trigonal-prismatic geometry. In A and B the molecular skeleton is represented by a hexagon and by a trigonal-prism, respectively, where the carbon bonded to the substituted hydrogen is represented by a black circle.

These posets meet the mathematical way of thinking as follows:

1. Variables: substituted sites, levels and number of isomers.

2. Symbols: *s*, *l* and *i*.

3. Functions: $s \longrightarrow l \longrightarrow i$.

Each molecule is characterised by its number of substituted sites, thence the values *s* can take are {0, 1, …, 6}. A level is defined as the set of molecules in the poset having equal number of substituted sites, *i.e.* $l = \{x \mid x$ of the poset having $l \in s$ substituted sites$\}$. The cardinalities of the different values *l* can take are {1, 1, 3, 3, 3, 1, 1} for the poset in Fig. (**3**)A and {1, 1, 4, 4, 4, 1, 1} for the poset in Fig. (**3**)B, with the levels arranged in increasing order according to their number of substituted sites. Finally, the function $l \longrightarrow i$ assigns to the cardinality of each level a number of isomers with *l* substituted sites.

Posetic Predictive Methods

The kinds of posets shown in Fig. **(3)** have been further explored by Klein and coworkers [31-35], who have devised three estimative methods of the properties of the molecules in the poset by mathematical algorithms able to move in the network created for the posets. These methods are called average, cluster expansion and splinoid and are further discussed in Ref. [38, 39].

As an example, we explain the cluster expansion method that calculates the property *P*(*x*) of molecule *x* using features *z*(*y*) of all *y* comparable to *x*, with *x*, *y* ∈ (*X*, ≼).

$$P(x) = \sum_{y}^{\succcurlyeq x} n(y, x) \cdot z(y)$$

The expansion makes use of the number *n*(*y*, *x*) of ways in which configurational arrangements *C'* ∈ *y* occur as substructures in a configuration *C* ∈ *x*. Parameters *z*(*y*) may be derived by a fitting procedure where the cluster expansion may be conveniently truncated to a limited sequence of non-zero cluster terms *z*(*y*), and so applied when the earlier terms alone give a good approximation for the property *P*.

These posets meet the mathematical way of thinking as follows:

1. Variables: comparabilities, configurational arrangements, fitting parameters and property to be estimated.

2. Symbols: $x \preccurlyeq y$, n, z and P.

3. Functions: $x \preccurlyeq y \longrightarrow n(y, x) \longrightarrow P(x)$.

Some examples of application of these posetic approaches are in the estimation of properties of substituted benzenes, toluenes, cubanes [39] and hemoglobins [38], to name but a few cases.

Multicriteria Situations

In multicriteria analysis, important for decision making processes [40], the elements of a set are ordered based on their characterisation that considers several criteria or features of the elements. This approach has been further explored in the so-called *Hasse diagram technique* (HDT) [1,29,41], whose seeds are found in Halfon *et al.* studies [28].

In HDT, elements $x, y \in X$ are characterised by features $q_1(x), q_2(x), \ldots, q_i(x)$ and $q_1(y), q_2(y), \ldots, q_i(y)$, respectively; x is ordered lower than y ($x \preccurlyeq y$) if all its features are lower in magnitude than those of y, or if at least one feature is lower for x while all others are equal, which gives place to comparabilities. If all features of x and y are equal, both objects are called equivalent. If at least one feature q_j satisfies $q_j(x) < q_j(y)$ while the others are opposite ($q_i(x) \geq q_i(y)$), x and y are incomparable.

A recent application example of the HDT was the estimation of octanol/water partition coefficients of chlorophenols [42], where each chlorophenol, including phenol, was associated to a multi-indicator system by considering the five positions around the phenyl-group as components q_i of a vector characterising each chlorophenol (the numbering system used is shown in Fig. **(4)** along with the derived poset). The value q_i for the substance x is assigned as follows:

$$q_i(x) = \begin{cases} 1 & \text{if in the } i\text{-th position a H atom is bonded} \\ 0 & \text{otherwise} \end{cases}$$

Hence, the characterisation of *o*-chlorophenol is (0, 1, 1, 1, 1), while (0, 1, 0, 1, 0) is the one of 1,3,5-trichloro-phenol.

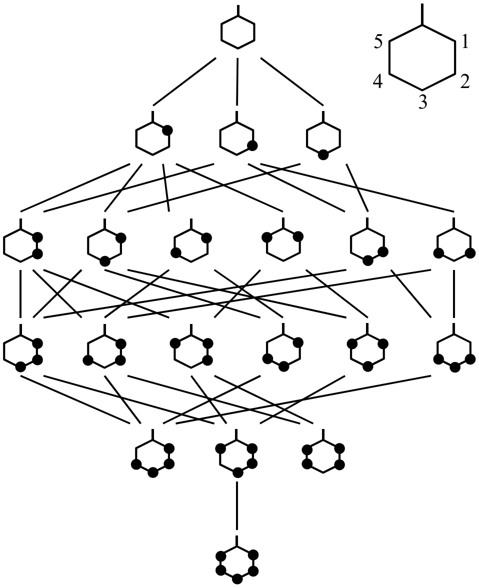


Figure 4: Poset of chlorophenols where the aromatic ring is represented by a hexagon and the carbon bonded to the substituted hydrogen is represented by a black circle. Top right phenol indicates the numbering convention used in the current chapter.

Once building up the poset, linear extensions and heights are introduced as follows: let L_i be a linear extension, *i.e.* a total order preserving all order relations in the poset. A *total order* is a poset where each couple of elements is comparable. In a finite set X, each linear extension has a *least* element. The height $h(x, L_i)$ of an element x regarding the linear extension L_i is given by counting the number of elements in $I(least, x)$ fulfilling $least \preccurlyeq y \preccurlyeq x$, being $I(least, x)$ the interval corresponding to objects least and x (the interval is defined as $\{y \in X \mid least \preccurlyeq y \preccurlyeq x\}$, with $least, x \in X$). The average height of x, $h_{av}(x)$, is calculated [43] as follows:

$$h_{av}(x) = \frac{\sum_{L_i} h(x, L_i)}{LT}$$

with LT being the total number of linear extensions of the respective poset through the HDT.

Brüggemann *et al.* [42] devised a method to relate h_{av} with the octanol/water partition coefficients (K_{OW}) of chlorophenols in Fig. **(4)**. In general, the model is of the form $K_{OW} = f(h_{av})$.

This HDT approach meets the mathematical way of thinking as follows:

1. Variables: comparabilities, average height and property to be estimated.

2. Symbols: $x \preccurlyeq y$, h_{av} and K_{OW}.

3. Functions: $x \preccurlyeq y \longrightarrow h_{av}(x) \longrightarrow K_{OW}(x)$.

Derivation of Concepts

One of the successes of mathematical chemistry, exemplified in several chapters of the current book, has been the development of methodologies for QSAR (Quantitative Structure-Activity Relationships) or more generally speaking QSPR (Quantitative Property-Activity Relationships), which in several cases circumvent the high costs in time and resources of experimental tests, *e.g.* toxicity. QSPR methods are based on the mathematical description of the molecular structure associated to the substances under study [44-47], where mathematical models depending on descriptors characterising the molecular structure are devised. In reference [17] we show how this approach meets the mathematical way of thinking in chemistry. QSPR models look for relating structural features of molecules with their properties; however the contrary relation is not an easy task by using these approaches [48-50]. This is mainly caused by the complexity of the models, which are customarily algebraic combinations of molecular descriptors that do not allow a straightforward interpretation of the model. The interpretation is also difficult as the meaning of molecular descriptors is not always an easy task.

These shortcomings have led several authors to develop alternative methods avoiding the numerical description of molecular structure (descriptors) and instead considering the graph associated to the structure as the molecular representation. One example of these approaches is the one by Bemis and Murcko [51], where the molecular structure is regarded as a framework made of ring systems and linkers, disregarding side chains. In this methodology atom identitites are not considered, *i.e.* all atoms are regarded as equivalent ones (there is no distinction between N, O, C, *etc.*). Some other non-numerical characterisations of molecules are found in references [52, 53].

One approach to relate the aforementioned characterisations of molecular structure with properties and the reverse relation is through Formal Concept Analysis, a data analysis technique based on order theory, where elements to order are described by attributes. The technique was introduced by Wille [54] and a brief description of it is the following [55] (for more details, see references [54,56, 57]):

i). The method starts with a context *i.e.* a matrix where rows are labelled with elements (in this case, molecules) and columns with attributes (non-numerical molecular characterisations and different scales of the property of interest). The entries of the matrix are 1s or 0s; 1 indicating the presence of the attribute of the column for the attribute of the row and 0 the absence of the attribute for that element.

ii). Based on the context, all possible concepts are found. A concept being a subset of elements (extent) uniquely characterised by a given subset of attributes (intent), which in turn uniquely describe the given set of elements.

iii). The concepts are ordered by set inclusion of their extents and intents, which leads to a particular poset, a lattice.

iv). Different implications *i.e.* relationships between attributes are obtained based on the order relations between concepts.

v). Associations are also derived; in this case it is possible to calculate the probability of a statement relating attributes by taking advantage of the number of objects meeting the respective association.

In Ref. [55] we applied FCA to the study of 95 heteroaromatic amines characterised by molecular frameworks and experimental mutagenicity scales for the amines. The results are summarised in Fig. **(5)**, where each node in the lattice contains two kinds of information: a set of mutagenicity scales (coloured boxes) and a set of frameworks. Once a node is selected, all those nodes found in a downward path give information about the node selected, *e.g.* if we are interested in the node marked with *, then the respective node shows that the molecular frameworks of two 6-ring systems connected by two atoms, two fused 6-ring systems and one 6-ring system are related to low-low mutagenicity, *i.e.* lowest mutagenicity. One of the associations found in the study shows that lowest mutagenicity is related to 6-ring systems with 83% probability [55].

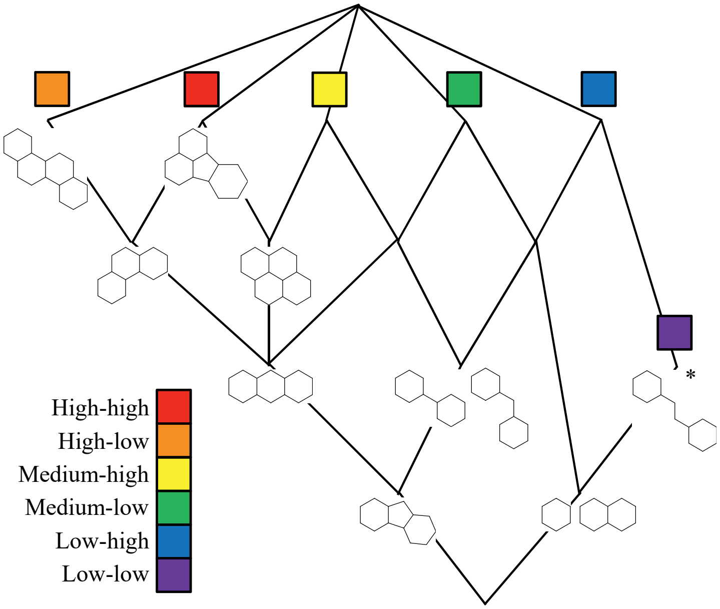


Figure 5: Lattice of molecular frameworks and mutagenicity scales, where the node (a concept) marked with * is further explained in the text.

This posetic approach meets the mathematical way of thinking as follows:

1. Variables: molecular frameworks and mutagenicity.

2. Symbols: *fr* and *m*.

3. Functions: $fr \longrightarrow m$ and $m \longrightarrow fr$.

Another FCA application, this time to radionuclides used in diagnosis, is found in reference [58].

CONCLUDING REMARKS

The examples of order theory in chemistry shown in the current chapter are a clear illustration of the presence of order theory in chemistry and of their different opportunities in mathematical chemistry. Perhaps the most important conclusion from this chapter is that by using the mathematical thinking in chemistry (particularly restricted to order theory) several chemical situations can be formalised. Hosoya [59] has claimed that all in all, chemists and mathematicians do not differ to a large extend in the kind of logic used. Something early recognised by Kant when stating that chemistry is the "paradigm for the method of critical philosophy"[1].

Through the chapter we have tried to make clear that it is not always possible to end up with a total order. As we have shown, incomparabilities are common in order studies, *e.g.* substituted benzenes in a particular level of Klein's posets. A particular popular kind of total order is the already mentioned ranking, which as discussed by Solomon [15], Klein and Babić [31, 61] and Restrepo [62], entails a partial order. The point to stress, and the criticism to rankings, is the fact of disregarding the underlying poset, which makes the mapping of the poset onto the reals a subjective process. This subjectivity justifies the distress of several institutions and authors to rankings [63, 64]. Solomon [15], for example, has shown that the discrepancies in diversity measures used in ecology are due to the different linear extensions a poset resulting from abundance vectors may produce. In other words, each diversity measure may show a particular linear extension of the same poset. Klein and Babić [31, 61], point out that posets may be deeply related to experimental sciences through the measuring process, where ambiguities resulting from measurements might be explained as the result of measuring elements, which in reality must be considered as incomparable. Hence, the measuring method may force the incomparabilities to be comparable and, because of the different possibilities to do this [65, 66], the outcomes may be different, therefore "ambiguous", in a systematically controlled way. All these examples make ponder on the reason and tendency of humans to map everything

[1] This contrasts with Kant's, unfortunately, famous assertion that "chemistry can become nothing more than systematic art or experimental doctrine, but never science proper". In [60] we showed that Lavoisier and his systematization of a big part of 18th-century chemistry played an important role in Kant's change of mind.

onto the reals [62]. As Klein [35] has pointed out, perhaps not everything is numerical in nature. Perhaps the urgency for mapping onto the reals distorts the object to study and it is better to look for other ways of understanding the object avoiding the reals. All in all, it is possible to do mathematics and understand phenomena not necessarily going to the kingdom of numbers [67].

ACKNOWLEDGEMENTS

We thank R. Brüggemann (Leibniz-Institute of Freshwater Ecology and Inland Fishery, Germany), A. Kerber (University of Bayreuth, Germany), D. J. Klein (Texas A&M University at Galveston, USA) and J. L. Villaveces (Universidad de los Andes, Colombia) for all their valuable and inspiring ideas on posets and mathematics in chemistry.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

REFERENCES

[1] Restrepo, G.; Brüggemann, R.; Weckert, M.; Gerstmann, S.; Frank, H. Ranking patterns, an application to refrigerants. *MATCH Commun. Math. Comput. Chem.* **2008**, *59*, 555-584.

[2] Wright, D.B.; Carlucci, M.E. The response order effect: People believe the first person who remembers an event. *Psychon. B. Rev*. **2011**, *18*, 805-812.

[3] Arnheim, R. *Toward a psychology of art*, University of California Press, Berkeley, **1966**, p. 123.

[4] Lorand, R. *Aesthetic order: a philosophy of order, beauty and art*, Routledge, London, **2000**.

[5] Taubes, G. Measure for measure in science. *Science* **1993**, *260*, 884-886.

[6] Ball, P. Index aims for fair ranking of scientists. *Nature* **2005**, *436*, 900-900.

[7] Jamison, D.T.; Sandbu, M.E. WHO Ranking of health system performance. *Science* **2001**, *293*, 1595-1596.

[8] Roberts, L. Ranking the rain forests. *Science* **1991**, *251*, 1559-1560.

[9] Shewchuk, R.M.; O'Connor, S.J.; Williams, E.S.; Savage, G.T. Beyond rankings: using cognitive mapping to understand what health care journals represent. *Soc. Sci. Med*. **2006**, *62*, 1192-1204.

[10] Hirsch, J.E. An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. USA*. **2005**, *102*, 16569-16572.

[11] Broder, A.Z.; Lempel, R.; Maghoul, F.; Pedersen, J. Efficient PageRank approximation *via* graph aggregation. *Inf. Retrieval* **2006**, *9*, 123-138.

[12] Willett, P. Textual and chemical information processing: different domains but similar algorithms. *Inform. Res*. **2000**, *5*(2), URL: http://InformationR.net/ir/5-2/infres52.html (accessed Jul. 2012).

[13] Page, L.; Brin, S.; Motwani, R.; Winograd, T. The PageRank citation ranking: bringing order to the web. URL: http://dbpubs.stanford.edu/pub/1999-66 (accessed Jul. 2012).

[14] Annoni, P. Different ranking methods: potentialities and pitfalls for the case of European opinion poll. *Environ. Ecol. Stat*. **2007**, *14*, 453-471.

[15] Solomon, D.L. In: *Ecological diversity in theory and practice*; Grassie, J.F.; Patil, G.P.; Smith, W.; Taillie, C. Eds.; International Co-operative Publishing House: Fairland, **1979**; pp. 29-35.

[16] Rosenbaum, P.R. *Observational studies*, Springer, New York, **1995**.

[17] Restrepo, G.; Villaveces, J.L. Mathematical thinking in chemistry. *HYLE – Int. J. Phil. Chem*., **2012**, *18*, 3-22.

[18] Weyl, H. The mathematical way of thinking. *Science* **1940**, *2394*, 437-446.

[19] Klein, F. *Vergleichende Betrachtungen über neuere geometrische Forschungen*, Verlag von Andreas Deichert, Erlangen, **1872**.

[20] Schummer, J. The chemical core of chemistry I: a conceptual approach. *HYLE – Int. J. Phil. Chem*., **1998**, *4*, 129-162.

[21] Hansch, C.; Fujita, T. p-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc*., **1964**, 86, 1616-1626.

[22] Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H. *Molecular descriptors for chemoinformatics*, 2 vols., Wiley-VCH, Weinheim, **2009**.

[23] Leach, A.R.; Gillet, V.J. *An introduction to chemoinformatics*, Kluwer, Dordrecht, **2007**.

[24] Ruch, E.; Schönhofer, A. Theorie der Chiralitätsfunktionen. *Theoret. Chim. Acta* **1970**, *19*, 225-287.

[25] Ruch, E. Algebraic aspects of the chirality phenomenon in chemistry. *Acc. Chem. Res*., **1972**, *5*, 49-56.

[26] Ruch, E. The diagram lattice as structural principle A. New aspects for representations and group algebra of the symmetric group B. Definition of classification character, mixing character, statistical order, statistical disorder; A general principle for the time evolution of irreversible processes. *Theoret. Chim. Acta*, **1975**, *38*, 167-183.

[27] Randić, M. On comparability of structures. *Chem. Phys. Lett*., **1978**, *55*, 547-551.

[28] Halfon, E.; Reggiani, M.G. On ranking chemicals for environmental hazard. *Environ. Sci. & Technol*., **1986**, *20*, 1173-1179.

[29] Brüggemann, R.; Bartel, H.G. A theoretical concept to rank environmentally significant chemicals. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 211-217.

[30] Brüggemann, R.; Sørensen, P.B.; Lerche, D.; Carlsen, L. Estimation of averaged ranks by a local partial order model. *J. Chem. Inf. Comput. Sci*., **2004**, *44*, 618-625.

[31] Klein, D.J.; Babić, D. Partial orderings in chemistry. *J. Chem. Inf. Comput. Sci*., **1997**, *37*, 656-671.

[32] Klein, D.J.; Bytautas, L. Directed reaction graphs as posets. *MATCH Commun. Math. Comput. Chem*., **2000**, *42*, 261-290.

[33] Došlić, T; Klein, D.J. Splinoid interpolation on finite posets. *J. Comput. Appl. Math*., **2005**, *177*, 175-185.

[34] Ivanciuc, T.; Klein, D.J.; Ivanciuc, O. Posetic cluster expansion for substitution-reaction diagrams and its application to cyclobutane. *J. Math. Chem*., **2007**, *41*, 355-379.

[35] Klein, D.J. Prolegomenon on partial orderings in chemistry. *MATCH Commun. Math. Comput. Chem.*, **2000**, *42*, 7-21.

[36] Kim, M.G. *Affinity, That Elusive Dream*, MIT Press, Cambridge, **2003**, p. 135.

[37] Hässelbarth, W. The inverse problem of isomer enumeration. *J. Comput. Chem.*, **1987**, *8*, 700-717.

[38] Restrepo, G.; Brüggemann, R.; Klein, D. Partially ordered sets: ranking and prediction of substances' properties. *Curr. Comput-Aid Drug*. **2011**, *7*, 133-145.

[39] Restrepo, G.; Klein, D.J. Predicting densities of nitrocubanes using partial orders. *J. Math. Chem.* **2011**, *49*, 1311-1321.

[40] Brüggemann, R.; Patil, G.P. *Ranking and prioritization for multi-indicator systems*, Springer, New York, **2011**.

[41] Restrepo, G.; Brüggemann, R. Dominance and separability in posets, their application to isoelectronic species with equal total nuclear charge, *J. Math. Chem.* **2008**, *44*, 577-602.

[42] Brüggemann, R.; Restrepo, G. Estimating Octanol / Water partition coefficients by order preserving mappings. *Croat. Chem. Acta*. **2013**, 86, 509-517.

[43] Winkler, P. Average height in a partially ordered set, *Discrete Math*. **1982**, *39*, 337-341.

[44] Greene, N.; Fisk, L.; Naven, R.; Note, R.; Patel, M.; Pelletier, D. Developing structure-activity relationships for the prediction of hepatotoxicity. *Chem. Res. Toxicol.* **2010**, *23*, 1215-1222.

[45] Low, L.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuz'min, V.; Fourches, D.; Zhu, H.; Rusyn, I.; Tropsha, A. Predicting drug-induced hepatotoxicity using QSAR and toxicogenomics approaches. *Chem. Res. Toxicol.* **2011**, *24*, 1251-1262.

[46] Rodgers, A.; Zhu, H.; Fourches, D.; Rusyn, I.; Tropsha, A. Modeling liver-related adverse effects of drugs using k-nearest neighbor quantitative structure−activity relationship method. *Chem. Res. Toxicol.* **2010**, *23*, 724-732.

[47] Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform*. **2010**, *29*, 476-488.

[48] Hou, T.; Wang, J. Structure – ADME relationship: still a long way to go? *Expert Opin. Drug Met.* **2008**, *4*, 759-770.

[49] Puzyn, T.; Leszczynski, J.; Cronin, M.T. *Recent advances in QSAR studies: methods and applications (challenges and advances in computational chemistry and physics)*, Springer, Dordrecht, **2010**, *8*, 57-314.

[50] Stanton, D. QSAR and QSPR model interpretation using partial least squares (PLS) analysis. *Curr. Comput-Aid Drug*. **2012**, *8*, 107-127.

[51] Bemis, G.W.; Murcko, M.A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem*. **1996**, *39*, 2887-2893.

[52] Wilkens, S.J.; Janes, J.; Su, A.I. HierS: Hierarchical scaffold clustering using topological chemical graphs. *J. Med. Chem*. **2005**, *48*, 3182-3193.

[53] Xu, J. A new approach to finding natural chemical structure classes. *J. Med. Chem*. **2002**, *45*, 5311-5320.

[54] Ganter, B.; Wille, R. *Formal concept analysis: mathematical foundations*, Springer-Verlag, Berlin, **1998**.

[55] Restrepo, G.; Basak, S.C.; Mills, D. Comparison of SAR and QSAR approaches to mutagenicity of aromatic and heteroaromatic amines. *Curr. Comput-Aid Drug*. **2011**, *7*, 109-121.

[56] Carpineto, C.; Romano, G. *Concept data analysis: Theory and applications*, John Wiley & Sons, Hoboken, **2004**.

[57] Kerber, A. Posets and lattices, contexts and concepts. *MATCH Commun. Math. Comput. Chem*. **2005**, *54*, 551-560.

[58] Quintero, N.; Cohen, I. M.; Restrepo, G. Relating β+ radionuclides' properties by order theory. *J. Radioanal. Nucl. Ch.*, **2013**, *298*, 1937-1946.

[59] Hosoya, H. What can mathematical chemistry contribute to the development of mathematics?. *HYLE – Int. J. Phil. Chem.,* **2013**, *19*, 87-105.

[60] Restrepo, G.; Villaveces, J. L. Chemistry, a lingua philosophica. *Found. Chem.* **2011**, *13*, 2 33-249.

[61] Klein, D.J. Similarity and dissimilarity in posets. *J. Math. Chem*., **1995**, *18*, 321-348.

[62] Restrepo, G. Quantifying complexity of partially ordered sets. In *Multi-indicator systems and modelling in partial order*; Bruggemann, R.; Carlsen, L.; Wittmann, J., Eds.; Springer: Berlin, Germany, **2014**; Chapter 5, 85-106

[63] Butler, D. Academics strike back at spurious rankings. *Nature* **2007**, *447*, 514-515.

[64] Butler, D. Experts question rankings of journals. *Nature* **2011**, *478*, 7367.

[65] Rival, I.; Zaguia, N. Constructing N-free, jump-critical ordered sets. *Congressus numerantium* **1986**, *55*, 199-204.

[66] Brightwell, G.; Winkler, P. Counting linear extensions. *Order* **1991**, *8*, 225-242.

[67] Kemeny, J.G. Mathematics without numbers. *Daedalus* **1959**, *88*, 577-591.

CHAPTER 3

On the Concept for Overall Topological Representation of Molecular Structure

Danail Bonchev*

Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284-2030, USA

Abstract: Graph theory based descriptors of molecular structure play important role in QSPR/ QSAR models. This chapter reviews some attempts to optimize the characterization of molecular structure *via* an integrated representation that accounts in a systemic manner for the contributions of all substructures. In its simplest version this approach counts the subgraphs of all sizes, the resulted single number being shown to be a very sensitive measure of structural complexity. The most complete version builds (i) an ordered set of counts of subgraphs of increasing number of edges, (ii) weights each subgraph with the value of selected graph-invariant, building a weighted ordered set, and (iii) sums up all the subgraph contributions to produce the overall value of the graph-invariant. The invariants tested include vertex degrees, vertex distances, and the graph non-adjacency numbers, the corresponding overall topological indices being called overall connectivity, overall Wiener, overall Zagreb and overall Hosoya indices. Their properties are analyzed in detail in acyclic and cyclic graphs. It is shown that they all are reliable measures of molecular structural complexity, increasing in value with the basic complexifying patterns of branching and cyclicity of molecular skeleton. The structure-property models derived for 10 physicochemical properties of alkane compounds show considerable improvement compared to models derived from molecular connectivity indices. The latest extension of these ideas is demonstrated with extended connectivities, walk counts, and Bourgas indices, the latter of which are the first integrated measures of graph complexity and vertex centrality.

Keywords: Molecular structure, molecular topology, molecular descriptors, graph theory, topological indices, overall connectivity, overall Wiener index, overall Zagreb indices, overall Hosoya index, Bourgas indices, structural complexity, structure-property models, vertex centrality, molecular branching, molecular cyclicity.

***Corresponding author Danail Bonchev:** Center for the Study of Biological Complexity, Virginia Commonwealth University, P.O.BOX 842030, Richmond, VA 23284, USA; Tel: 804-827-7375; Fax: 804-828-1961; Email: dgbonchev@vcu.edu

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50003-5

INTRODUCTION

Molecules of chemical compounds exist in a stunning variety of shapes described by their 3D-geometrical structure. A simplified but very effective way to characterize molecular structure is based on graph theory [1, 2]. Although using a 2D-representation, molecular graphs retain the most essential structural information of molecules - the manner in which the atoms are connected. This topological structure correlates amazingly well with the physicochemical properties (QSPR) and biological activities (QSAR) of chemical compounds [3-5], a finding that stimulated greatly the development of chemical graph theory [6,7] since the beginning of the 1970s [8-11]. A large number of graph-invariants (called also topological indices or topological descriptors) have been identified and applied to the structural analysis of chemical properties [12-14]. The quantitative structure-activity relationships (QSAR) became a basic tool in the area of drug discovery [15, 16], a trend additionally reinforced with the development of the high-throughput methods for identifying lead compounds in drug design [17, 18]. The software tools created for calculating topological descriptors of molecules [19, 20] have also contributed to the rapid development of this area of applied graph theory.

FROM SIMPLE GRAPH-INVARIANTS TO A MORE GENERAL REPRESENTATION OF MOLECULAR TOPOLOGY

All seemingly chaotic development of the world of topological indices raises challenging questions for theoreticians. Is there a "best graph-invariant", a "magic bullet" that would capture in a single number the most essential topological features of molecules? If "yes", is there a "magic" mathematical function that would produce descriptor(s) highly correlating with physicochemical properties and biological activities of chemical compounds?

It is natural in such a search to proceed from the most basic graph matrices - adjacency matrix **A** and distance matrix **D**. These matrices generate both local and global invariants. The simplest local invariants are the vertex degree a_i - the number of the vertex nearest neighbors, and the node distance d_i - the sum of the distances from the given vertex *i* to all other vertices in the graph *G* (the distance between a pair of vertices being defined as the number of edges along the shortest path connecting these vertices). Summing up the values of these local invariants over all vertices *V*, one defines the *total adjacency*, *A*, and the *total distance*, *D*, respectively:

$$A(G) = \sum_{i=1}^{V} a_i \; ; \; D(G) = \sum_{i=1}^{V} d_i \qquad \text{(1a; 1b)}$$

While providing important information on the totality of connectivity and distances in the graph, these indices have the pitfall of being (highly) degenerate, *i.e.*, providing the same value for different molecular structures. Thus, total adjacency has the same value for all molecular graphs having the same number of vertices and cycles, while the capacity of the total distance to discriminate isomers reduces rapidly with the increase of the number of atoms (Fig. **(1)**).

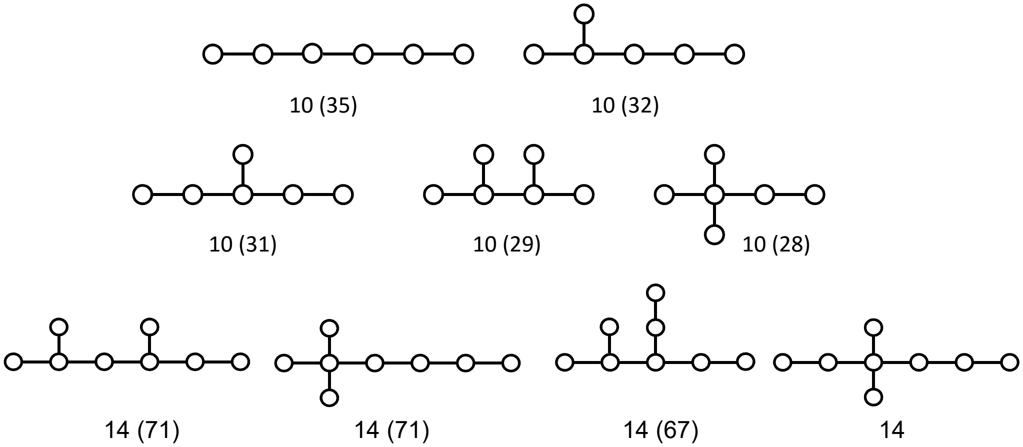


Figure 1: Examples for the low discriminatory power of the total adjacency A of a graph, and the more discriminative but still degenerate Wiener number given in parenthesis. (The Wiener number [21, 22] is half of the total distance in undirected graphs).

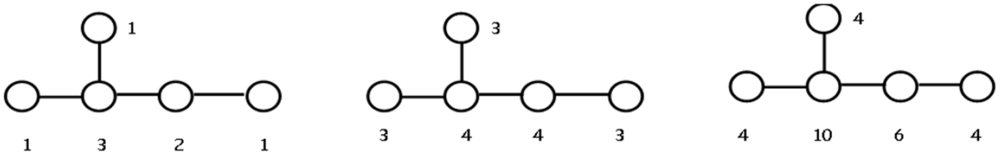
Different approaches have been used to reduce the degeneracy of topological indices. The earliest attempt of this kind was done by Morgan [23], who was searching for unique signature ID for molecules for the purposes of chemical documentation. The *extended connectivity* of Morgan is calculated by a simple iterative algorithm. The null iteration calculates vertex degrees, then in the first iteration each vertex is assigned an extended connectivity index calculated as the sum of the degrees its nearest neighbors have in the null iteration, and this mechanism is repeated until obtaining the maximum possible diverse values of vertex extended connectivity allowed by the symmetry of the graph. Another definition (Razinger [24]) presents the k^{th} extended connectivity of a vertex i as the sum over the i^{th} row elements of the k^{th} degree of adjacency matrix, $\mathbf{A}^{\mathrm{k}}$ (eq. 2a):

$$ {}^{k}EC(i) = \sum_{jadji} {}^{k-1}a_j; \quad {}^{k}EC(i) = \sum_{j=1}^{V} a_{ij}(A^k) \tag{2a,b} $$

As illustrated in Fig. **(2)**, the Morgan algorithm in some cases cannot reach an end and oscillates between two alternating assignments. Despite its failure to provide unique ID for each molecule, Morgan's approach contains a valuable idea - to create a more comprehensive topological representation of the molecular structure by extending local connectivity to a series of partially extended connectivities ${}^{k}EC(G)$:

$$ {}^{k}EC(G) = \sum_{i=1}^{V} {}^{k}EC(i) \tag{2c} $$

The series ends in a term in which connectivity is extended till the most distant neighborhood of any vertex in the graph, which corresponds to the largest distance in graph G, $k = d(max) = diam(G)$. Other realizations of this idea will be analyzed in the last Section.



$$ {}^{1}\mathrm{A(G)} = \left|\begin{array}{ccccc|l} 0&0&1&0&0&1 \\ 0&0&1&0&0&1 \\ 1&1&0&1&0&3 \\ 0&0&1&0&1&2 \\ 0&0&0&1&0&1 \end{array}\right. \qquad {}^{2}\mathrm{A(G)} = \left|\begin{array}{ccccc|l} 1&1&0&1&0&3 \\ 1&1&0&1&0&3 \\ 0&0&3&0&1&4 \\ 1&1&0&2&0&4 \\ 0&0&1&0&1&3 \end{array}\right. \qquad {}^{3}\mathrm{A(G)} = \left|\begin{array}{ccccc|l} 0&0&3&0&1&4 \\ 0&0&3&0&1&4 \\ 3&3&0&4&0&10 \\ 0&0&4&0&2&6 \\ 1&1&0&2&0&4 \end{array}\right. $$

Figure 2: The Morgan iterative algorithm [23] recalculates at each step k the vertex degree of each vertex i as the sum of the degrees of the vertex nearest neighbors in the (k-1)$^{\text{th}}$ iteration step (eq. 2a). Razinger's algorithm (eq. 2b) [24] calculates vertex extended connectivities as sums of the rows in the adjacency matrix k^{th} powers.

A major source of the degeneracy of molecular topological indices is their definition as *a sum* of the values of the corresponding local (mostly vertex) graph-invariant. A natural way of reducing this degeneracy is to probe a more sensitive mathematical operation. This has been done in the first and second Zagreb indices, M1 and M2, using squared degrees of all vertices, and products of vertex degrees of all pairs of adjacent vertices, respectively [25]. Randić modified the M2 index by using the inverse square root of the product of vertex degrees to define his highly discriminative branching index χ [26]:

$$M1(G) = \sum_{i=1}^{V} a_i^{\;2} \text{ (3a); } M2(G) = \sum_{i,j-adj} (a_i a_j) \text{ (3b); } \chi(G) = \sum_{i,j-adj} (a_i a_j)^{-1/2} \qquad \text{(3c)}$$

An essential conceptual improvement of these attempts was advanced by Kier and Hall's *molecular connectivity* approach [27-29]. The latter generalized Randić's function (3c), renamed as *first-order molecular connectivity*, to a series of indices of increasing order *t* = 0, 1, 2, 3,. which includes a generalized product of degrees of a variable number of adjacent vertices:

$${}^{t}\chi(G) = \sum_{a,j,k,..adj} (a_i a_j a_k ...)^{-1/2} \qquad (4)$$

The advantage of molecular connectivity (MC) concept is that it makes use of molecular fragments (or subgraphs in terms of graph theory) of increasing size, beginning with isolated vertices (t=0), edges (t=1), two-edge fragments (t=2), *etc.* This might be considered as a generalization of the pioneering ideas of Smolenskii [30] and Gordon [31], who first used molecular subgraphs for a systematic characterization of molecular properties. The Molecular Connectivity concept provided the basis for successful QSPR and QSAR models, and thus became an important component of drug discovery process [32]. As analyzed in occasion of the 25th anniversary of this concept [33], the great success of eq. (4) in modeling physicochemical properties and biological activities of chemical compounds does not result from the well discriminating branching index χ, but from the better representation of molecular topology using subgraphs of increasing size and shape. (The inverse-square-root function (eq. 4) was shown to provide slightly inferior statistics compared to some other exponent values in modeling molecular properties). Several different molecular shapes, denoted as path, cluster, path-cluster, and cycle have also been discerned within molecular connectivity approach, improving further the statistics of QSPR/QSAR molecular connectivity models (Fig. **(3)**).

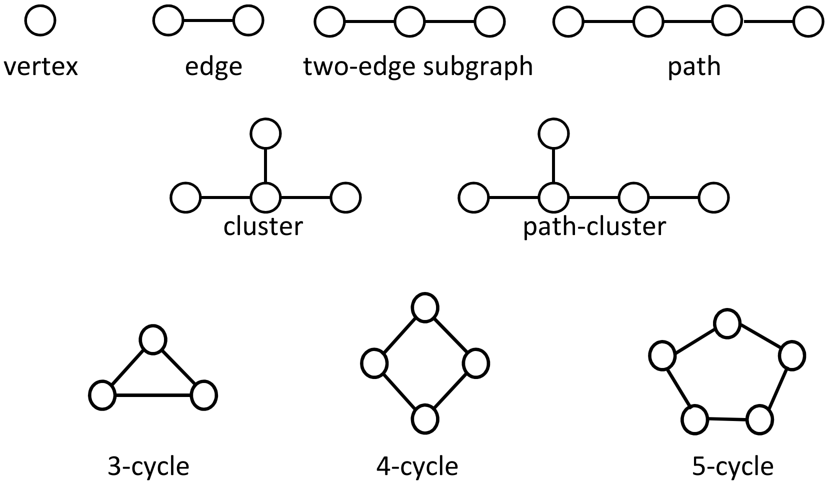


Figure 3: Subgraphs of increasing sizes and variety of shapes used in molecular connectivity approach: a) vertices, b) edges, c) two-edge subgraphs, d) three-edge subgraphs: path and cluster, e) four edge subgraphs: path, path-cluster, cluster, and n-cycle.

Another approach aimed at increasing the discriminating capacity of topological descriptors has been based on the mathematical framework of information theory [34, 35]. Redefined for a graph of *V* vertices, partitioned into *k* classes of V_1, V_2,., V_k vertices, this approach [36] makes use of the Shannon equation interpreted as equation for the mean and total information content $\bar{I}(G)$ and $I(G)$, respectively:

$$\bar{I}(G) = -\sum_{i=1}^{k} p_i \log_2 p_i = -\sum_{i=1}^{k} \frac{V_i}{V} \log_2 \frac{V_i}{V}, bits / vertex \tag{5a}$$

$$I(G) = Vx\bar{I} = V \log_2 V - \sum_{i=1}^{k} V_i \log_2 V_i, \ bits \tag{5b}$$

Here, $p_i = V_i / V$ is the probability of a randomly selected graph vertex to belong to the class *i* having V_i vertices. A more complete representation of molecular topology can be achieved within this framework as a vector termed *information-*

theoretic superindex [36, 37]. The latter contains a set of diverse topological descriptors translated into the information theory language:

$$SI = \{I_{orb}, I_{chr}, I_C, I_{edge}, I_D, I_Z\} \tag{6}$$

where the six supervector components represent the information on vertex distribution into the orbits of the automorphisms group of the graph, into the graph chromatic classes, and into the centrically ordered classes, as well as the information on the edge degree distribution, the graph distances distribution according to their magnitudes, and the partition of the Hosoya number *Z* into classes of nonadjacent vertices. The superindex expresses in the language of information theory six different aspects of molecular topology, those of symmetry, chromaticity, centrality, connectivity, distances, and nonadjacency, and provides an extended basis for structure-property and structure-activity correlations.

TOPOLOGICAL COMPLEXITY AS A GUIDE IN THE SEARCH FOR A GENERALIZED TOPOLOGICAL CHARACTERIZATION OF MOLECULAR STRUCTURE

The idea of overall topological indices, describing more adequately molecular structure, crystalized in mid 1990s from the search for reliable measures of molecular complexity. The first attempts to measure complexity of systems (human body, living cell, molecules) have been done in the 1950s by a group of US scientists related to the journal "Bulletin of Mathematical Biophysics" (later renamed to "Bulletin of Mathematical Biology"). Applying the then-recently developed information theory, Dancoff *et al.* [38] proposed to use the *information content* of a system as a measure of its complexity. While different aspects of complexity have been considered, the *topological information content* I_{top} introduced by Rashevsky in 1955 combined the information on the chemical nature of atoms with the atom's equivalence based on identical neighborhood relationships [39]:

$$\bar{I}_{top} = -\sum_{i=1}^{k} \frac{N_i}{N} \log_2 \frac{N_i}{N}, bits / atom \tag{7}$$

Here, N is the total number of atoms in the molecule, N_i is the number of atoms in class i having the same chemical nature and the same atomic neighborhoods at distance 1, 2,., d_{max} (See Fig. **(4)**. The topological aspects of Rashevsky's index have been reformulated by Trucco [40] in terms of the orbits of the automorphism group of the graph.

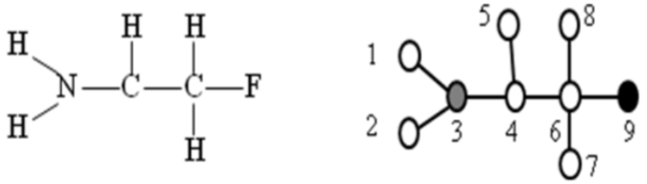


Figure 4: Illustration of the topological information of Rashevsky [39]. The nine vertices are partitioned into 5 classes of one vertex (## 3,4,5,6,9) and two classes of two vertices ({1,2} and {7,8}) equivalent by both chemical nature and identical neighborhood. The average topological information of this molecule calculated by eq. (7) is 2.73 bits per atom.

The next step down the road of measuring complexity was done by Mowshowitz, who analyzed in 1968 the relative complexity of undirected and directed graphs based on their topological information content [41]. Minoli in 1975 was the first to assess complexity of graphs without resorting to information theory. He constructed an index (eq. 8a) to measure the *combinatorial complexity* of graphs, based on the number of vertices, edges and paths of the graph [42]. To reduce strongly the degeneracy of the Minoli index, MI, Bonchev modified it replacing the number of graph paths ΣP_i by the paths total length ΣL_i (eq. 8b) [43, 44]:

$$MI = \frac{V\,x\,E}{V+E}\sum_i P_i \tag{8a}$$

$$MB = \frac{V\,x\,E}{V+E}\sum_i L_i \tag{8b}$$

In 1980, and in more detail in 1987, Bonchev *et al.* used the number of spanning trees to measure complexity of cyclic graphs [45, 46], an approach also developed independently by Mallion *et al.* in 1983 and 1998 [47, 48]. In 1981, Bertz used the Shannon information equation to design a sensitive index of molecular structural complexity [49]. Instead of graph vertices and edges, the Bertz Index *BI* grouped all two-edge subgraphs n in equivalence classes of cardinality n_i (eq. 9):

$$BI = 2n\log_2 n - \sum n_i \log_2 n_i, bits \tag{9}$$

This allowed not only to reduce the complexity index degeneracy, but also to mirror some of the complexifying structural patterns.

Following the founding study of Minoli [42], the criteria for a mathematical function to be a complexity measure have been further developed and discussed by different authors during the next 25 years [43, 50-55]. A point of agreement between all of them is that topological complexity measures should follow the complexifying structural patterns of branching [56-59] and cyclicity [60-63]. This will be illustrated in the next sections with examples of the latest and most comprehensive measures proposed since the beginning of the 1990s.

In what follows we present the concept of overall topological indices, which develops further the ideas for a more complete topological characterization of molecular structure by accounting for subgraphs of increasing size *weighted* by some of the basic graph-invariants. It will be shown that this integrated approach orders molecular graphs according to their increasing complexity, and provides QSPR/QSAR models with high statistics. The latest of the indices shown derives overall graph complexity from centrality of the graph vertices, which in turn is defined from both their connectivity and distances. Other complexity measures sharing the property of detailed characterization of molecular topology will also be analyzed.

THE CONCEPT FOR OVERALL TOPOLOGICAL DESCRIPTORS OF MOLECULAR STRUCTURE

During the second part of the 1990s, Bonchev [64-66] and Bertz [67, 68] independently and almost simultaneously developed in detail the idea to use the total subgraph count, $SC(G)$, as a measure of *complexity* of graph G. (The possibility of using the total subgraph count as a measure of molecular *similarity* has been briefly mentioned earlier by Bertz and Herndon [69]). Bertz applied the SC index to the synthesis planning in organic chemistry, while Bonchev represented the total subgraph count as an ordered set $\{SC\}$ of counts of subgraphs having an increasing number of edges, and applied it to QSPR/QSAR.

The set begins with the number of isolated vertices V, regarded as a null-order index, ^{0}SC, followed by the number of edges E, as a first-order index, ^{1}SC, the two-edge subgraphs, as a second-order index, ^{2}SC, *etc.*:

$$SCG = {}^{0}SCG + {}^{1}SCG + {}^{2}SCG + ... + {}^{E}SCG; \ \{SCG\} = \{{}^{0}SCG, {}^{1}SCG, {}^{2}SCG, ..., {}^{E}SCG\} \quad \text{(10a,b)}$$

In addition to counting the subgraphs, Bonchev proposed [33, 65, 66, 70] to weight each of the subgraphs i with the value of its total adjacency A_i, defined as the sum of the entries of the subgraph adjacency matrix. By summing up the adjacencies $A_i(^{e}G_i)$ of all e^{th}-order subgraphs $^{e}G_i$, with $e = 0, 1, 2, 3, …, E$, one defines the *overall connectivity* $OC(G)$ of the graph G with its null-, first-, second, *etc.* order components $^{e}OC(G)$, and the *overall connectivity vector* $\{OC\}$:

$$OCG = {}^{0}OCG + {}^{1}OCG + {}^{2}OCG + ... + {}^{E}OCG; \ \{OCG\} = \{{}^{0}OCG, {}^{1}OCG, {}^{2}OCG, ..., {}^{E}OCG\} \quad \text{(11a,b)}$$

The overall connectivity index was initially called *topological* complexity and defined in two versions, *TC* and *TC1*, differing in the vertex degrees used - those in the parent graph G or those in the subgraphs, respectively. The first approach was adopted as better from the point of view of systems theory and applied later as overall topological representation of molecular structure to other graph-invariants. Included here were the *overall Wiener index*, $OW(G)$ [71], the *overall Zagreb indices*, *OM1* and *OM2* [72], and the *overall Hosoya topological index*, $OZ(G)$ [73]. As well known, the Wiener index W [21, 22] is the half-sum of all graph distances, d_{ij}, whereas the Hosoya $Z(G)$ index [74] is the sum of all graph matchings (sets of edges without common vertices). Generalized definitions of the overall topological indices are given below, with graph-invariant X standing for any of the above listed invariants: subgraph count SC, connectivity C, Wiener index W, Zagreb indices $M1$ and $M2$, and Hosoya index Z. Besides the partitioning into ordered set according to the number of edges in the subgraphs, a more detailed grouping of subgraphs according to their specific shapes, such as paths, clusters, path-clusters, and cycles (proposed earlier by Kier and Hall in their classical molecular connectivity concept [27, 28]) was also adopted.

Viewing the overall connectivity concept as an extension of the earlier ideas of Kier and Hall, (other extensions have been proposed, to mention just few here

[75, 76]) this concept was termed "next generation molecular connectivity" [33]. The genetic link between the two concepts was also extended beyond topological complexity, measured by overall topological indices. In order to account for the different chemical nature of molecular graph vertices for potential QSPR/QSAR applications, the simplest among a variety of ways is to follow the valence connectivity approach of Kier *et al.* [29, 77], as shown below in Definition 5. The *overall valence connectivity index OVC* introduced in this definition emerges thus as a measure of molecular complexity, which integrate the information on molecular topology with that on the elemental composition of molecules.

Definition 1: The overall topological index *OX(G)* of any graph *G* is defined as the sum of the values of graph-invariant $X_i(G_i)$ of all *K* subgraphs of *G*:

$$OX(G) = \sum_{i=1}^{K} X_i(G_i \in G) \tag{12}$$

Definition 2: The e^{th}-order overall topological index $^{e}OX(G)$ of any graph *G* is defined as the sum of the values of graph-invariant $X_j\,(^{e}G_j)$ of all ^{e}K subgraphs $^{e}G_j \in G$ that have *e* edges:

$$^{e}OX(G) = \sum_{i=1}^{^{e}K} X_i(^{e}G_i \in G) \tag{13}$$

Definition 3: The e^{th}-order overall topological index $^{e}OX(G)$ can be partitioned into a sum of terms, $^{e}OX_k(G)$, representing the sum of the values of the vertex graph-invariant *X* in subgraphs of specified type *k* having the same number of *e* edges. For acyclic graphs (as in molecular connectivity approach of Kier *et al.* [27, 28], the subgraph types are *path* (k = p), *cluster* (k = c), and *pathcluster* (k = pc) type, while for cyclic graphs these are *n-cycles* with *n* being the cycle size:

$$\begin{aligned} {}^{e}OX(G) &= {}^{e}OX_p(G) + {}^{e}OX_c(G) + {}^{e}OX_{pc}(G) + {}^{e}OX_{n-c}(G) = \\ &= \sum_{j=1}^{^{e}K_p} X_{p,j} + \sum_{l=1}^{^{e}K_c} X_{c,l} + \sum_{m=1}^{^{e}K_{pc}} X_{pc,m} + \sum_{r=1}^{^{e}K_{n-c}} X_{n-c,r} \end{aligned} \tag{14}$$

Definition 4: The overall topological vector *OXV(G)* of any graph *G* is the sequence of all $^{e}OX(G)$s listed in an ascending order of the number of edges *e*:

$$\mathrm{OXV(G) = OX\ \{{}^{o}OX,\ {}^{1}OX, {}^{2}OX, \ldots, {}^{E}OX\}} \tag{15a}$$

or in more detail for the type of graphs $k \equiv p$ (path) or $k \equiv c$ (cluster) or $k \equiv pc$ (path-cluster) or $k \equiv n\text{-}c$ (cycle of size n):

$$\mathrm{OXV\ (G) = OX\ \{{}^{o}OX,\ {}^{1}OX, {}^{2}OX,\ {}^{3}OX_{p},\ {}^{3}OX_{c},\ {}^{3}OX_{3\text{-}c}, \ldots,\ {}^{E}OX_{p},\ {}^{E}OX_{c},\ {}^{E}OX_{pc}, {}^{E}OX_{n\text{-}c}\}} \tag{15b}$$

Definition 5: The *overall valence connectivity index* $OVC^{v}(G)$ of graph G is defined as the sum of the total valence adjacencies ${}^{e}A_{k}^{v}$ $({}^{e}G_{k})$ of all ${}^{e}K$ subgraphs ${}^{e}G_{k}$ of G having e edges and N(i) vertices:

$$OVC^{v}(G) = \sum_{e=1}^{E} {}^{e}OVC^{v}({}^{e}G) = \sum_{e=1}^{E}\sum_{k=1}^{K(e)} {}^{e}A_{i}^{\,v}({}^{e}G_{i}) = \sum_{e=1}^{E}\sum_{k=1}^{{}^{e}K}\sum_{i=1}^{N(i)} a_{i}^{\,v}({}^{e}G_{i}) \tag{16}$$

where a_{i}^{v} is the valence term of Kier and Hall [29, 77] replacing the vertex degree a_{i}.

Definition 6: The average overall index per vertex, $X_{a}(G)$, and the normalized overall index $X_{n}(G)$, $0 \leq X_{n}(G) \leq 1$ are defined as:

$$X_{a}(G) = \frac{X(G)}{V}; \quad {}^{e}X_{a}(G_{e}) = \frac{{}^{e}X(G_{e})}{V} \tag{17a,b}$$

$$X_{n}(G) = \frac{X(G)}{X(K_{V})}; \quad {}^{e}X_{n} = \frac{{}^{e}X(G)}{{}^{e}X(K_{V})} \tag{18a,b}$$

where K_{V} stands for the complete graph having the same number of vertices V with G.

Definition 7: C*umulative* p^{th} order overall indices, ${}^{p}X(G)$, are defined [78] as the sum of the values of the ${}^{e}X(G)$ indices (Definition 2) for e = 1, 2, …, p:

$${}^{p}X(G) = \sum_{e=0}^{p} {}^{e}X(G) \tag{19}$$

The overall indices defined by Definitions 1-4 and 6 were introduced for a more complete topological characterization of molecular structure. The rapid development of network theory after the year 2000 [79-83] imposed some modification of the concept for applications to networks. Due to the very large size of the majority of networks, particularly those in biology and social sciences, the requirement to account for all subgraphs would lead for these networks to combinatorial explosion. The modification proposed in Definition 7 [78] limits the size of the subgraphs to such having no more than a predefined number of edges *p*, which for the complex networks having over a hundred nodes should not be larger than 3, thus making use of only first-, second-, and third-order overall complexity indices. This proved to be computationally feasible and sufficient for assigning different ID numbers to networks with the same number of nodes, since there is no degeneracy problem for large complex networks [78, 84]. The cumulative indices could also be used as additional descriptors in case of large molecules as well.

FORMULAS FOR THE OVERALL TOPOLOGICAL INDICES OF SOME CLASSES OF GRAPHS

Some classes of molecular structures, such as n-alkanes, monocyclic compounds, and others have a regular structure, which allows deriving analytical expressions for overall topological indic*es. Presented below are such expressions for subgraph count, SC, overall connectivity, OC*, and overall Wiener *OW* indices [70, 71, 78]. The parameters used are the total number of vertices *n*, the total number of edges *q*, and the number of edges e in the given class of subgraphs.

Path graphs (P_n)

$$^{e}SC(P_n) = n - e; \quad SC(P_n) = n(n+1)/2 \tag{20a,b}$$

$$^{e}OC(P_n) = 2[q(e+1) - e^2]; \quad OC(P_n) = n(n-1)(n+4)/3 \tag{21a,b}$$

$$^{e}OW(P_n) = e(e+1)(n-e)/6; \quad OW(P_n) = (n+3)(n+2)(n+1)n(n-1/120 \tag{22a,b}$$

Cycles (C_n)

$$^{e}SC(C_n) = n; \quad SC(C_n) = n^2 + 1 \tag{23a,b}$$

$$^{e}OC(C_n) = 2n(e+1);\quad OC(C_n) = n(n^2 + n + 2) \tag{24a,b}$$

$$\begin{aligned} &^{e}OW(C_n) = e(e+1)(e+2)n/6fore<n \\ &OW(C_n) = (n^5 + 2n^4 + 2n^3 - 2n^2 - an)/24;\quad a(n\,even) = 0;\quad a(n\,odd) = 3 \end{aligned} \tag{25a,b}$$

Star Graphs (S_n)

$$^{e}SC(S_n) = \binom{q}{e};\quad SC(S_n) = 2^q + q \tag{26a,b}$$

$$^{e}OC(S_n) = (q+e)\binom{q}{e}; OC(S_n) = 2^q + \sum_{e=1}^{q}(q+e)\binom{q}{e} \tag{27a,b}$$

$$^{e}OW(S_n) = e^2\binom{q}{e};\quad OW(S_n) = \sum_{i=0}^{n-2}(n-i-1)^2\binom{n-1}{i} \tag{28a,b}$$

To facilitate the application of the overall indices introduced in Section 3 as measures of the network topological complexity to complex networks, equations were derived for the first several terms of these indices for complete graphs as a function of the number of graph vertices *n*, thus enabling the calculation of the normalized complexity indices within the 0 to 1 range.

$$^{1}OC(K_V) = n(n-1)^2 \tag{29}$$

$$^{2}OC(K_V) = \frac{3}{2}n(n-1)^2(n-2) \tag{30}$$

$$^{3}OC(K_V) = \frac{1}{6}n(n-1)^2(n-2)(16n-45) \tag{31}$$

$$^{3}OC(K_V, triangle) = \frac{1}{2}n(n-1)^2(n-2) \tag{32}$$

$$^{3}OC(K_V, linear) = 2n(n-1)^2(n-2)(n-3) \tag{33}$$

$$^{3}OC(K_V, star) = \frac{2}{3}n(n-1)^2(n-2)(n-3) \tag{34}$$

OVERALL TOPOLOGICAL INDICES CAPTURE THE PATTERNS OF INCREASING MOLECULAR COMPLEXITY

Table **1** below presents the values of six overall topological indices, SC, OC, OW, OM1, OM2, and OZ for the acyclic graphs depicting in Fig. **(5)** the carbon skeleton of alkane molecules having 2 to 7 carbon atoms. The table shows that there are no degenerate overall indices values within the groups of isomeric alkanes, in contrast to the values of graph-invariants used in the design of overall indices. Thus, the total adjacency has the same value for all graphs with the same number of vertices, whereas for the set of 21 acyclic graphs the same Wiener number values are found for two pairs, two triplets, and even one quadruplet of structures.

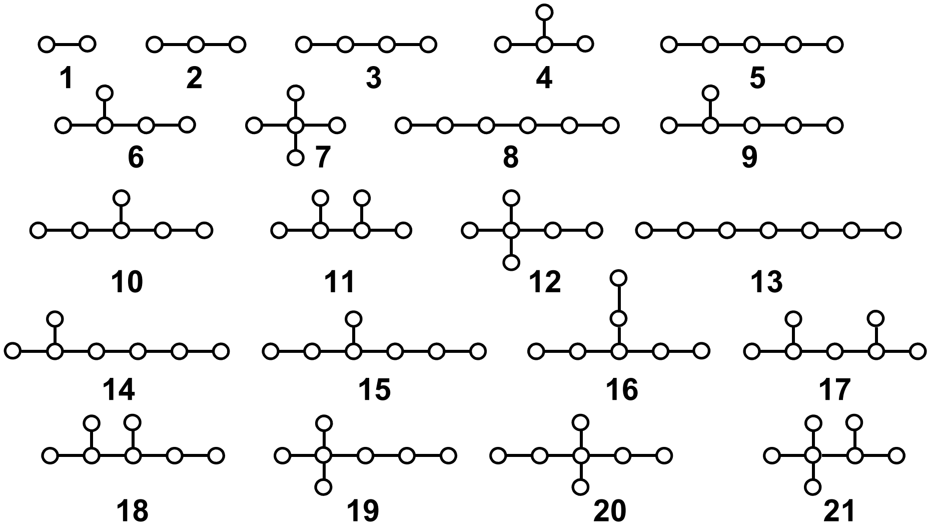


Figure 5. Acyclic graphs depicting hydrocarbon molecules with two to seven carbon atoms used for illustrating structural patterns of increasing complexity captured by the overall topological indices.

Table **1** and Fig. **(5)** demonstrate that the patterns of increasing complexity in acyclic graphs having up to 7 vertices are mirrored adequately by increasing values of the overall descriptors of topological complexity (OI). These branching patterns analyzed in details earlier [56, 59] evidence that at a constant number of nodes *n* the complexity of acyclic graphs increases with the increase in:

(i) *Number of branches*: linear < monobranched < dibranched < tribranched < tetrabranched.

Examples (See Table **1**): OI(21) > OI(17-20) > OI(14,15) > OI(13); OI(11,12) > OI(9,10) > OI(8); OI(7) > OI(6) > OI(5).

Table 1: Overall complexity values of graphs 1 – 21, representing the carbon skeleton of alkane molecules with 2-7 atoms. Included are the subgraph count, *SC*, the overall connectivity, *OC* [70], the overall Wiener index, *OW* [71], the overall Zagreb indices *OM*1 and *OM*2 [72], and the overall Hosoya index, *OZ* [73]. All overall indices completely discriminate the molecules with the same number of atoms

Graphs	***SC***	***OC***	***OW***	***OM1***	***OM2***	***OZ***
1	3	4	1	1	1	4
2	6	14	6	22	10	10
3	10	32	21	56	26	21
4	11	39	24	87	27	23
5	15	60	56	110	60	40
6	17	76	67	168	67	46
7	20	100	80	292	68	52
8	21	100	126	188	130	72
9	24	127	154	277	149	84
10	25	136	161	300	161	89
11	28	164	188	404	172	100
12	30	181	197	505	168	103
13	28	154	252	294	272	125
14	32	194	311	418	315	147
15	34	214	333	468	351	159
16	36	234	354	516	390	172
17	37	246	384	584	366	173
18	40	276	411	668	410	191
19	41	284	414	762	370	185
20	44	314	440	850	412	202
21	49	369	510	1075	433	225

(ii) *Branch centrality*: 2M < 3M < 4M < …; 2,2MM < 3,3MM < …

Here and below, branches of length 1 and 2 are denoted by letters *M* for methyl and *E* for ethyl, respectively. Examples: OI(10) > OI(9); OI(15 > OI(14); OI20) > OI(19).

(iii) *Branch length*: 3M < 3E < … Example: OI(16) > OI(14,15). Also for the graphs of octane molecules: 2,3MMC6 < 2,3MEC5; 3,3MMC6 < 3,3MEC5.

(iv) *Branch multiplicity* (number of branches attached to a vertex): OI(19,20) > OI(17,18); OI(12) > OI(11). 2,3MMC4 < 2,2MMC4; 2,3MMC5 < 2,2MMC5.

(v) *Branch clustering* (grouping of vertices to closer located chain vertices): OI(2,5MMC6) < OI(2,4MMC6) < OI(2,3MMC6);

Besides these five basic branching patterns, there are more complex cases in which two or more patterns are combined. The relative importance of branching factors in such cases varies with the increase of the number of atoms. The latter increases the role of branch centrality, which at higher number of atoms becomes dominant relative to branch multiplicity (*e.g.*, 2,3MMC6 > 2,2MMC6, compared to 2,3MMC5 < 2,2MMC5). The branch multiplicity pattern holds only when a branch is shifted to a vertex with equivalent or higher centrality. Examples of two pairs of graphs of alkanes having eight carbon atoms: OI(3,3MMC6) > OI(3,4MMC6); OI(2,3,4MMMC5) > OI(2,2,4MMMC5). The branch centrality in larger graphs becomes a dominant factor even to the number of branches (first examples appear in alkanes having more than eight carbon atoms).

All six overall indices given in Table **1** obey these five patterns of increasing complexity in series of acyclic graphs having the same number of vertices, with a single exception, OM2, which do not follow pattern (iv). When the number of graph vertices increases, there are no clear trends but more and more highly branched graphs with a smaller number of vertices become characterized by larger values of the overall indices than larger graphs with simpler topology. Thus, three of the indices examined (SC, OC, and OM1) have lower values for linear graph corresponding to n-C7 alkane (#13 in **Fig.(5)**) than the most branched C6-graph (#12). The overall Wiener index reaches this value inversion for n-C8 with respect to 2,2,3MMMC4, while for OM2 and OZ indices such inversion occurs at higher numbers of carbon atoms.

Complexity trends in cyclic molecules (discussed in detail earlier as "cyclicity rules" [60-62]) are briefly analyzed below with a representative sample of 16 cyclic graphs having five vertices shown in Fig. **(6)**.

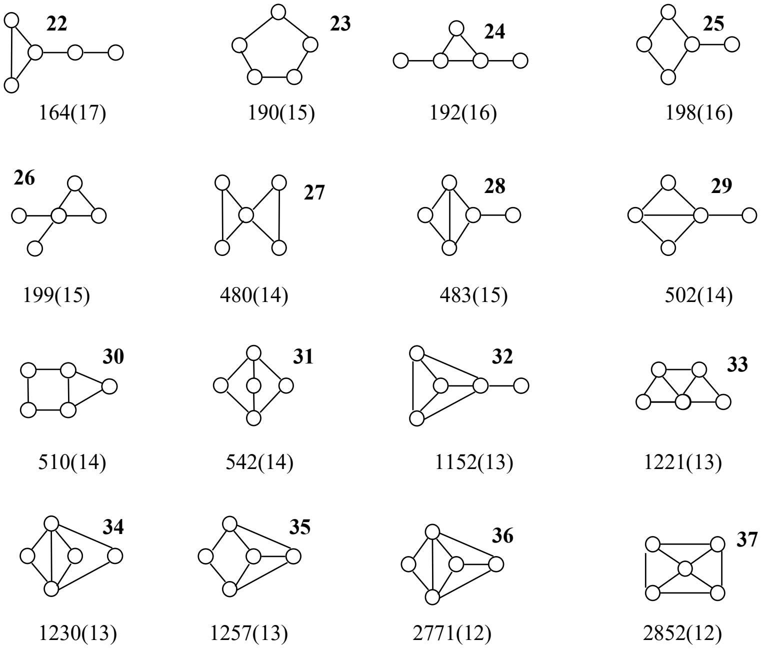


Figure 6: Illustration of complexity trends in structures containing cycles, as characterized by the overall Wiener index (the highly degenerate values of the original Wiener index are given for comparison in parentheses).

Two basic complexity patterns are clearly manifested. The topological complexity of cyclic molecules increases with:

(i) The number of cycles:

OW(**22-26**) < OW(**27-31**) < OW(**32-35**) < OW(**36,37**)

(ii) The degree of connectedness of a pair of cycles:

OW(**27**) < OW(**28-30**) < OW(**31**)

This pattern which shows the increase in complexity in pairs of cycles connected by a common vertex, common edge and two common edges is of specific interest in comparing kinetics of reaction mechanisms [46, 84]. The subgraph count *SC* and overall connectivity *OC* increase in value with the number of cycles, while the pattern of increasing complexity with the stronger connectivity of two cycles is captured by the overall Wiener index only.

A third pattern, that of complexity increasing with the increase in cycle size at a constant number of vertices is again identified by the overall Wiener index, *e.g.* when comparing tricyclic structures: OW(**32, 33**) < OW(**34, 35**). The same pattern is also observed in the bicyclic structures, although in these cases it acts jointly with pattern (ii): OW(**27-29**) < OW(**30,31**). More subtle cases again appear when several complexity factors are involved. Such is the case with the five monocyclic structures **22-26** for which the cycle size is intermingled with branching factors - number of branches and multiplicity of branches at the same vertex. The presence of acyclic branch(es) also mixes with cyclicity factors in tricyclic graphs OW (**32**) < OW(**33-35)**, indicating again the need of more detailed complexity studies of molecules containing cycles.

OVERALL TOPOLOGICAL INDICES (*OI*) PROVIDE A BASIS FOR HIGH STRUCTURE-PROPERTY AND STRUCTURE-ACTIVITY CORRELATIONS

Using the carbon skeleton of alkane molecules as a basis for testing our integrated representation of molecular topology, we modeled ten alkane physicochemical properties by the overall topological indices *OI*. The properties modeled were boiling points [65], T_B in °C; critical temperatures [85] T_c in °C; critical pressures [85] P_c in atm; critical volume V_c in [86] L/mole; molar volume V_m in cm^3/mol [28]; molecular refraction R_m in cm^3/mol [54]; surface tension [86] *ST* in dyn/cm; the heat of formation in gaseous state, $\Delta H_f(g)$ in kJ/mol [87]; the heat of vaporization, ΔH_v in kJ/mol [87]; the heat of atomization, ΔH_a in kcal/mol [28].

The values of the molecular connectivity indices were taken from Kier and Hall's monograph [28]. While complete details of the modeling, including the values of all parameters and properties used can be found in our previous publications [66, 71, 78],

Table **2** presents the performance of the different *OI*s by the standard deviation of the best five-parameter models obtained. The overall Wiener index *OW* and the two Zagreb Indices *OM1* and *OM2* were used, along with the composite overall connectivity index *OC**, which includes all *SC* and *OC* terms defined by eqs. (12, 13). They were compared with the best molecular connectivity χ model to demonstrate the level of improvement achieved by the overall topological indices, which are based on a more complete topological representation of molecular structure.

Overall connectivity best models dominate in Table **2** with seven best performances and three second ones, followed by the overall Wiener best models showing two best performances, six second, and two third ones. The overall Zagreb indices perform less successful, with one best model (OM2, critical volume) and one second best model (OM1, heat of formation). The improvement against the best molecular connectivity models is considerable, *e.g.*, the standard deviation *sd* of the heat of atomization was reduced 20-fold (0.30 *vs.* 5.80), the boiling point *sd* was cut to a half (1.60K *vs.* 3.31K), *etc.*

The analysis of the specific parameters incorporated into *molecular connectivity models* reveals that the most significant term is $^{o}\chi$ (the square root of the number of vertices), which is included in seven of the ten models. In contrast, only three of the properties (boiling point, molecular refraction, and critical volume) are found to be strongly size dependent in models with overall topological indices, as evidenced by the ^{0}SC term, directly expressing the number of atoms (graph vertices). Three other properties (heat of atomization, molecular volume and critical pressure) are most significantly dependent on the *number of edges* (atom-atom connectivity), which are represented by the first-order overall connectivity index ^{1}OC. The heat of formation depends strongly on another, more intricate measure of atom-atom connectivity, expressed by the number of two-edge subgraphs ^{2}SC. The remaining three properties examined (surface tension, heat of vaporization and critical temperature) were found to depend either on the entirety of molecular subgraphs (*SC*) or on the entirety of their total adjacencies as characterized by the OC index. The second most significant terms in the overall indices models include one size term, and three terms related to the number of edges, whereas in the remaining six cases more complex topological indices are involved.

Table 2: Standard deviations of the best five-parameter models of ten C3-C8 alkanes physicochemical properties with different overall topological descriptors, compared to those obtained by molecular connectivity indices

Property	***Standard Deviation***								
Boiling Point	***OC****	<	*OW*	<	*OM2*	<	*OM1*	<	χ
	1.60		1.70		2.71		2.76		3.31
Heat of Formation	*OC**	**<**	***OM1***	**<**	***OW***	**<**	χ	**<**	***OM2***
	1.02		1.08		1.33		1.37		1.55
Heat of Vaporization	***OC****	<	*OW*	=	*OM2*	<	*OM1*	<	χ
	0.67		0.70		0.70		0.79		0.79
Heat of Atomization	***OC****	<	*OW*	<	*OM2*	<	*OM1*	<	χ
	0.30		0.34		0.39		2.96		5.80
Surface Tension	***OC****	<	*OW*	<	χ	<	*OM1*	<	*OM2*
	0.17		0.18		0.22		0.23		0.27
Molecular Refraction	***OC****	<	*OW*	<	χ	<	*OM2*	<	*OM1*
	0.041		0.042		0.044		0.050		0.057
Molar Volume	**OW**	<	*OC**	<	χ	<	*OM1*	<	*OM2*
	0.23		0.33		0.36		0.45		0.57
Critical Volume	***OM2***	<	*OC**	<	*OW*	<	*OM1*	<	χ
	0.0076		0.0079		0.0080		0.0081		0.0087
Critical Pressure	***OC****	<	*OW*	=	*OM1*	<	*OM2*	<	χ
	0.37		0.40		0.40		0.43		0.50
Critical Temperature	**OW**	<	*OC**	<	χ	<	*OM1*	<	*OM2*
	3.23		3.25		4.76		4.82		5.13

*OC** (abbreviation for overall connectivity) is used for the set of all *OC* and *SC* terms.

A comparison between the best model statistics of the octane properties models produced by the series of overall connectivity, overall Wiener, overall Zagreb indices and molecular connectivity is made in Table **3**. The smaller set of 18 compounds restricted to four the maximum number of variables to be included in the overall Wiener models, but did not prevent the derivation of models with five parameters for the other four types of topological descriptors. The overall Wiener *four*-parameter model (data not shown) compared favorably with the other four-parameter models of nine of the ten alkane properties examined [71]. However, the *five*-parameter models based on overall connectivity indices restored their dominance, and showed the best statistics for eight of the simulated properties. The classical molecular connectivity indices of Kier and Hall could not provide any best or second-best performing four- or five-variable model.

Table 3: Standard deviations of the best 5-parameter models of ten C3-C8 alkane physico-chemical properties with different overall topological descriptors compared to those obtained by molecular connectivity indices

Property	***Standard Deviation***								
Boiling Point	***OC****	<	*OW*	<	χ	<	*OM1*	<	*OM2*
	0.55		0.61		0.73		0.78		1.40
Heat of Formation	*OC**	**<**	***OW***	**<**	***OM1***	**<**	***OM2***	**<**	$\boldsymbol{\chi}$
	0.70		1.10		1.23		1.06		1.25
Heat of Vaporization	***OC****	<	*OW*	=	*OM1*	<	χ	<	*OM2*
	0.28		0.32		0.32		0.34		0.36
Heat of Atomization	***OC****	<	*OW*	<	*OM1*	<	*OM2*	<	χ
	0.17		0.26		0.26		0.30		0.31
Surface Tension	***OC****	=	***OM1***	<	*OW*	<	χ	<	*OM2*
	0.09		0.09		0.10		0.13		0.22
Molecular Refraction	***OC****	<	*OW*	=	*OM1*	<	χ	<	*OM2*
	0.012		0.013		0.013		0.020		0.027
Molar Volume	**OC***	<	*OW*	<	χ	<	*OM1*	<	*OM2*
	0.27		0.34		0.35		0.36		0.45
Critical Volume	***OM1***	=	***OM2***	<	χ	<	*OW*	<	*OC**
	0.0080		0.0080		0.0083		0.0084		0.0088
Critical Pressure	***OW***	<	*OM1*	<	*OM2*	=	χ	<	*OC**
	0.15		0.17		0.20		0.20		0.23
Critical Temperature	***OC****	<	OW	<	χ	<	*OM1*	<	*OM2*
	0.86		0.87		1.12		1.46		2.65

*OC** (abbreviation for overall connectivity) is used for the set of *OC and SC* terms. The data for overall Wiener indices refer to four-variables models.

There is an essential difference in the parameters dominating the C3-C8 models and those of the octane models. The higher-order indices, which capture more details in the variations in molecular topology, have considerably higher weights in octane models than in the models that vary both size and topology. This indicates that models obtained for a single series of isomers cannot be a good basis for prediction of properties of chemical compounds.

Some property-specific indices patterns were identified. Thus, the best C3-C8 *OM1* models for surface tension and critical temperature incorporated the same five variables: $^{o}OM1$, $^{1}OM1$, $^{2}OM1$, $^{3}OM1_{p}$ and $^{5}OM1_{p}$. Similar coupling of these two properties was found in the octane models, which included several more structure-specific indices: $^{1}OM1$, $^{2}OM1$, $^{3}OM1_{c}$, $^{6}OM1_{p}$, and $^{6}OM1_{pc}$. Other pairs of coupled properties were identified though with four, instead of five identical indices: boiling point and heat of vaporization, as well as molar volume and critical volume. Heat of atomization and heat of formation of isomeric octanes, which are linearly dependent in isomers, are also described by models involving a set of the same five indices: $^{1}OM1$, $^{3}OM1_{c}$, $^{4}OM1_{c}$, $^{5}OM1_{p}$, and $^{5}OM1_{c}$. This analysis confirms the capacity of overall topological indices to capture physicochemical properties that have common molecular mechanism.

ALTERNATIVE APPROACHES TO A MORE COMPLETE TOPOLOGICAL REPRESENTATION OF MOLECULAR STRUCTURE

Extended Connectivity and Its Overall Version

The overall connectivity concept is not the only approach in the search for a more general representation of molecular topology. In this approach, different graph invariants, such as graph adjacency, distance, number of matchings, *etc.* are generalized by summing up their values in all subgraphs, including the graph itself considered as proper subgraph.

An alternative approach could be developed by using the Morgan's extended connectivity algorithm [23,24], discussed in Section 1. Instead of using subgraphs of increasing size this algorithm recalculates vertex degrees in successive steps by using the degrees of each vertex in the layers of neighboring vertices at a distance k = 1, 2,., d(max). The sum of *vertex extended connectivities* (perhaps a more precise term would be *extended vertex degree*) for each distance *k* defines the graph k^{th} order extended connectivity, ^{k}EC (eq. 2b), the initial vertex degrees being thus considered as null-order degrees. In order to compare the complexity of graphs it seems reasonable to select the number of iteration steps so as to the same as in the path graph of the same size. thus, for all graphs of same number of vertices V, the number of Morgan's iteration is taken equal to V-1. As shown in Fig. **(7)**, where all acyclic graphs having 3 to 7 vertices are shown with their *extended connectivity vector*, the $^{k(max)}EC$-values reflect fairly well the increase in graph size and complexity. This alternative general representation of molecular topology leads to ordering of graphs rather close to that produced by the overall complexity indices. It captures similarly the complexity patterns of increasing number and length of branches, branches more central location and preferred attachment to the same vertex. We can treat the vector of ^{k}EC terms in the manner

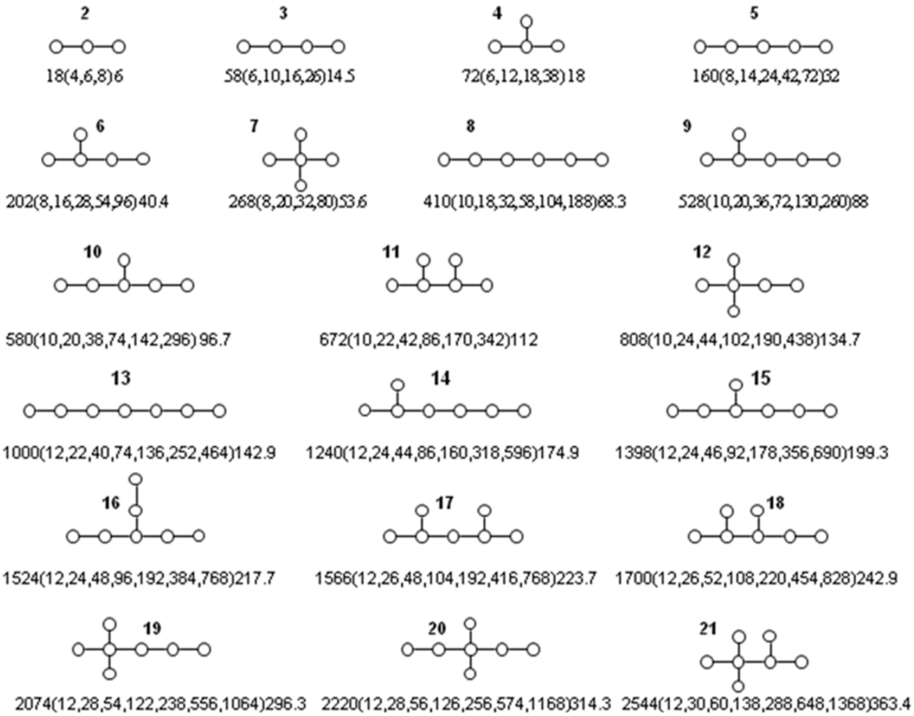


Figure 7: Acyclic graphs with 2-7 vertices, their k^{th}-order series of extended connectivity indices $^{k}EC(G)$ (k = 0, 1, 2,., V-1), and their overall extended connectivity OEC(G). Both the maximal (V-1)th order extended connectivity and the OEC(G), demonstrate high sensitivity to complexifying structural patterns. Shown after the parenthesis is also the overall extended connectivity index of each graph averaged per vertex, which also captures precisely all the patterns of increasing complexity due to graph size and the manner in which the graph vertices are connected.

of overall topological indices concept and define the *overall extended connectivity* of graph *G*, *OEC(G)*, as the sum over all *k* terms:

$$OEC(G) = \sum_{k=0}^{V-1} {}^{k}EC(G) \tag{35}$$

In Fig. **(7)**, the overall version of extended connectivity indices demonstrates a discriminatory potential higher than that of the maximal ${}^{k}EC$ index and produces different values for the single pair of graphs **16** and **17** with degenerate ${}^{k(max)}EC$ - value of 768.

The Total Walk Count

A very detailed description of molecular structure, and hence a very sensitive measure of molecular complexity, can also be reached by counting all walks *w* of different length *l*, an approach developed by Rücker and Rücker [88, 89]. A *walk W* in a graph *G* is an alternating sequence of vertices and edges, $W = v_0, e_1, v_1, e_2, ., e_n, v_n$, such that for $j = 1,., n$, the vertices v_{j-1} and v_j are the endpoints of the edge e_j. Repetitions of vertices and edges in a walk are allowed. The length *l* of a walk is the total number of edges in it, repetitions included. The number of walks of length *l* between vertices *i* and *j* is given by the *ij*-element $(\mathbf{A}^l)_{ij}$ in the l^{th} power of the adjacency matrix **A**. The sum over the row *i* entries in $\mathbf{A}^l$ defines the vertex *i* walk count of length *l*. It is denoted for molecular graphs by Rücker and Rücker as *atomic walk count of length l*, $awc_l(i)$:

$$awc_l(i) = \sum_{j-1}^{V} (A^l)_{ij} \tag{36}$$

Summing up over all lengths *l*, defines the *atomic walk count awcs*(*i*) of atom *i*:

$$awcs(i) = \sum_{l=1}^{V-1} awc_l(i) = \sum_{l=1}^{V-1} \sum_{j=1}^{V} (A^l)_{ij} \tag{37}$$

Alternatively, the sum of walk counts of length *l* of all atoms *i* defines the *molecular walk count of length l*, mwc_l:

$$mwc_l = \sum_{i=1}^{V} awc(i) = \sum_{i=1}^{V}\sum_{j=1}^{V}(A^l)_{ij} \tag{38}$$

The total walk count *twc* is defined from eqs. (37, 38) by summing over all atoms *i* or by summing over all lengths *l*, respectively:

$$twc = \sum_{i=1}^{V} awcs(i) = \sum_{l=1}^{V-1} mcw_l = \sum_{l=1}^{V-1}\sum_{i=1}^{V}\sum_{j=1}^{V}(A^l)_{ij} \tag{39}$$

An example illustrating the total walk count approach to the more complete representation of molecular topology, and the counts of the simplest walks of length 1, 2 and 3, is shown in Fig. **(8)** with the hydrogen depleted graph of isobutane molecule.

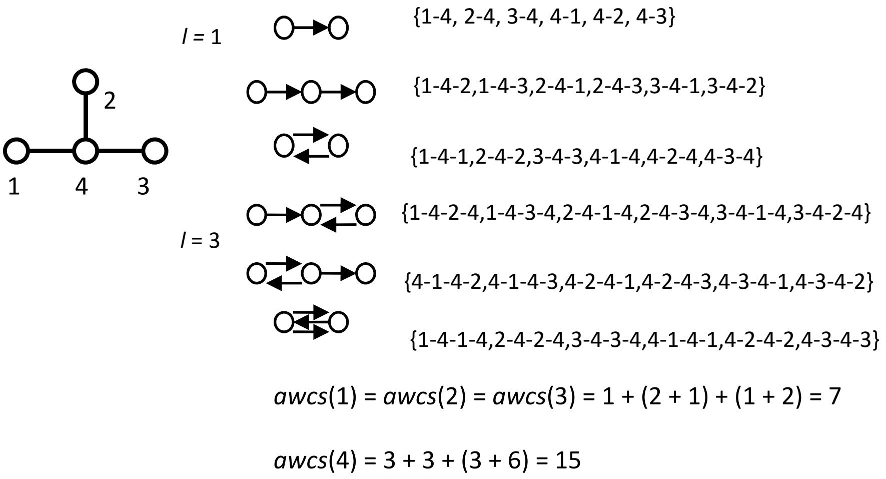


Figure 8: The graph of isobutane molecule, all the walks in it, and the derived atomic, molecular and total walk counts.

The comparison made in Table **4** indicates that while all four complexity measures order in a very similar manner the increasing complexity of acyclic graphs. Subgraph count *SC* and overall connectivity *OC* are slightly more discriminatory showing only a single degenerate pairs of values. The total walk count also follows the basic patterns of increased complexity with the number of branches (M < MM < MMM < MMMM), branch length (M < E; 22MM < 2M3E), centrality (2M < 3M < 4M; 22M < 33MM) and multiplicity at given branch site (23MM < 22MM). As discussed in analyzing the overall topological indices, when two or more complexifying factors act in opposing directions different complexity measures tends to produce conflicting ordering of isomeric compounds. In Table **4**, such disagreements marked in bold indicate conflicts between centrality and multiplicity factors, for example in comparing 2,3,4MMMC5 to 2,2,4MMMC5, for which *twc* favors branch multiplicity to vertex 2 rather than having branch in the more central position 3. Similar competition between the number of branches and their length is seen in the overall Wiener index (OW(3EC6) < OW(2,5MMC6)).

Table 4: Comparison of the total walk count *twc* to three other complexity indices: subgraph count *SC*, overall connectivity *OC*, and overall Wiener *OW*

#	***Compounds***	***twc***	***SC***	***OC***	***OW***
22	nC8	627	36	224	462
23	2MC7	764	41	279	572
24	3MC7	838	44	312	622
25	4MC7	856	45	323	636
26	25MMC6	911	47	348	**709**
27	3EC6	928	48	356	**683**
28	24MMC6	997	51	393	771
29	22MMC6	**1142**	**53**	411	781
30	23MMC6	**1068**	**53**	414	790
31	34MMC6	1136	56	448	837
32	2M3EC5	1152	57	459	848
33	33MMC6	1301	59	477	859
34	224MMMC5	**1317**	62	519	**968**
35	234MMMC5	**1296**	63	**532**	**981**
36	3M3EC5	1441	64	**532**	**921**
37	223MMMC5	1536	69	597	1049
38	233MMMC5	1609	71	618	1065
39	2233MMMMC4	2047	86	798	1312

Overall Bourgas Indices

Several years ago an intriguingly simple way was found to characterize both graph centrality and complexity by local and overall descriptors of the same graph invariant. The idea behind the approach came from the theory of complex dynamic networks. These networks are characterized among other common features by high connectivity and small diameter. The latter property was termed network "small-worldness" [79]. It was conjectured that graph (network) complexity increases with the ratio of the graph total adjacency A and total distance D. This simple graph descriptor was named *first Bourgas complexity index*, and denoted by *B*1. (The name *small-world connectivity* was also alternatively used.) The B1 descriptor was shown to present a good approximate and easily calculable measure of graph complexity [78, 90, 91]. The only pitfall of this descriptor is some degeneracy of calculated values of graphs of equal size and similar structure. A more discriminating complexity measure (termed *second Bourgas complexity index*, *B*2) was simultaneously proposed based on the individual ratios *b*(*i*) of degree *a*(*i*) and distance *d*(*i*) of all graph vertices *i*:

$$B1(G) = \frac{A}{D}; \; B2(G) = \sum_{i=1}^{V} b(i) = \sum_{i=1}^{V} \frac{a(i)}{d(i)} \tag{40a,b}$$

A third version *B3*(*G*) of these indices was defined [90, 92] as the information content of the *B2*(*G*) distribution in the set of all b_i values:

$$B3(G) = B2\log_2 B2 - \sum_{i=1}^{V} b(i)\log_2 b_i \tag{41}$$

The properties of these complexity indices were analyzed and analytical formulas were derived for several classes of graphs. It was also shown that the *b*(*i*) ratios emerged as new, adequate measure of network nodes centrality [78, 91], so that both complexity and centrality of a graph or network are described within the same topological framework.

Here we present a very recent development of these ideas by constructing *overall Bourgas complexity index*, *OB*(*G*) [93].

Definition: The overall Bourgas complexity index is the sum of vertex degree/vertex distance ratios, b(i) = a(i)/d(i), for all vertices i in all k subgraphs to which the vertex belongs:

$$ob(i) = \sum_{k=1}^{k(\max)} \frac{a_i}{d_{i,k}}\,;\; OB(G) = \sum_{i=1}^{V} ob(i) = \sum_{i=1}^{V} \sum_{k=1}^{k(\max)} \frac{a_i}{d_{i,k}} \qquad \text{(42a,b)}$$

Fig. **(9)** illustrates the concept with the series of acyclic graphs having six vertices (isomeric C6 alkanes). The increase in complexity of these structures from the linear one to monobranched, to monobranched with a more central location of the branch, to symmetrical dibranched, and to dibranched with branches to the same vertex is matched by both B2 and OB descriptors. Similarly, the centrality of vertices in these graphs is precisely matched by the local *b*(*i*) and *ob*(*i*) indices to increase from terminal vertices to the most centrally located and having higher degree vertices in agreement with intuition. Yet, *ob*(*i*) is more discriminative, showing different values for the central and terminal vertices in the third and fourth structure compared to the 3/7 and 1/11 values for both produced by the *b*(*i*) descriptor.

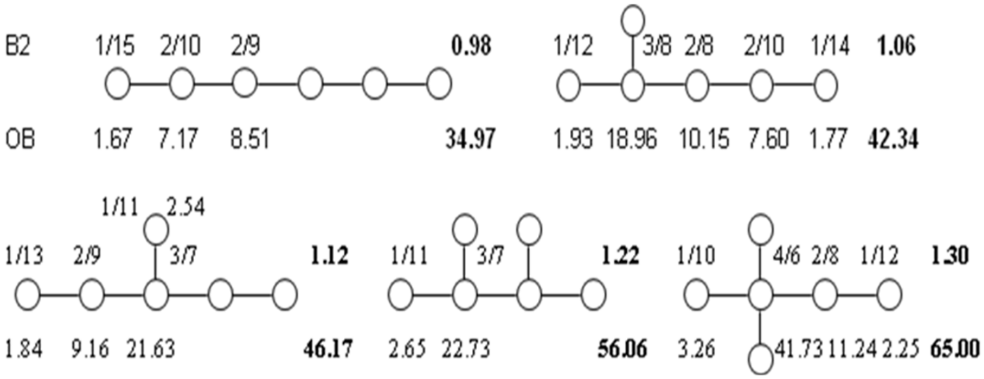


Figure 9: Molecular graphs of isomeric C6 alkanes illustrate the ability of the Bourgas complexity index *B2* and its overall version *OB* (shown in bold) to increase with the increase in the number of branches, their more central location and their location at a higher degree vertex. Simultaneously, the respective vertex indices b(i) and ob(i) increase in value when shifting from a terminal to a central and higher vertex degree location. (The values of these indices for symmetrically located vertices are shown only once for avoiding overcrowding).

The centrality measure *OB*(*i*) defined above by eq. (42a) may be regarded to some extent as weighted version of network *subgraph centrality* concept introduced by Estrada [94, 95], in which the larger the number of subgraphs a given vertex is involved in, the more central its location in the graph.

ACKNOWLEDGEMENTS

Declared none.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

REFERENCES

[1] Harary, F. *Graph theory*, 2nd ed; Addison-Wesley: Reading, MA, **1969**.

[2] Gross, J.L.; Yellen, J., Eds. *Handbook of Graph Theory*; CRC Press: Boca Raton, FL, **2004**, pp.1167.

[3] Balaban, A.T.; Motoc, I.; Bonchev, D.; Mekenyan, O. *Topological indices for QSAR*. In: *Topics Curr. Chem.*; Springer: Berlin, **1983**, Vol. 114, p.21-56.

[4] Devillers, J.; Balaban, A.T., Eds. *Topological Indices and Related Descriptors in QSAR and* QSPR; Gordon and Breach: Amsterdam, **1999**, pp. 811.

[5] Diudea, M.V., Ed. *QSPR / QSAR Studies by Molecular Descriptors*; Nova: Huntington, N.Y., **2001**, pp. 438.

[6] Bonchev, D.; Rouvray, D.H., Eds. *Mathematical Chemistry*, Vol. 1, *Chemical Graph Theory. Introduction and Fundamentals*; Gordon and Breach: Reading, U.K., **1991**, pp. 288.

[7] Trinajstić, N. *Chemical Graph Theory*, 2nd ed.; CRC Press: Boca Raton, FL, **1992**, pp. 352.

[8] King, R.B.; Rouvray, D.H., Eds. *Graph Theory and Topology in Chemistry*; Elsevier: Amsterdam, **1987**, pp. 476.

[9] Diudea, M.V.; Gutman, I.; Jäntschi, L. *Molecular Topology*; Nova: Huntington, N.Y., **2001**, pp. 332.

[10] Rouvray, D.H.; King, R.B., Eds. *Topology in Chemistry. Discrete Mathematics of Molecules*; Horwood: Chichester, England, **2002**, pp. 387.

[11] Janežič, D.; Miličević, A.; Nikolić, S.; Trinajstić, N. *Graph Theoretical Matrices in Chemistry*; MCM: Kraguevac, **2007**, pp. 205.

[12] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, **2009**, Vol. I, pp. 967; Vol. II, pp. 257.

[13] Kier, L.B.; Hall, L. *Molecular Structure Description:* The Electrotopological State; Academic Press: San Diego, CA, **1999**, pp. 245.

[14] Balaban, A.T., Ed. *From Chemical Topology To Three-Dimensional Geometry*; Plenum Press: **1997**, pp. 420.

[15] Kubinyi, H. QSAR and 3D-QSAR in Drug Design. *DDT* **1997**, *2*(11), 457-467.

[16] Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, **2000**, pp. 448.

[17] Bleicher, K.H.; Bohm, H.J.; Muller, K.; Alanine, A.I., Hit and Lead Generation: Beyond High- Throughput Screening. *Nat. Rev. Drug. Discov*. **2003**, *2*(5), 369-378.

[18] Brehme, M.; Hantschel, O.; Colinge, J.; Kaupe, I.; Planyavsky, M.; Köcher, T.; Mechtler, K.; Bennett, K.L.; Superti-Furga, G. Charting the Molecular Network of the Drug Target Bcr-Abl. *PNAS*, **2009**, *106* (18), 7414-7419.

[19] *MollConnZ*, Version 4.05, **2003**; Hall Ass. Consult.; Quincy, MA.

[20] DRAGON - Software for the Calculation of Molecular Descriptors, Version 5.4, **2006;** Todeschini, R.; Consonni, V.; Mauri, A. *et al*., Talete srl.; Milan, Italy.

[21] Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc*. **1947**, *69*, 17-20.

[22] Wiener, H. Relation of the Physical Properties of the Isomeric Alkanes to Molecular Structure. *J. . Phys. Chem*. **1948**, *52*, 1082-1089.

[23] Morgan, H.L. The Generation of a Unique Machine Description for Chemical Structures - A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc*. **1965**, *5*, 107-113.

[24] Razinger, M. Extended Connectivity in Chemical Graphs. *Theor. Chim. Acta* **1982**, *61*, 581-586.

[25] Gutman, I.; Ruščić, B.; Trinajstić, N.; Wilcox, Jr., C.W. Graph Theory and Molecular Orbitals. 12. Acyclic Polyenes. *J. Chem. Phys.* **1975**, *62*, 3399-3405.

[26] Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc*. **1975**, *97*, 6609-6615.

[27] Kier, L.B.; Hall, L.H.; Murray, W.J.; Randić, M. Molecular Connectivity. I. Relationship to Nonspecific Local Anesthesia. *J. Pharm. Sci.* **1975**, *64*, 1971-1974.

[28] Kier, L.B.; Hall, L.H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, **1976**, pp. 257.

[29] Kier, L.B.; Hall, L.H. Derivation and Significance of Valence Molecular Connectivity. *J. Pharm. Sci.* **1981**, *70*, 583-589.

[30] Smolenskii, G.A. Graph Theory Application to the Calculation of Structure-Additive Properties of Hydrocarbons. *Zh. Fiz. Khim.* **1964**, *38*, 1288-1291.

[31] Gordon, M.; Kennedy, J.W. The Graph-like State of Matter. Part 2. TCGI Schemes for the Thermodynamics of Alkanes and the Theory of Inductive Inference. *J.C.S. Faraday Trans. II* **1973**, *69*, 484-504.

[32] Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure Activity Analysis*; Wiley: London, **1986**, p.262.

[33] Bonchev, D. Overall Connectivity - A Next Generation Molecular Connectivity, *J. Mol. Graphics Model.* **2001**, *20*(10), 65-75.

[34] Shannon, C.; Weaver, W. *Mathematical Theory of Communications*; University of Illinois Press: Urbana, MI, **1949**.

[35] Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*; Research Studies Press: Chichester, U.K.; **1983**.

[36] Bonchev, D.; Mekenyan, O.; Trinajstić, N. Isomer Discrimination by Topological Information Approach. *J. Comput. Chem*. **1981**, *2*, 127-148.

[37] Bonchev, D.; Trinajstić, N. Chemical Information Theory. Structural Aspects. *Intern. J. Quantum Chem. Symp*. **1982**, *16*, 463-480.

[38] Kastler, H.; Ed. *Essays on the Use of Information Theory in Biology*; University of Illinois Press: Urbana, IL, **1953**.

[39] Rashevsky, N. Life, Information Theory, and Topology. *Bull. Math. Biophys.* **1955**, *17*, 229-235.

[40] Trucco, E. A Note on the Information Content of Graphs. *Bull. Math. Biophys.* **1956**, *18*, 129-135; On the Information Content of Graphs: Compound Symbols; Different States for Each Point. *Bull. Math. Biophys.* **1956**, *18*, 237-253.

[41] Mowshovitz, A. Entropy and the Complexity of Graphs. 1. An Index of the Relative Complexity of a Graph. *Bull. Math. Biophys.* **1968**, *30*, 175-204.

[42] Minoli, D. Combinatorial Graph Complexity. *Atti. Acad. Naz. Lincei Rend.* **1976**, *59*, 651-661.

[43] Bonchev, D. The Problems of Computing Molecular Complexity, In: *Computational Chemical Graph Theory*; Rouvray, D.H. Ed.; Nova: New York, **1990**, pp. 34-67.

[44] Bonchev, D. Shannon's Information and Complexity. In: *Mathematical Chemistry Series*, Vol. 7, *Complexity in Chemistry*; Bonchev, D.; Rouvray, D.H., Eds.; Taylor & Francis: London, **2003**, pp. 155- 187.

[45] Bonchev, D.; Temkin, O.N.; Kamenski, D. On the Classification and Coding of Linear Reaction Mechanisms. *React. Kinet. Catal. Lett.* **1980**, *15*, 113-118.

[46] Bonchev, D.; Kamensky, D.; Temkin, O.N. Complexity Index for the Linear Mechanisms of Chemical Reactions. *J. Math. Chem.* **1987**, *1*, 345-388.

[47] Gutman, I.; Mallion, R.B.; Essam, J.W. *Counting Spanning Trees of a Labeled Molecular Graph, Mol. Phys.* **1983**, *50*, 859-877.

[48] John, P.E.; Mallion, R.B.; Gutman, I. An Algorithm for Counting Spanning Trees in Labeled Molecular Graphs Homeomorphic to Cata-Conjugated Systems. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 108-112.

[49] Bertz, S.H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599- 3601.

[50] Bertz, S.H. A Mathematical Model of Molecular Complexity. In: *Chemical Applications of Topology and Graph Theory*; King, R.B., Ed.; Elsevier: Amsterdam, **1983**, pp. 206-221.

[51] Bonchev, D.; Polansky, O.E. On the Topological Complexity of Chemical Systems. In: *Graph Theory and Topology in Chemistry*; King, R.B.; Rouvray, D.H. Eds.; Elsevier: Amsterdam, **1987**, pp. 126-158.

[52] Basak, S.C. Use of Molecular Complexity Indices in Predictive Pharmacology and Toxicology: A QSAR Approach. *Med. Sci.Res.* **1987**, *15*, 605-609.

[53] Randić M.; Plavšić, D. On Characterization of Molecular Complexity. *Int. J. Quantum Chem.* **2003**, *91*(1), 20-31.

[54] Randić, M.; Guo, X.; Plavšić, D.; Balaban, A.T. On the Complexity of Fullerenes and Nanotubes. In: *Mathematical Chemistry Series. Vol. 7, Complexity in Chemistry*; Bonchev, D.; Rouvray, D., Eds.; Gordon and Breach: Amsterdam, **2005**.

[55] Nikolić, S.; Trinajstić, N.; Tolić, I. M.; Rücker, G.; Rücker, C. On Molecular Complexity Indices. In: *Mathematical Chemistry Series. Vol. 7, Complexity in Chemistry*; Bonchev, D.; Rouvray, D., Eds.; Gordon and Breach: Amsterdam, **2005**.

[56] Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.* **1977**, *67*, 4517-4533.

[57] Bonchev, D.; Knop, J.V.; Trinajstić, N. Mathematical Models of Branching. *MATCH Commun. Math. Comput. Chem.* **1979**, *6*, 21-47.

[58] Ruch, E.; Gutman, I. The Branching Extent of Graphs. *J. Comb. Inf. System Sci.* **1979**, *4*, 285-295.

[59] Bonchev, D. Topological Order in Molecules. 1. Molecular Branching Revisited. *Theochem* **1995**, *336*, 137-156.

[60] Bonchev, D.; Mekenyan, O.,; Trinajstić, N. Topological Characterization of Cyclic Structures. *Intern. J. Quantum Chem.* **1980**, *17*, 845-893.

[61] Mekenyan, O.; Bonchev, D.; Trinajstić, N. On Algebraic Characterization of Bridged Polycyclic Compounds. *Intern. J. Quantum Chem.* **1981**, *19*, 929-955.

[62] Balaban, A.T.; Bonchev, D.; Liu, X.; Klein, D.J. Molecular Cyclicity and Centricity of Polycyclic Graphs. I. Cyclicity Based on Resistance Distances or Reciprocal Distances. *Int. J. Quantum Chem.* **1994**, *50*, 1- 20.

[63] Pisanski, T.; Plavšić, D.; Randić, M. On Numerical Characterization of Cyclicity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*(3), 520-523.

[64] Bonchev, D. Kolmogorov's Information, Shannon's Entropy, and Topological Complexity of Molecules. *Bulg. Chem. Commun.* **1995**, *28*, 567-582.

[65] Bonchev, D. Novel Indices for the Topological Complexity of Molecules. *SAR QSAR Environ. Res.* **1997**, *7*, 23-43.

[66] Bonchev, D. Overall Connectivity and Molecular Complexity. In: *Topological Indices And Related Descriptors*; Devillers, J.; Balaban, A.T., Eds.; Gordon and Breach: Reading, U.K., **1999**, p. 361- 401.

[67] Bertz, S.H.; Sommer, T.J. Rigorous Mathematical Approaches To Strategic Bonds and Synthetic Analysis Based On Conceptually Simple New Complexity Indices. *Chem. Commun.***1997**, 2409-2410.

[68] Bertz, S.H.; Wright, W.F. The Graph Theory Approach To Synthetic Analysis: Definition and Application of Molecular Complexity and Synthetic Complexity. *Graph Theory Notes New York Acad. Sci.* **199**8, *35*, 32-48.

[69] Bertz, S.; Herndon, W. C. Similarity of Graphs and Molecules. In: A*rtificial Intelligence Applications in Chemistry*; ACS: Washington, D.C., **1986**, pp.169-175.

[70] Bonchev, D. Overall Connectivities / Topological Complexities: A New Powerful Tool for QSPR/QSAR. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 934-941.

[71] Bonchev, D. The Overall Wiener Index - A New Tool for Characterization of Molecular Topology, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 582-592.

[72] Bonchev, D.; Trinajstić, N. Overall Molecular Descriptors. 3. Overall Zagreb Indices. *SAR/QSAR Envir. Res.;* **2001**, *12*, 213-235.

[73] Bonchev, D. The Overall Topological Complexity Indices. *Lect. Ser. Comp. Comput. Sci.* **2005**, *4*, 1554-1557.

[74] Hosoya**,** H. Topological Index. A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn***. 1971,** *44*, 2332-2339.

[75] Randić M. The Connectivity Index 25 Years After. *J. Mol. Graphics Modelling.* **2001**, *20*, 19-35.

[76] Estrada, E.; Guevara, N.; Gutman, I. Extension of Edge Connectivity Index. Relationship to Line Graph Indices and QSPR Applications. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 428-431.

[77] Kier, L.B.; Hall, L.H. The Nature of Structure-Activity Relationships and Their Relation to Molecular Connectivity. *Eur. J. Med. Chem*. **1977,** *12*, 307-312.

[78] Bonchev, D.; Buck, G.A. Quantitative Measures of Network Complexity. In: *Chemistry, Biology and Ecology*; Bonchev, D., Rouvray, D.H., Eds.; Springer: New York, **2005**, pp. 191-235.

[79] Watts, D.J.; Strogatz, S.H. Collective Dynamics of "Small-World" Networks. *Nature* **1998**, *393*, 440- 442.

[80] Barabási, A.-L.; Albert, R. Emergence of Scaling in Random Networks. *Science* **1999**, *286*, 509-512.

[81] Dorogovtsev, S.N.; Mendes, J.F.F. Evolution of Networks. *Adv. Phys.* **2002**, *51*, 1079-1187.

[82] Kitano, H. Computational Systems Biology. *Nature* **2002**, *420*(11), 206-210.

[83] Alon, U. Network Motifs: Theory and Experimental Approaches. *Nature Rev. Genet.* **2007**, *8*, 450- 461.

[84] Temkin, O.N.; Zeigarnik, A.V.; Bonchev, D. *Chemical Reaction Networks. A Graph Theoretical Approach*; CRC Press: Boca Raton, FL, **1996**, pp. 286.

[85] Needham, D.E.; Wei, I.-C.; Seybold, P.G. Molecular Modeling of the Physical Properties of the Alkanes. *J. Am. Chem. Soc.* **1988**, *110*, 4186-4194.

[86] American Petroleum Institute, Research Project 44. Selected Values of Properties of Hydrocarbons and Related Compounds; Pittsburgh, PA, **1977**.

[87] Garbalena, M.; Herndon, W. C. Optimum Graph-Theoretical Models for Enthalpic Properties of Alkanes. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 37-42.

[88] Rücker, G.; Rücker, C. Walk Counts, Labyrinthicity, and Complexity of Acyclic and Cyclic Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 99-106.

[89] Rücker, G.; Rücker, C. Substructure, Subgraph and Walk Counts as Measures of the Complexity of Graphs and Molecules. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1457-1462.

[90] Bonchev, D.; Buck, G.A. From Molecular to Biological Structure and Back. *J. Chem. Inform. Model.* **2007**, *47*, 909-917.

[91] Bonchev, D. A Simple Integrated Approach to Network Complexity and Node Centrality. In: *Analysis of Complex Networks. From Biology to Linquistics*; Dehmer, M.; Emmert-Streib, F., Eds.; Wiley- Blackwell: Weinheim, **2009**, p. 47-53.

[92] Bonchev, D. Information Theoretic Measures of Complexity. In: *Encyclopedia of Complexity and System Science*; R. Meyers, Ed.; Springer: Heidelberg, Germany, **2009**, *5*, pp. 4820-4838.

[93] Bonchev, D. On the Integrated Representation of Network Complexity and Node Centrality. *Current Computer-Aided Drug Design* (*CCADD*), submitted.

[94] Estrada, E.; Rodriguez, J.A. Subgraph Centrality in Complex Networks. *Phys. Rev. E* 2005, *71*, 056103. 9 pp.

[95] Estrada, E. Generalized Walks-Based Centrality Measures for Complex Biological Networks. *J. Theor. Biol.* **2011**, *285*(1), 147-155.

CHAPTER 4

The Four Connectivity Matrices, Their Indices, Polynomials and Spectra

Bono Lučić, Ivan Sović and Nenad Trinajstić*

The Ruđer Bošković Institute, P.O.Box 180, HR-10 002 Zagreb, Croatia

Abstract: The four connectivity matrices are presented: the *vertex*-product-connectivity matrix, the *edge*-product-connectivity matrix, the *vertex*-sum-connectivity matrix and the *edge*-sum-connectivity matrix. The half-sum of their matrix elements are the corresponding connectivity indices: the *vertex*-product-connectivity index, the *edge*-product-connectivity index, ve*rtex*-sum-connectivity index and the *edge*-sum-connectivity index. The suitability of these four forms of connectivity indices in developing structure-property relationships is illustrated on four sets of alkanes for 14 experimental physico-chemical properties. Their polynomials and spectra are also given. The method used for constructing polynomials of the connectivity matrices considered is the Le Verrier-Fadeev-Frame method, that has been modified by Balasubramanian and Živković.

Keywords: Connectivity matrix, connectivity index, vertex-product, edge-product, vertex-sum, edge-sum, connectivity matrix spectra, connectivity matrix polynomial, matrix polynomial computation, Le Verrier-Fadeev-Frame method, structure-property relationship, inter-correlation, comparative correlational analysis, alkane data sets, branched alkanes, non-branched alkanes, physico-chemical properties, boiling points, melting points, vaporization enthalpy, standard Gibbs energy of formation, refractive index, density, molar heat capacity.

INTRODUCTION

Milan Randić introduced the product-connectivity matrix in his paper entitled *Similarity based on extended basis descriptors*, published in 1992 [1] with the following text: "We start with the χ-matrix, which is based on the adjacency matrix, but instead of the binary entries 0 and 1 we assign the value of $1/\surd(m,n)$ to

***Corresponding author Nenad Trinajstić:** Ruđer Bošković Institute, Bijenička c. 54, HR-10000 Zagreb, Croatia; Tel: ++385-1-4680095; Fax: ++385-1-4680245; E-mail: trina@irb.hr

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50004-7

nonzero matrix elements, where *m* and *n* are the valencies of the vertices involved." This result remained unknown. Thirteen years later the product-connectivity matrix has been rediscovered [2-5], but Randić's paper was not mentioned. In these reports the product-connectivity matrix has been discussed under different names, *e.g.*, the weighted adjacency matrix [2], the degree-adjacency matrix [3], the normalized adjacency matrix [4] and the Randić matrix [5]. We use the name the *product*-connectivity matrix to differ it from the related matrix that we call the *sum*-connectivity matrix [6]. Since we consider two forms of the product-connectivity matrix, one based on vertices and the other on edges, the first matrix is called the *vertex*-product-connectivity matrix and the other *edge*-product-connectivity matrix, respectively.

The non-vanishing elements of the vertex-product-connectivity matrix are the vertex-product-connectivity indices. In 1975, Randić [7] introduced the vertex-product-connectivity index and it has been since the most used molecular descriptor in QSPR and QSAR modeling. Todeschini and Consonni in their two *Handbooks* [8, 9] reviewed the uses of the vertex-product-connectivity index and related molecular descriptors in modeling properties and activities of various classes of molecules. Additionally, Todeschini and Consonni also discussed in their *Handbooks* the role of graph-theoretical matrices in deriving molecular descriptors (topological indices) [10, 11].

In 2010, we introduced the sum-connectivity matrix in parallel to the product-connectivity matrix [12]. Its non-vanishing elements are the sum-connectivity indices [6] whose uses in the QSPR and QSAR modeling parallels that of the product-connectivity indices [13, 14]. Randić and his co-workers [15] introduced independently the sum-connectivity matrix, but named this matrix the *distance-weighted adjacency matrix*. In the same report Randić *et al.* [15] also introduced the sum-connectivity index which they called the *modified connectivity index*. In their paper, they did not refer to our work. In this report we call this descriptor as the *vertex*-sum-connectivity index and the corresponding matrix vertex-sum-connectivity matrix. Similarly, the sum-connectivity matrix based on edges is called here the *edge*-sum-connectivity matrix. Randić *et al.* in their paper also listed 13 matrices of interest in chemistry [15]. Most of these matrices are also included in our monograph Graph-Theoretical Matrices in Chemistry [16].

In this report, we present four types of connectivity matrices, that is, the *vertex*-product-connectivity matrix, the *edge*-product-connectivity matrix, the *vertex*-sum-connectivity matrix and the *edge*-sum-connectivity matrix, respectively. We also give their polynomials and spectra. We use in this report the (chemical) graph-theoretical concepts and terminology [17-19].

PRODUCT-CONNECTIVITY MATRICES

The *vertex*-Product-Connectivity Matrix

The *vertex*-product-connectivity matrix of a molecular graph G, denoted by $\mathbf{V} = \mathbf{V}(G)$, is a symmetric square matrix defined as:

$$\left[\mathbf{V}(G)\right]_{ij} = \begin{cases} [d(i)d(j)]^{-1/2} & \text{if vertices } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

where $d(i)$ and $d(j)$ are the degrees of vertices i and j. The vertex-product-connectivity index ${}^{v}\chi$ based on this matrix is given by:

$$ {}^{v}\chi = (1/2) \sum_{i,j} [\mathbf{V}(G)]_{ij} \qquad (2)$$

As an example, we give below the vertex-product-connectivity matrix $\mathbf{V}(G_1)$ of G_1 depicting carbon skeleton of 1,1,2-trimethylcyclopropane.

$$\mathbf{V}(G_1) = \begin{bmatrix} 0 & 0.2887 & 0.3536 & 0.5 & 0.5 & 0 \\ 0.2887 & 0 & 0.4082 & 0 & 0 & 0.5774 \\ 0.3536 & 0.4082 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5774 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The vertex-labeled graph G_1 is shown in Fig. (**1**) and the vertex degrees in G_1 in Fig. (**2**), respectively.

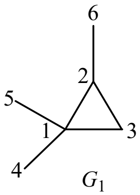


Figure 1: The vertex-labeled graph G_1 representing the carbon skeleton of 1,1,2-trimethylcyclopropane.

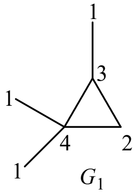


Figure 2: The vertex-degrees in G_1.

The spectrum of this matrix is given by: 1.0000, 0.5132, 0, 0, -0.6921, -0.8212. The corresponding vertex-product-connectivity index ${}^{v}\chi$ is 2.6279.

The *edge*-Product-Connectivity Matrix

The *edge*-product-connectivity matrix of a molecular graph G, denoted by $\mathbf{E} = \mathbf{E}(G)$, is a symmetric square matrix defined as:

$$\left[\mathbf{E}(G)\right]_{ij} = \begin{cases} [d(e_i)d(e_j)]^{-1/2} & \text{if edges } e_i \text{ and } e_j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $d(e_i)$ and $d(e_j)$ are the degrees of edges i and j. The edge-product-connectivity index ${}^{e}\chi$ based on this matrix is given by

$${}^{e}\chi = (1/2) \sum_{i,j} [\mathbf{E}(G)]_{ij} \tag{4}$$

As an example, we give below the edge-product-connectivity matrix $\mathbf{E}(G_1)$ of G_1 depicting carbon skeleton of 1,1,2-trimethylcyclopropane. The edge-degrees of G_1 are shown in Fig. (**3**).

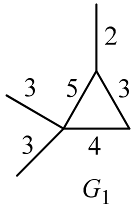


Figure 3: The edge-degrees in G_1.

However, the *edge*-product-connectivity matrix of graph G, $\mathbf{E}(G)$, is the *vertex*-product-connectivity matrix of the corresponding line graph L(G), $\mathbf{V}$[L(G)]:

$$\mathbf{E}(G) = \mathbf{V}[\mathrm{L}(G)] \tag{5}$$

The line graph L(G) of a graph G is the graph generated from G in such a way that the edges in G are replaced by vertices in L(G). Two vertices in L(G) are connected if the corresponding edges in G are adjacent. In Fig. (**4**), we show the edge-labeled graph G_1 and the corresponding vertex-labeled line graph L(G_1), and in Fig. (**5**), the vertex-degrees in L(G_1), respectively

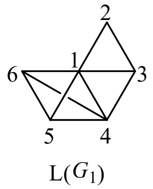


Figure 4: The vertex-labeled line graph L(G_1).

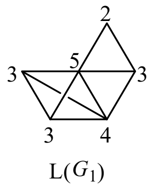


Figure 5: The vertex-degrees in L(G_1).

As an example, we give below the edge-product-connectivity matrix of L(G_1) and the corresponding spectrum.

$$\mathbf{V}[\mathrm{L}(G_1)]=\begin{bmatrix} 0 & 0.3162 & 0.2582 & 0.2236 & 0.2582 & 0.2582 \\ 0.3162 & 0 & 0.4082 & 0 & 0 & 0 \\ 0.2582 & 0.4082 & 0 & 0.2887 & 0 & 0 \\ 0.2236 & 0 & 0.2887 & 0 & 0.2887 & 0.2887 \\ 0.2582 & 0 & 0 & 0.2887 & 0 & 0.3333 \\ 0.2582 & 0 & 0 & 0.2887 & 0.3333 & 0 \end{bmatrix}$$

The spectrum of this matrix is given by: 1, 0.4082, -0.1225, -0.3333, -0.4083, -0.5441. The corresponding edge-product-connectivity index $^{e}\chi$ is 2.9220.

SUM-CONNECTIVITY MATRICES

The *vertex*-Sum-Connectivity Matrix

The *vertex*-sum-connectivity matrix of a molecular graph *G*, denoted by **S** = **S**(*G*), is a symmetric square matrix defined as:

$$\left[\mathbf{S}(G)\right]_{ij}=\begin{cases}[d(i)+d(j)]^{-1/2} & \text{if vertices } i \text{ and } j \text{ are adjacent}\\ 0 & \text{otherwise}\end{cases} \tag{6}$$

where *d*(*i*) and *d*(*j*) are the degrees of vertices *i* and *j*. The vertex-sum-connectivity index $^{v}\varepsilon$ based on this matrix is given by:

$$^{v}\varepsilon=(1/2)\sum_{\mathrm{i,j}}[\mathbf{S}(\mathrm{G})]_{\mathrm{ij}} \tag{7}$$

As an example, we give below the vertex-sum-connectivity matrix **S**(G_1) of G_1 depicting carbon skeleton of 1,1,2-trimethylcyclopropane (see Figs. (**1**) and (**2**)).

$$\mathbf{S}(G_1)=\begin{bmatrix} 0 & 0.3780 & 0.4082 & 0.4472 & 0.4472 & 0 \\ 0.3780 & 0 & 0.4472 & 0 & 0 & 0.5 \\ 0.4082 & 0.4472 & 0 & 0 & 0 & 0 \\ 0.4472 & 0 & 0 & 0 & 0 & 0 \\ 0.4472 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The spectrum of this matrix is given by: 1.0431, 0.4080, 0, 0, -0.6499, -0.8013, and the corresponding vertex-sum-connectivity index $^{v}\varepsilon$ is 2.6278, respectively.

The *edge*-Sum-Connectivity Matrix

The *edge*-sum-connectivity matrix of a molecular graph *G*, denoted by **C** = **C**(*G*), is a symmetric square matrix defined as:

$$\left[\mathbf{C}(G)\right]_{ij} = \begin{cases} [d(e_i)+d(e_j)]^{-1/2} & \text{if edges } e_i \text{ and } e_j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

where $d(e_i)$ and $d(e_j)$ are the degrees of edges *i* and *j*. As an example, we give below the edge-sum-connectivity matrix of L(G_1).

$$\mathbf{C}[L(G_1)] = \begin{bmatrix} 0 & 0.3780 & 0.3536 & 0.3333 & 0.3536 & 0.3536 \\ 0.3780 & 0 & 0.4472 & 0 & 0 & 0 \\ 0.3536 & 0.4472 & 0 & 0.3780 & 0 & 0 \\ 0.3333 & 0 & 0.3780 & 0 & 0.3780 & 0.3780 \\ 0.3536 & 0 & 0 & 0.3780 & 0 & 0.4082 \\ 0.3536 & 0 & 0 & 0.3780 & 0.4082 & 0 \end{bmatrix}$$

The edge-sum-connectivity index $^{e}\varepsilon$ based on this matrix is given by:

$$^{e}\varepsilon = (1/2) \sum_{i,j} [\mathbf{C}(G)]_{ij} \tag{9}$$

The spectrum of this matrix is given by: 1.3075, 0.4656, -0.1642, -0.4082, -0.5373, -0.6634, and the corresponding edge-sum-connectivity index $^{e}\varepsilon$ is 3.7615, respectively.

COMPARISONS BETWEEN THE CONNECTIVITY INDICES

In order to illustrate the dependence of vertex- and edge-connectivity indices on the size of compounds we computed the connectivity matrices and corresponding connectivity indices for the series of 10 acyclyc alkanes from C3 to C12 (Table **1**).

The discriminating potency of the edge-connectivity indices compared with the vertex-connectivity indices was illustrated on the size-independent set of acyclyc isomers of C7 alkanes (Table **1**).

Table 1: The values of product- and sum- vertex-connectivity and the corresponding edge-connectivity indices for 10 non-branched acyclyc alkanes (C3-C12) and 9 C7 isomers, together with the experimental melting and boiling points

Alkane	MP (°C)	BP (°C)	${}^{v}\chi$	${}^{e}\chi$	${}^{v}\varepsilon$	${}^{e}\varepsilon$
non-branched (C3-C12)						
n-propane	-188	-42.05	1.1547	1.0000	1.1547	0.7071
n-butane	-137	-0.15	1.6547	1.4142	1.6547	1.1547
n-pentane	-129.5	36	2.1547	1.9142	2.1547	1.6547
n-hexane	-95	68.5	2.6547	2.4142	2.6547	2.1547
n-heptane	-90.5	98.5	3.1547	2.9142	3.1547	2.6547
n-octane	-57	125.68	3.6547	3.4142	3.6547	3.1547
n-nonane	-53.5	150.5	4.1547	3.9142	4.1547	3.6547
n-decane	-29.5	174	4.6547	4.4142	4.6547	4.1547
n-undecane	-26	195	5.1547	4.9142	5.1547	4.6547
n-dodecane	-9.5	216	5.6547	5.4142	5.6547	5.1547
branched alkanes (C7)						
n-heptane	-90.5	98.5	3.1547	2.9142	3.1547	2.6547
2-methylhexane	-118.5	90	3.0246	2.9319	3.0246	2.9190
3-methylhexane	-119	92	3.0491	2.8425	3.0491	2.8272
2,2-dimethylpentane	-123.7	79.2	2.8272	2.9267	2.8272	3.3442
2,3-dimethylpentane	-135	89.8	2.9328	2.8349	2.9328	3.0499
2,4-dimethylpentane	-123	80.5	2.8944	2.9663	2.8944	3.1971
3,3-dimethylpentane	-135	86.1	2.8656	2.7380	2.8656	3.1681
3-ethylpentane	-119	93.5	3.0737	2.7321	3.0737	2.7247
2,2,3-trimethylbutane	-25	80.9	2.7196	2.9071	2.7196	3.5413

${}^{v}\chi$ is the vertex-product-connectivity index; ${}^{e}\chi$ is the edge-product-connectivity index; ${}^{v}\varepsilon$ is the vertex-sum-connectivity index; ${}^{e}\varepsilon$ is the edge-sum-connectivity index.

Results of correlational analysis between computed indices and experimental properties for these data sets are given in the first part of Table **2**. High inter-correlations (r_{ij} = 1.0) exist between all indices for C3-C12 alkanes, and, for C7 alkanes, the same inter-correlations (r_{ij} ~ 1.0) are between all pairs of indices

Table 2: Inter-correlation coefficients between connectivity indices for different data sets and their correlation coefficients with experimental properties

	${}^{v}\chi$	${}^{e}\chi$	${}^{v}\varepsilon$	${}^{e}\varepsilon$
non-branched alkanes(C3-C12), *n* = 10				
MP	**0.9774**	0.9747	**0.9774**	0.9758
BP	**0.9936**	0.9922	**0.9936**	0.9928
${}^{v}\chi$		0.9999	1.0000	1.0000
${}^{e}\chi$			0.9999	1.0000
${}^{v}\varepsilon$				1.0000
branched alkanes (C7), *n* = 9				
MP	-0.3515	0.3049	-0.3515	**0.3859**
BP	0.9106	-0.3511	0.9106	**-0.9300**
${}^{v}\chi$		-0.1631	1.0000	-0.9904
${}^{e}\chi$			-0.1631	0.2936
${}^{v}\varepsilon$				-0.9904
branched acyclyc alkanes from [14, 20, 22], *n* = 137				
RT	0.9791	0.9532	**0.9792**	0.8584
MP (n= 63)	0.4476	0.5125	0.4536	**0.5518**
BP	0.9785	0.9633	**0.9796**	0.8755
${}^{v}\chi$		0.9574	0.9997	0.8196
${}^{e}\chi$			0.9638	0.9366
${}^{v}\varepsilon$				0.8299
branched acyclyc alkanes between C6 and C10 from [20-22], *n* = 134				
Δ_{vap}H	0.9751	0.9195	**0.9769**	0.7090
Δ_{f}G° (n = 130)	0.6422	0.6807	0.6431	**0.7517**
n_{D}^{25} (n = 133)	0.8048	0.8149	0.8053	**0.8302**
ρ□(n = 133)	0.7903	0.7855	0.7891	**0.8033**
C_{p}	0.9496	**0.9698**	0.9553	0.8736
BP (from [20])	0.9643	0.9400	**0.9661**	0.8200
BP (from [21])	0.9655	0.9344	**0.9666**	0.8113
${}^{v}\chi$		0.9250	0.9995	0.7233
${}^{e}\chi$			0.9363	0.9088
${}^{v}\varepsilon$				0.7397

RI is the acronym for retention index values; MP and BP are the acronyms for melting and boiling point values; *n* denotes the total number of corresponding experimental values used in correlations. $\Delta_{vap}H$, $\Delta_{f}G$°, n_{D}^{25}, ρ, and C_{p} denote: vaporization enthalpy at 300 K (kJ mol^{-1}), the standard Gibbs energy of formation in the gas phase at 300 K (kJ mol^{-1}), refractive index at 25 °C, density at 25 °C (kg m^{-3}), and molar heat capacity at 300 K (J K^{-1} mol^{-1}), respectively. The highest value of correlation coefficient in correlation of indices with experimental properties are given in bold.

except those pairs containing the edge-product connectivity index ($^{e}\chi$), whose inter-correlation with other three indices are below 0.3. Analogously, correlations of all indices for C3-C13 alkanes with melting and boiling points are mutualy (very) similar and high (> 0.97). For the set of C7 alkanes, where all molecules have the same number of bonds, all correlation coefficients with melting points are poor (< 0.4), and all those with boiling points are reasonably high ($r > 0.9$), except the correlation with $^{e}\chi$ ($r = 0.35$).

We included in Table **2** results of inter-correlational analysis of four connectivity indices studied here that were computed for a larger data set of 137 hadrocarbons (DS-137) taken from literature [14, 20, 22], and their correlation with three experimental properties of alkanes: reteintion index, melting and boiling point values. Additionally, we also give results of inter-correlational analysis between vertex- and edge-connectivity indices, as well as their correlation with six experimental properties of data sets of 134 alkanes (DS-137) [20-22]. Two sets of experimental values from literature were collected for boiling points in analysis on data set of 134 alkanes [20, 21], in order to test the dependence of correlation coefficients on the accuracy of experimental values used in comparative study. These both data sets [14, 21] were used in literature to test the applicability and suitability of topological descriptors in structure-property relationships.

In both data sets the lowest inter-correlation is between the vertex-product- ($^{v}\chi$) and edge-sum- ($^{e}\varepsilon$) connectivity indices ($r \sim 0.82$ and 0.72, for DS-137 and DS-134, respectively). Inter-correlation between all other pairs of indices is higher, having very high values ($r > 0.91$), with the highest value between the vertex-product- and vertex-sum-connectivity indices ($r > 0.999$).

Analysis of correlations between computed connectivity indices and experimental properties for data sets DS-137 and DS-134 clearly shows that in most cases (9 out of 10), the highest correlation coefficients are obtained with the sum-connectivity indices, *i.e.* five with the vertex-sum- ($^{v}\boldsymbol{\varepsilon}$) and four with the edge-sum- ($^{e}\boldsymbol{\varepsilon}$) connectivity indices.

From correlation of connectivity indices with boiling points on DS-134 it is evident that the level of correlation is not significantly dependent on the random error caused by using different measurements of experimental data. For both BP values, *i.e.* those from ref. 20 and from ref. 21, corresponding correlation coefficients are similar, and for the best one (with $^{v}\boldsymbol{\varepsilon}$) the difference is negligible (0.9661 *versus* 0.9666, respectively).

CONNECTIVITY POLYNOMIALS

An convenient method for computing the characteristic polynomial of any real and symmetric square matrix is the recursive approach by Le Verrier [23], Faddeev [24] and Frame [25] that was made very efficient by computer-based modifications of Balasubramanian [26] and Živković [27], respectively. Below we briefly delineate the modified Le Verrier-Faddeev-Frame method.

The *characteristic polynomial* P(*G*;x) of a graph *G* is defined as:

$$P(G;x) = \det \left| x\mathbf{I} - \mathbf{M} \right| \tag{10}$$

where **I** is the $N \times N$ unit matrix and **M** is a real, symmetric matrix. In this report **M** is one of the connectivity matrices considered.

The coefficient form of the characteristic polynomial is given by:

$$P(G;x) = c_0\, x^N - \sum_{i=1}^{N} c_n\, x^{N-n} \tag{11}$$

The coefficients of the characteristic polynomial can be obtained by the expansion of the determinant (10):

$$P(G;x) = c_0\, x^N - c_1\, x^{N-1} - c_2\, x^{N-2} - \ldots - c_{N-1}\, x - c_N \tag{12}$$

The polynomial coefficients can be computed using the following scheme:

$$c_n = \frac{1}{n}\sum_{i=1}^{N}(\mathbf{M}_n)_{ii} \tag{13}$$

where the diagonal form of the matrix $\mathbf{M}_n$ is given in terms of the diagonal forms of the initial matrix **M** and an auxiliary matrix $\mathbf{B}_n$:

$$(\mathbf{M}_n)_{ii} = (\mathbf{M})_{ii}\,(\mathbf{B}_n)_{ii} \tag{14}$$

Table 3: The computation of the vertex-product-connectivity polynomial of molecular graph G_1 given in Fig. (**1**), using the modified Le Verrier-Fadeev-Frame method

(1)	The spectrum of the matrix $\mathbf{V}(G_1)$ $\{1, 0.5132, 0, 0, -0.6921, -0.8212\}$
(2)	$c_0 = 1$ by definition
(3)	$\sum_{i=1}^{N} (\mathbf{V})_{ii} = \sum_{i=1}^{N} (\mathbf{V}_1)_{ii} = 0$ $c_1 = 0$
(4)	$\{(\mathbf{B}_1)_{ii} = (\mathbf{V}_1)_{ii} - (c_1\mathbf{I})\}_{i=1,\ldots,6} = \{1, 0.5132, 0, 0, -0.6921, -0.8212\}$ $\{(\mathbf{V}_2)_{ii} = (\mathbf{V})_{ii}\,(\mathbf{B}_1)_{ii}\}_{i=1,\ldots,6} = \{1, 0.2634, 0, 0, 0.4790, 0.6744\}$ $c_2 = ½ \sum_{i=1}^{N} (\mathbf{V}_2)_{ii} = 1.2084$
(5)	$\{(\mathbf{B}_2)_{ii} = (\mathbf{V}_2)_{ii} - (c_2\mathbf{I})\}_{i=1,\ldots,6} = \{-0.2084, -0.9450, -1.2084, -1.2084, -0.7294, -0.5340\}$ $\{(\mathbf{V}_3)_{ii} = (\mathbf{V})_{ii}\,(\mathbf{B}_2)_{ii}\}_{i=1,\ldots,6} = \{-0.2084, -0.4850, 0, 0, 0.5048, 0.4385\}$ $c_3 = 1/3 \sum_{i=1}^{N} (\mathbf{V}_3)_{ii} = 0.0833$
(6)	$\{(\mathbf{B}_3)_{ii} = (\mathbf{V}_3)_{ii} - (c_3\mathbf{I})\}_{i=1,\ldots,6} = \{-0.2917, -0.5683, -0.0833, -0.0833, 0.4215, 0.3552\}$ $\{(\mathbf{V}_4)_{ii} = (\mathbf{V})_{ii}\,(\mathbf{B}_3)_{ii}\}_{i=1,\ldots,6} = \{-0.2917, -0.2917, 0, 0, -0.2917, -0.2917\}$ $c_4 = 1/4 \sum_{i=1}^{N} (\mathbf{V}_4)_{ii} = -0.2917$
(7)	$\{(\mathbf{B}_4)_{ii} = (\mathbf{V}_4)_{ii} - (c_4\mathbf{I})\}_{i=1,\ldots,6} = \{0, 0, 0, 0, 0, 0\}$ $\{(\mathbf{V}_5)_{ii} = (\mathbf{V})_{ii}\,(\mathbf{B}_4)_{ii}\}_{i=1,\ldots,6} = \{0, 0, 0, 0, 0, 0\}$ $c_5 = 1/5 \sum_{i=1}^{N} (\mathbf{V}_5)_{ii} = 0$
(8)	The vertex-product-connectivity polynomial of G_1 $\mathrm{V}(G_1) = x^6 - 1.2084\,x^4 - 0.0833\,x^3 + 0.2917\,x^2$

The diagonal form of the auxiliary matrix $\mathbf{B}_n$ is given by:

$$(\mathbf{B}_n)_{ii} = (\mathbf{M}_n)_{ii} - (c_n\,\mathbf{I})_{ii} \tag{15}$$

The procedure ends when the **B**-matrix vanishes, *i.e.* when $n = N$:

$$(\mathbf{B}_N)_{ii} = (\mathbf{M}_N)_{ii} - (c_N\,\mathbf{I})_{ii} \tag{16}$$

As an illustrative example, we compute the vertex-product-connectivity polynomial of the graph G_1 (see Fig. (**1**)) using the above procedure. This is shown in Table **3**.

In the same way other three polynomials are computed. The edge-product-connectivity polynomial of G_1 is given by:

$$\mathrm{E}(G_1) = \mathrm{x}^6 - 0.8777\ \mathrm{x}^4 - 0.2667\ \mathrm{x}^3 + 0.0963\ \mathrm{x}^2 + 0.0444\ \mathrm{x} + 0.0037$$

Similarly, the polynomials of the vertex-sum-connectivity polynomial and edge-sum-connectivity polynomial of G_1 are given below:

$$\mathrm{S}(G_1) = \mathrm{x}^6 - 1.1595\ \mathrm{x}^4 - 0.1380\ \mathrm{x}^3 + 0.2216\ \mathrm{x}^2$$

$$\mathrm{C}(G_1) = \mathrm{x}^6 - 1.4243\ \mathrm{x}^4 - 0.6056\ \mathrm{x}^3 + 0.1956\ \mathrm{x}^2 + 0.1308\ \mathrm{x} + 0.0145$$

We list below several properties of connectivity polynomals that are easily detectable (note symbol M stands for the connectivity matrices considered in this report):

(1) The coefficients c_0 and c_1 are, respectively, always unity and zero. The c_1 coefficient is equal to zero because of the relationship:

$$c_1 = \sum_{i=1}^{N} x_i = \mathrm{tr}\ \mathbf{M} \tag{17}$$

where tr **M** is the trace of a given connectivity matrix.

(2) The coefficient c_2 is equal to the half-sum of the squares of elements of a given connectivity matrix:

$$c_2 = \frac{1}{2}\sum_{i,j=1}^{N} (\mathbf{M})_{ij}^2 \tag{18}$$

(3) The last coefficient c_N is equal to the determinant of a given connectivity matrix:

$$c_N = (-1)^N \det |\mathbf{M}| \tag{19}$$

(4) The sum of squares of the elements in the Harary matrix is equal to the trace of the squared Harary matrix:

$$\sum_{i,j=1}^{N} (\mathbf{M})_{ii}^2 = \mathrm{tr}\ \mathbf{M}^2 \tag{20}$$

CONCLUDING REMARKS

The following four connectivity matrices are reviewed: *vertex*-product-connectivity matrix, *edge*-product-connectivity matrix, *vertex*-sum-connectivity matrix and *edge*-sum-connectivity matrix. They are generated using degrees of vertices making up a simple graph. Half-sums of these four matrices give the four related connectivity indices, that is, *vertex*-product-connectivity index, *edge*-product-connectivity index, *vertex*-sum-connectivity index and *edge*-sum-connectivity index, respectively.

The inter-correlations and the usefulness and suitability of all forms of connectivity indices in developing structure-property relationships, especially of newer edge- and sum- forms of connectivity indices, are illustrated and confirmed on four data sets of alkanes for 14 experimental physico-chemical properties.

The corresponding connectivity polynomials are also generated using the time-honored approach by Le Verrier (1840) [23] with adaptations from Frame (1949) [24], Fadeeva (1959) [25], Balasubramanian (1986) [26] and Živković (1990) [27].

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science, Education and Sports of Croatia through Grants 098-1770495-2919 (BL and NT) and 098-0982562-2567 (IS).

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

ABBREVIATIONS

QSAR	=	Quantitative Structure-Activity Relationship
QSPR	=	Quantitative Structure-Property Relationship

REFERENCES

[1] Randić, M. Similarity based on extended basis descriptors. *J. Chem. Inf. Comput. Sci.*, **1992**, *32*, 686-692.

[2] Rodriguez, J.A. A spectral approach to the Randić index. *Lin. Alg. Appl*., **2005**, *400*, 330-344.

[3] Rodriguez, J.A.; Sigaretta J.M. On the Randić index and conditional parameters of a graph. *MATCH Commun. Math. Comput. Chem*., **2005**, *54*, 403-418.

[4] Hogben, L. Spectral graph theory and the inverse eigenvalue problem of a graph. *Electron. J. Lin. Algebra*, **2005**, *14*, 12-31.

[5] Bozkurt, Ş.B.; Güngör, A.D.; Gutman, I.; Cevik, A.S. Randić matrix and Randić energy. *MATCH Commun. Math. Comput. Chem*., **2010**, *64*, 239-250.

[6] Zhou, B.; Trinajstić, N. On a novel connectivity index. *J. Math. Chem*., **2009**, *46*, 1252-1270.

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

[8] Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, Wiley-VCH: Weinheim, **2000**, pp. 84-90.

[9] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemometrics*, Wiley-VCH: Weinheim, **2009**, pp. 161-172.

[10] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemometrics*, Wiley-VCH: Weinheim, **2009**, pp. 283-286.

[11] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemometrics*, Wiley-VCH: Weinheim, **2009**, pp. 478-487.

[12] Zhou, B.; Trinajstić, N. On sum-connectivity matrix and sum-connectivity energy of (molecular) graph. *Acta Chim. Slov*., **2010**, *57*, 518-523.

[13] Lučić, B.; Trinajstić, N.; Zhou, B. Comparison between the sum-connectivity and product-connectivity indices for benzenoid hydrocarbons. *Chem. Phys. Lett*., **2009**, *475*, 146-148.

[14] Lučić, B.; Nikolić, S.; Trinajstić, N.; Zhou, B.; Ivaniš Turk, S. Sum-connectivity index. In: *Novel Molecular Structure Descriptors – Theory and Applications I*; Gutman, I.; Furtula, B. Eds.; University of Kragujevac: Kragujevac, Serbia, **2011**, pp. 101-136.

[15] Randić, M.; Pisanski, T.; Novič M.; Plavšić, D. Novel graph distance matrix. *J. Comput. Chem*., **2010**, *31*, 1832-1841.

[16] Harary, F. *Graph Theory*, 2nd ed.; Addison-Wesley: Reading, PA, **1971**.

[17] Wilson, R.J. *Introduction to Graph Theory*, Oliver & Boyd: Edinburgh, **1972**.

[18] Trinajstić, N. *Chemical Graph Theory*, 2nd rev. ed.; CRC: Boca Raton, **1992**.

[19] Graph (mathematics). Wikipedia: The Free Encyclopedia. Wikimedia Foundation Inc. http://en.wikipedia.org/wiki/Graph_(mathematics) (accessed June 06, 2014).

[20] Rücker, G.; Rücker, C. On topological indices, boiling points, and cycloalkanes. *J. Chem. Inf. Comput. Sci*., **1999**, *39*, 788–802.

[21] Ivanciuc, O.; Ivanciuc, T.; Cabrol-Bass, D.; Balaban, A.T. Evaluation in quantitative structure-property relationship models of structural descriptors derived from information-theory operators. *J. Chem. Inf. Comput. Sci.*, **2000**, *40*, 631–643.

[22] Lučić, B.; Sović, I; Batista, J; Skala, K; Plavšić, D; Vikić-Topić, D; Bešlo, D; Nikolić, S.; Trinajstić, N. The additive variant of the Randić connectivity index. *Curr. Comput. Aided Drug Des.*, **2013**, *9*, 184-194.

[23] Le Verrier, U.J.J. Sur les variations séculaires des éléments elliptiques des sept planètes principales: Mercure, Vénus, la Terre, Jupiter, Saturne et Uranus. *J. Math. Pures Appl*., **1840**, *5*, 220-254.

[24] Frame, J.S. A simple recursion formula for inverting a matrix, contribution presented to American Mathematical Society at Boulder, Colorado (September 1, 1949) as referred to in P.S. Dwyer, *Linear Computations*, Wiley: New York, **1951**, pp. 225-235.

[25] Fadeeva, V.N. *Computational Methods in Linear Algebra*, Dover: New York, **1959**.

[26] Balasubramnian, K. On graph-theoretical polynomials in chemistry. In: *Mathematics and Computational Concepts in Chemistry*, Trinajstić. N. Ed.; Horwood: Chichester, **1986**, pp. 20.-33.

[27] Živković, T. On the evaluation of the characteristic polynomial of a chemical graph. *J. Comput. Chem*. **1990**, *11*, 217-222.

CHAPTER 5

The Use of Weighted 2D Fingerprints in Similarity-Based Virtual Screening

Shereena M. Arif[1,2], John D. Holliday[1] and Peter Willett[1,*]

[1] *Information School, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK and* [2] *Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia*

Abstract: The fingerprints that are widely used for similarity-based virtual screening typically encode the presence or absence of fragments, without any indication as to their relative importance. This chapter discusses the use of weighted fingerprints, where each fragment is associated with a weight denoting its degree of importance in quantifying the degree of similarity between a reference structure and a database structure. Extensive studies using the *World of Molecular Bioactivity* and *MDL Drug Data Report* databases show that weighting fragments according to their frequency of occurrence within a molecule can increase the effectiveness of screening, but that this is not the case when fragments are weighted according to their frequency of occurrence within a database.

Keywords: Chemoinformatics, ECFC4 fingerprint, extended connectivity fingerprint counts fingerprint, fingerprint, fragment weighting scheme, frequency weighting, IDF weighting, information retrieval, inverse frequency weighting, ligand-based virtual screening,/MDL Drug Data Report/database, similarity-based virtual screening, similarity coefficient, similarity searching, TF weighting, virtual screening, weighting scheme,/*World of Molecular Bioactivity*/database.

INTRODUCTION

An important component of modern drug-discovery programmes is *virtual screening* [1-5]. This involves the use of computational methods to rank a chemical database in order of decreasing probability of bioactivity, so that chemical synthesis and biological testing programmes can focus on those classes of molecules that are most likely to be relevant to a drug discovery programme.

***Corresponding author Peter Willett:** Information School, 211 Portobello Street, Sheffield S1 4DP, UK; Tel: 0044-114-2222633; Fax: 0044-114-2780300; E-mail: p.willett@sheffield.ac.uk

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50005-9

Database molecules are often represented by a *fingerprint*, a vector that encodes the presence or absence of a range of substructural fragments [6, 7], and one of the most common types of virtual screening involves the calculation of similarity measures based on such fingerprints [8-13]. Specifically, the fingerprint describing a *reference structure* that is known to exhibit the bioactivity of interest is compared with the fingerprints describing each of the database structures to identify those that are most similar, and that are hence the most likely to be active. The fragments encoded in a fingerprint can represent either two-dimensional (2D) or three-dimensional (3D) fragment substructures. Although the latter clearly provide a more detailed representation of molecular structure, 2D fingerprints, where the encoded fragments describe small patterns of atoms, bonds or rings [6, 7], have been found to be effective in operation. They continue to be very widely used for virtual screening, and hence form the basis for the work reported here.

Most types of fingerprints are binary in character, with each element of the vector denoting the presence (or absence) of a particular fragment substructure by the setting (or not setting) of one or more bits. However, this is not necessarily so, and the vector elements may instead contain a weight that reflects the importance of a particular fragment in determining the degree of similarity between a database structure, and the reference structure. Thus, a high-weight fragment occurring in both a reference structure and a database structure would contribute more to the overall degree of resemblance than would a low-weight fragment.

Fragment weighting has been widely used in some machine learning approaches to virtual screening [14], *e.g.*, substructural analysis involves calculating weights that represent probability of activity of a molecule that contains a specific fragment [15]. This is a powerful technique but one that requires the availability of extensive amounts of activity data to compute the probabilities. In similarity-based virtual screening (hereafter SBVS), conversely, the only such information available is the knowledge that the solitary reference structure is active. There is, however, an additional source of information available to the search algorithm in SBVS, *viz* the frequencies of occurrence of the fragments encoded in the fingerprints describing the reference structure and the database structures. This chapter describes an extended series of experiments to determine how such frequency information can best be used to increase the effectiveness of screening

[16, 17]. In the next section, we describe previous work on frequency-based weighting schemes, considering not only studies in SBVS but also in the design of text search engines, where such frequency information has long played a key role. We then briefly describe the experimental set-up we have used, before considering two types of weighting scheme: frequencies of occurrence within individual molecules, and frequencies of occurrence in a database as a whole. The chapter concludes with a summary of our major findings.

PREVIOUS STUDIES

The starting point for our work has been the extensive studies that have been carried out over many years in *information retrieval* (hereafter IR), which provides the technology for the search engines that are used to access the wealth of textual information now available on the World Wide Web [18, 19]. There are several ways in which IR is analogous to chemical virtual screening. For example, a textual document is indexed by a set of words selected from the much larger number that might possibly be used, and a chemical fingerprint encodes just those few fragments that are present in that molecule. Again, only a few documents in a text database are likely to be relevant to a user's query, and only a few molecules in a chemical database are likely to have the same bioactivity as a reference structure. Finally, and related to the second point, the analogy between relevance and bioactivity means that performance measures developed for evaluating the effectiveness of IR systems (in terms of the numbers of relevant and non-relevant documents retrieved) can also be used for evaluating the performance of systems for virtual screening (in terms of the numbers of active and inactive molecules retrieved). Given these resemblances, which have been discussed in some detail by Willett [20], it seems reasonable to consider whether the extensive work that has been carried out on frequency-based weighting schemes in IR might be applicable to SBVS.

Attempts to use frequency information in IR date back to the early Seventies when Spärck Jones carried out the first of a long series of experiments to investigate the extent to which word-frequency information could be used to enhance the effectiveness of retrieval systems [21, 22]. This work, and subsequent studies by Robertson [23, 24] and Salton [25, 26], demonstrated the effectiveness of two types of weighting scheme, called *idf* (for *inverse document frequency*) and *tf* (for *term frequency*) weighting. In idf weighting, each word in a document (or query) is associated with its inverse frequency of occurrence in the database that is being searched, so that rare words have larger associated weights than do words that are used more commonly. In tf weighting, each word in a document (or query) is associated with its frequency of occurrence, so that the document representation indicates not just a term's presence (or *incidence*) but also how often it is present (its *occurrence*). In the SBVS context, idf weighting makes the assumption that two molecules sharing a fragment that occurs only rarely in the database resemble each other more closely than if they share a commonly occurring fragment; and tf weighting makes the assumption that two molecules sharing multiple occurrences of a fragment in common resemble each other more closely than if they share just a single occurrence. Both of these assumptions seem entirely plausible.

Of the two types, idf-like weighting has been less studied in chemoinformatics. Both Adamson and Bush [27] and Willett and Winterman [28] found that it was not helpful in small-scale property prediction experiments. For database searching, Moock *et al*. found that it was highly effective for similarity searching in a large reactions database [29], Abdo and Salim have included inverse frequency counts as part of a scoring function to model probabilities of bioactivity in their work on Bayesian inference networks for SBVS [30], whilst Downs *et al*. obtained equivocal results in searches of three Pfizer screening datasets [31]. There have been several studies of tf-like weighting. The first report was by Willett and Winterman, who found that occurrence-based fingerprints were significantly superior to incidence-based fingerprints in property prediction experiments on small QSAR and QSPR datasets [28], and this finding was confirmed in other subsequent property prediction studies [32-35]. A similar conclusion in the context of database searching was reported by Chen and Reynolds [36], although Fechner *et al*. [37] and Stiefl *et al*. [38] found little difference between the two types of fingerprint.

The available evidence hence suggests that tf-like fingerprint weighting may enhance SBVS performance, when compared to conventional, incidence-based fingerprints, but that this may not be the case with idf-like weighting. That said, there is a fair degree of inconsistency in the results to date, and the experiments have often been limited in the sizes of the datasets used or in the extent to which the weighted and binary fingerprints differed. In the remainder of this chapter, we describe the experiments we have carried out to investigate these two types of weighting, which we shall refer to subsequently as *frequency weighting* (for tf) and *inverse frequency weighting* (for idf).

METHODS

The experimental set-up is typical of current work in SBVS. A bioactive reference structure is submitted, its similarity calculated with each of the database structures, the database ranked in order of decreasing similarity, and a cut-off applied to retrieve some fixed percentage of the top-ranked molecules. These nearest neighbours are then checked to determine whether they exhibit the same activity as the reference structure, and a measure of retrieval effectiveness determined. Two databases were used here: the *MDL Drug Data Report* database (hereafter MDDR, from Accelrys Inc. at http:/accelrys.com/) and the *World of Molecular Bioactivity* database (hereafter WOMBAT, from Sunset Molecular Discovery LLC at http:/sunsetmolecular.com/). There were 102,535 molecules in the MDDR dataset and 138,127 molecules in the WOMBAT dataset, as described in detail by Arif *et al*. [16].

Several activity classes were chosen for each of the two databases, as listed in Tables **1a** and **1b** for MDDR and WOMBAT, respectively. Each row of the table contains an activity class, the number of database molecules that have been listed as exhibiting that activity, and an indication of the class's diversity. The diversity figures listed are the mean intra-class similarities when all the members of a class are compared with each other and the similarities calculated using Tripos Unity 2D fingerprints and the Tanimoto coefficient (*vide infra*).

For each activity class, ten disparate example molecules were chosen to act as reference structures for simulated virtual screening. The numbers of actives retrieved in these similarity searches were then averaged over the ten reference structures, using cut-offs of the top-1% and the top-5% of the similarity rankings. We also noted the numbers of distinct ring systems in the active molecules that were retrieved, rather than just the number of active molecules, to assess the effectiveness of the various weighting schemes for scaffold-hopping applications [39, 40]. However, we found that the differences in screening effectiveness that were identified using the numbers of actives retrieved as the performance criterion mirrored closely the differences identified using the numbers of ring systems in those actives. We have hence included here the results only for the numbers of active molecules; full results are provided by Arif *et al*. [16, 17]. The MDDR and WOMBAT molecules were represented by ECFC4 circular substructure fingerprints (available from Accelrys Inc. at http:/www.accelrys.com), where ECFC denotes Extended Connectivity Fingerprint Counts. ECFC4 fingerprints encode circular substructures describing a central atom and all the atoms within a two-bond radius of it [41]: in our experiments, the integer codes representing a circular substructure were hashed to give a fixed-length fingerprint containing 1024 elements. The hashing results in very few collisions since ECFC fingerprints have a very low bit-density; and previous experiments have shown that the use of the 1024-element fingerprint results in only a minimal reduction in effectiveness compared to that obtained from the use of variable-length fingerprints where hashing is not used [42]. The fingerprint elements contained the weight associated with each fragment according to one of the various weighting schemes that are discussed in detail in the remainder of the chapter. Thus, an ECFC4 fingerprint can be considered as a vector, X (where X can denote either the reference structure or a database structure), with the i-th element denoting a fragment occurring f_i times in a molecule ($f_i \geq 0$). The f_i values may then be modified by the application of a weighting scheme to the fingerprint.

Table 1: Activity classes used in the virtual screening experiments, chosen from the (a) MDDR and (b) WOMBAT databases

Activity Class	Active Molecules	Mean Similarity
5HT3 antagonists	752	0.35
5HT1A agonists	827	0.34
5HT reuptake inhibitors	359	0.35
D2 antagonists	395	0.35
Renin inhibitors	1125	0.57
Angiotensin II AT1 antagonists	943	0.40
Thrombin inhibitors	803	0.42
Substance P antagonists	1246	0.40

Table 1: contd….

HIV protease inhibitors	750	0.45
Cyclooxygenase inhibitors	636	0.27
Protein kinase C inhibitors	453	0.32

(a)

Activity Class	**Active Molecules**	**Mean Similarity**
5HT3 antagonists	220	0.38
5HT1A antagonists	592	0.40
D2 antagonists	910	0.37
Renin inhibitors	474	0.59
Angiotensin II AT1 antagonists	724	0.44
Thrombin inhibitors	421	0.42
Substance P antagonists	558	0.43
HIV protease inhibitors	1128	0.44
Cyclooxygenase inhibitors	965	0.32
Protein kinase C inhibitors	142	0.57
Acetylcholine esterase inhibitors	503	0.37
Factor Xa inhibitors	842	0.39
Matrix metalloprotease inhibitors	694	0.44
Phosphodiesterase inhibitors	596	0.36

(b)

Given this representation, the similarity S_{XY} between the vectors R and D, representing a reference structure and a database structure respectively, was computed using the full form of the widely used Tanimoto coefficient [43-45],

$$S_{XY} = \frac{\sum r_i d_i}{\sum r_i^2 + \sum d_i^2 - \sum r_i d_i} \tag{1}$$

where the summations are over all of the elements in each fingerprint.

INVERSE FREQUENCY WEIGHTING

Inverse frequency weighting has been widely used in IR (see Previous Studies section), and we have used three IR weighting schemes here, as well as a further such weight that has previously been used specifically for chemoinformatics searching.

Considering each ECFC4 fingerprint as a vector, *X*, as described above, the inverse frequency weighting experiments used five different weighting schemes (w1-w5). The conventional, zero/one binary weight is represented by:

$$\mathrm{w1}: x_i = 1 .$$

This binary weight encodes just the incidence of the *i*-th fragment, and is obtained by setting to unity all elements in *X* for which the corresponding fragment occurred one or more times. The following weights (w2-w5) all involve replacing the value of 1 for a fragment occurring in a molecule by a weight reflecting the fragment's inverse frequency of occurrence in the database as a whole.

Assume that the *i*-th fragment occurs in a total of T_i molecules ($T_i \geq 0$) in the database and that the database comprises a total of *N* different molecules. Then the three IR-derived weights are as follows, where ln denotes natural logarithms.

$$\mathrm{w2}: x_i = \ln\left(\frac{N}{T_i + 1}\right) \tag{2}$$

$$\mathrm{w3}: x_i = \ln\left(\frac{N}{T_i}\right) + 1 \tag{3}$$

$$\mathrm{w4}: x_i = \ln\left(\frac{N + 0.5}{T_i + 0.5}\right). \tag{4}$$

There are clear relationships between these four weights: w2 and w4 yield comparable weights except where a fragment occurs very infrequently across the whole database; while w3 is simply w1 augmented by an inverse frequency component. In addition to their successful origins in IR, the use of logarithmic inverse frequency functions is further supported by consideration of the fragment weighting schemes used in substructural analysis approaches to virtual screening. As noted previously, substructural analysis requires large amounts of activity information; however, Arif *et al*. demonstrate that the mathematical formulations of two of these weights reduce to functions involving ln($1/T_i$) in the absence of such information [17].

The final weight, w5, is rather different in form, and is that used by Moock *et al.* in their work on similarity searching of chemical reaction databases [29]:

$$\text{w5}: x_i = \sqrt{\frac{\text{Max}\{T_i\}}{T_i}} \tag{5}$$

where Max$\{T_i\}$ is the number of molecules containing the most frequently occurring fragment.

Each of the five schemes, w1-5, can be applied to the reference structure and to each of the database structures, giving a total of 25 different combined weighting schemes that could be used for SBVS. Here, we have considered those schemes where both the reference structure and the database structures are weighted in the same manner; in addition, we have considered those where either the reference structure or the database structures are weighted using w1, *i.e.*, the use of conventional binary weighting. This gives a total of 13 different combinations for evaluation. An individual combined weight is referred to subsequently as W*ab*, where *a* denotes the weight applied to the database structures' fingerprints and *b* denotes the weight applied to the reference structure's fingerprint. For example, W15 represents the searches (ten of them for each of the chosen activity classes) where the database structures use w1 (*i.e.*, conventional binary weighting) and where the reference structures use the reaction searching weight, w5.

A typical set of search results is shown in Table **2**, for searches of the MDDR dataset in which the molecules comprising the top-1% of the ranked database are checked for activity and in which each of the columns in the main body of the table is headed by a three-letter abbreviation for the activity class. The results show the mean numbers of actives when averaged over the ten different reference structures for each of the eleven activity classes. The penultimate column gives the mean when averaged over all of the 110 searches for each weighting scheme W*ab*. The largest number of actives in each column is heavily-shaded, and elements that are within 5% of this largest value are lightly-shaded. The last column gives the total number of shaded elements for each combination of weights, and has been included for the following reason. Table **1** demonstrates that there are large variations in the numbers of actives and in the inter-molecular similarities for each of the activity classes. Most obviously, the MDDR renin dataset contains many highly similar molecules, making it easy for all the weighting schemes to retrieve large numbers of actives, as clearly demonstrated by the column headed REN in Table **2**. This could bias the results when the arithmetic mean is calculated, and hence the number of shaded cells has been included as this indicator is based solely upon the ranking of the various weighting schemes, and not upon the absolute numbers of actives.

Table **2** exemplifies the results obtained using one combination of parameters (top-1% of MDDR). The results obtained for all combinations (MDDR or WOMBAT, top-1% or top-5%) are shown in Table **3**, with the largest valued cells again being shaded. Inspection of Table **3** shows clearly that the best results, as denoted by the much greater prevalence of shaded cells in the upper part of the table, are generally obtained using weights of the form W1*b*, *i.e.*, with the database structures unweighted and with the inverse frequency weights applied only to the fragments occurring in the reference structure. Overall, the best results are obtained using W13 (for the MDDR activity classes) or W11 (for the WOMBAT activity classes) based on the number of actives retrieved. Based on the ranks, W11 and W13 again perform well (although several other combinations of weights are comparable to W13 in the WOMBAT searches). The best overall performance would seem to occur with W11, *i.e.*, not using weights for either the reference structure or the database structures.

Inspection of Tables **1a** and **2** suggests that W11 tends to perform well, when compared to W13, when the mean pair-wise similarities for the active molecules is low (*i.e.*, when searches are being carried out for the more diverse activity classes). To confirm this observation, the MDDR and WOMBAT activity classes were divided into two sets: those with similarities in Table **1** ≥ 0.40 (homogeneous activity classes, *i.e.*, where the molecules in a class are structurally similar) and those with similarities < 0.40 (heterogeneous activity classes, *i.e.*, where the molecules in a class are structurally diverse). The ten searches for each activity class means that there are 50 MDDR homogeneous searches and 60 heterogeneous searches, and 70 WOMBAT searches for both homogeneous and heterogeneous. We have then taken all the top-1% searches using W11 and W13 and compared the numbers of actives retrieved to see which of the two weighting schemes performed better. For each set of searches (W11 and W13) we have compared the results using both the Sign and Wilcoxon tests: the Sign test simply compares the number of searches where one of the measures was superior to the other, while the Wilcoxon test additionally takes account of the magnitude of the difference in each case. The significance of the differences in each case is assessed using a *Z* test [46]. For example, for the 50 MDDR homogeneous-class searches, W13 does better than W11 for 34 of the searches, W11 does better than W13 for 14 of the searches, and there is no difference in the remaining two searches. Both the Wilcoxon and Sign test show that W13 is significantly better ($p <= 0.01$) than W11 for these homogeneous activity classes. Conversely, both tests show that W11 is significantly better ($p <= 0.001$) for the MDDR heterogeneous classes. There are no significant differences for the WOMBAT homogeneous classes, while the Sign test (but not the Wilcoxon test) shows that W11 is better than W13 for the WOMBAT heterogeneous classes.

Table 2: Mean numbers of actives retrieved in the top-1% of the rankings obtained from 10 searches of each of the 11 MDDR activity classes, using inverse frequency weighting

	5HT3	5HT1	5HT	D2	REN	ANG	THR	SUBP	HIV	COX	PKC	Mean	Shaded
W11	90.2	81.1	24.0	27.2	419.8	236.1	56.5	121.3	86.5	28.3	35.5	109.7	7
W12	75.3	71.7	21.0	25.5	468.5	218.7	59.6	145.3	91.2	20.9	27.9	111.4	4
W13	73.4	71.5	20.4	25.9	480.6	231.5	60.1	140.5	92.9	19.1	27.6	113.1	5
W14	75.0	71.9	21.1	25.4	468.8	219.0	59.8	145.1	91.2	21.1	28.1	111.5	4
W15	71.9	65.4	19.7	24.1	464.7	209.6	58.9	144.0	92.3	18.0	26.0	108.6	4
W21	84.9	76.6	21.0	25.4	311.2	173.5	50.3	111.2	74.1	27.9	30.9	89.7	1
W22	78.3	73.9	21.0	23.7	398.7	176.6	56.2	136.5	87.6	23.8	27.7	100.4	0
W31	85.3	73.9	20.0	24.0	249.8	157.2	44.7	98.5	66.0	29.3	30.6	79.9	1
W33	83.2	77.4	21.2	26.0	415.8	196.1	57.5	137.4	89.9	25.3	29.7	105.4	3
W41	84.8	76.2	21.0	25.4	311.4	173.6	49.6	111.0	73.9	28.1	30.9	89.6	1
W44	78.3	73.8	21.0	23.8	398.3	176.6	56.4	136.9	87.6	23.7	27.7	100.4	0
W51	76.9	70.2	20.5	24.3	254.0	139.9	42.9	94.4	63.2	27.9	28.2	76.6	1
W55	72.5	61.6	20.0	20.6	356.3	158.5	54.9	122.6	83.6	21.5	24.8	90.6	0

Table 3: Mean numbers of actives retrieved and numbers of shaded cells using inverse frequency weighting

	MDDR				WOMBAT			
	Top-1%		Top-5%		Top-1%		Top-5%	
	Actives	Cells	Actives	Cells	Actives	Cells	Actives	Cells
W11	109.7	7	211.9	6	103.6	10	188.2	8
W12	111.4	4	205.5	3	101.2	7	184.9	6
W13	113.1	5	212.8	5	100.8	6	187.8	8
W14	111.5	4	198.2	3	101.0	7	184.6	7
W15	108.6	4	202.5	3	98.5	2	181.8	6
W21	96.3	1	183.9	1	85.6	3	166.2	3
W22	100.4	0	184.3	0	95.9	6	171.1	2
W31	79.9	1	173.8	3	75.6	2	155.1	2
W33	105.4	3	197.0	0	100.4	7	180.3	5
W41	89.6	1	183.9	1	85.3	2	166.6	4
W44	100.4	0	184.3	0	96.1	6	171.2	2
W51	76.6	1	166.7	1	68.5	1	143.5	0
W55	90.6	0	171.3	0	85.8	1	153.6	1

The experiments discussed here have used just a single type of fingerprint, *i.e.*, those based on the ECFC4 circular substructures. Arif *et al*. [17] report additional experiments that employed three other types of fingerprint: BCI keys (1052 elements and available from Digital Chemistry Ltd.) are selected to maximise discrimination in substructure searching using a frequency-based selection algorithm; MDL keys (166 elements and available from Accelrys Inc.) encode common fragment substructures; and Sunset keys (560 elements and available from Sunset Molecular Discovery LLC) combine chemical substructure recognition with topologically-relevant pharmacophore patterns based on atom-pairs. The conclusions that can be drawn from using these additional fingerprint types are broadly in line with those obtained with the ECFC4 fingerprints (although there are differences of detail occasioned by the very different statistical characteristics of the various molecular representations).

It hence seems reasonable to conclude that W13 provides an effective weighting scheme when searching for structurally homogeneous sets of actives, but that W11 is the method of choice for the more challenging, and pharmaceutically much more important, task of searching for structurally diverse sets of actives. A theoretical rationale for why the diversity of the actives could affect the relative performance of these two weighting schemes is outlined in the Appendix to this chapter.

FREQUENCY WEIGHTING

We now turn to frequency weighting where, as before, we consider each fingerprint as a vector, *X*, with the *i*-th fragment occurring f_i times in a molecule. The experiments used the following five weighting schemes (w1-w5). First, the incidence weight:

$$\text{w1: } x_i = 1 \tag{6}$$

Second, the occurrence weight is obtained by setting,

$$\text{w2: } x_i = f_i\,, \tag{7}$$

i.e., using the raw occurrence counts. w1 and w2 are the obvious weighting schemes, and the ones that are normally meant when binary and weighted fingerprints are referred to in the chemoinformatics literature. However, we also consider three further ways in which the occurrence frequencies can be used. The first two are standard normalisations in data analysis, and involve taking either the natural logarithm,

$$\text{w3: } x_i = \ln(f_i) \tag{8}$$

or the square root,

$$\text{w4: } x_i = \sqrt{f_i} \tag{9}$$

of the frequency of occurrence. The final scheme, w5, expresses the raw occurrence frequency as a fraction of the frequency for the most frequently occurring fragment in the molecule, and then normalised to give a value between 0.5 and 1. This approach has been tested here since it has been widely used in IR weighting studies [25, 26]:

$$\text{w5: } x_i = 0.5 + 0.5\frac{f_i}{\max\{f_i\}}. \tag{10}$$

The reference structure and each database structure can be weighted using each of these five weighting schemes, yielding a total of 25 possible different similarity measures. As with the inverse frequency weighting experiments, we have chosen to use those combinations where both the reference structure and the database structures were weighted using the same weighting scheme, together with all combinations involving w1 or w2 (*i.e.*, involving simple binary weighting or raw frequency counts). This gives a total of 19 different combinations for evaluation. As before, W*ab* describes the combined weighting scheme used for SBVS, where *a* denotes the weight applied to the database structures' fingerprints and *b* the weight applied to the reference structure's fingerprint.

The frequency weighting experiments have been analysed as described in the previous section for the inverse frequency weighting experiments, and the results in Table **4** for the numbers of retrieved actives are hence analogous to those shown in Table **3**. Inspection of Table **4** shows that effective screening is obtained with the following weighting schemes: W12, W14, W44, W51 and W52, with W12 being probably the best overall.

The results in Table **4** relate to the use of ECFC4 fingerprints. Arif *et al*. [16] report detailed experiments using two further fingerprints that provide occurrence information, these being the Sunset keys (*vide supra*) and Tripos holograms (997 elements available from Tripos Inc.). There is some variation in the results obtained from the three different types of representation (and more so than when the different fingerprints were tested in the inverse frequency weighting experiments discussed in the previous section). When the Sunset keys were used, the best results were obtained with W14, W44, W51 and W55; while W22, W33 and W44 were the weighting schemes of choice with the Tripos holograms. Taking the three fingerprints together, Arif *et al*. concluded that W44 gave the best overall screening performance, *i.e.*, that the fingerprints should encode the square root of the raw occurrence counts for both the reference structure and the database structures. The effect of the w4 weight is to lessen the contribution of the more generic fragments that can occur relatively frequently within molecules, and that thus yield high element values if the raw occurrence counts are used without some form of normalisation.

Table 4: Mean numbers of actives retrieved and numbers of shaded cells using frequency weighting

	MDDR				WOMBAT			
	Top-1%		Top-5%		Top-1%		Top-5%	
	Actives	Cells	Actives	Cells	Actives	Cells	Actives	Cells
W11	109.7	1	211.9	1	103.6	3	188.2	4
W12	118.7	5	227.2	2	108.2	6	193.4	4
W13	29.0	0	95.2	1	26.2	0	85.1	0
W14	114.9	3	219.4	2	105.8	3	191.1	3
W15	88.1	1	183.3	1	89.7	1	163.7	0
W21	50.7	1	126.4	1	50.0	0	116.0	0
W22	86.2	2	185.8	4	86.0	2	165.8	2
W23	13.6	0	59.1	0	9.0	0	40.7	0
W24	62.7	1	142.8	1	62.1	1	133.7	1
W25	25.0	0	76.2	1	26.4	0	66.8	0
W31	88.4	2	197.6	2	55.4	0	154.3	1
W32	55.0	0	171.0	1	25.0	0	122.8	0
W33	69.1	0	166.7	2	71.3	1	158.9	1
W41	109.3	2	215.0	3	100.8	0	186.7	1
W42	99.4	1	213.7	2	82.0	1	172.2	0
W44	114.6	5	223.5	3	103.0	1	192.6	3
W51	119.9	4	226.8	2	107.2	6	196.0	6
W52	115.6	2	222.5	1	104.7	4	193.7	6
W55	113.0	1	208.3	1	103.0	0	188.8	3

Arif *et al*. present a detailed discussion of how even quite small variations in the weighting scheme can affect the magnitudes of the Tanimoto coefficients that are calculated during a screening experiment [16]. Their analysis reveals a complex pattern of relationships that means, for example, that if there is a large discrepancy in the weights computed using the weighting schemes for the reference structure and for the database structure then screening effectiveness is likely to be less than if the two weights are comparable in magnitude.

The Tanimoto coefficient is normally regarded as being symmetric in character, in that the two objects that are being compared in a similarity calculation are treated analogously. However, this will only be so if the two object descriptions are weighted in the same way, and it will be clear from the formulations of w1-w5 that this is not always the case, with the result that there can be considerable differences in the magnitudes of the three components of the denominator of the Tanimoto expression. Thus, whereas the choice of the Tanimoto coefficient for virtual screening would normally be regarded as being non-problematic, the evidence here, as with the inverse frequency weights and as detailed in the Appendix, would suggest that the mathematical form of the coefficient may affect screening performance in ways that are far from obvious on first inspection.

CONCLUSION

Lead-discovery programmes in the pharmaceutical and agrochemical industries make extensive use of SBVS based on 2D fingerprints. The fingerprints are normally binary in character, encoding just the incidence of fragments in a molecule; however, they can also be weighted to reflect the relative degree of importance of different fragments in determining the degree of resemblance between two molecules.

In this chapter, we have summarised the principal findings of a project to evaluate the effectiveness of two approaches to the weighting of fingerprints. Drawing on previous work in IR, we have studied both inverse frequency weighting, where fragments that occur infrequently throughout a database are assumed to be more important than common fragments that occur in many molecules, and frequency weighting, where fragments that occur multiple times within a molecule are assumed to be important than fragments that occur only once in a molecule. Our conclusions are threefold. First, whilst inverse frequency weighting seems intuitively reasonable, it is of little practical use: it appears to be effective only where the active molecules are structurally similar but offers few benefits in the normal environment where the actives are structurally diverse. Second, frequency weighting is often beneficial, as long as the weights applied to the reference structure and database structure fingerprints are broadly comparable in magnitude, with the best results being obtained by taking the square roots of the occurrence frequencies. Third, although widely used, the effectiveness of the Tanimoto coefficient for virtual screening is crucially dependent on the precise weighting scheme that is used. As a consequence of this last finding, current work in our laboratory is investigating the use of other similarity coefficients for frequency-weighted SBVS.

APPENDIX

This appendix provides a theoretical rationale for why the search effectiveness of the inverse frequency weighting schemes W11 and W13 might be differentially affected by the structural diversity of the actives that are being sought in SBVS. Specifically, we believe that these differences arise from the nature of the Tanimoto coefficient.

For W11 the coefficient has the usual form, *i.e.*,

$$\frac{\sum r_i d_i}{\sum r_i^2 + \sum d_i^2 - \sum r_i d_i} \tag{11}$$

In a similarity search, a single reference structure is matched against each of the different database structures and hence the $\sum r_i^2$ term will be a constant, call it c in this context. When the actives are similar to each other, there are likely to be many fragments in common between the reference structure and the active database structures. This means that the $\sum r_i d_i$ term will be large for the top-ranked database structures; indeed, it may be comparable in magnitude to the $\sum d_i^2$ term, which would result in the coefficient using W11 being very approximately,

$$\frac{\sum r_i d_i}{c} \tag{12}$$

With a diverse set of actives the similarities are unlikely to be large on average, and thus the contribution from the matching fragments, *i.e.*, the $\sum r_i d_i$ term, is likely to be small. The coefficient using W11 in this situation is hence very approximately

$$\frac{\sum r_i d_i}{c+\sum d_i^2} \tag{13}$$

The form of the w3 weight means that fragments weighted using this scheme will have much higher weights than they will if weighted using w1, with the result that when W13 is used, *i.e.*, when the reference structure fingerprint is weighted and the database structure fingerprint is unweighted, $\sum r_i^2$ will typically be much greater than $\sum d_i^2$. The Tanimoto coefficient will hence be approximately,

$$\frac{\sum r_i d_i}{\sum r_i^2 - \sum r_i d_i} \tag{14}$$

As noted above, the $\sum r_i^2$ term in the denominator is a constant, and the Tanimoto coefficient is thus given by,

$$\frac{\sum r_i d_i}{c-\sum r_i d_i}. \tag{15}$$

Consider what happens when the actives are structurally diverse. Here, the reference structure and the active database structures will generally have few fragments in common: the *c* term in the denominator will then be much larger than $\sum r_i d_i$ and the coefficient will hence be approximately,

$$\frac{\sum r_i d_i}{c} \tag{16}$$

The approximations above are clearly gross; however, they do suggest that the Tanimoto similarities calculated using the W11 and W13 inverse frequency weighting schemes may be differentially affected by changes in the diversity of the sets of active molecules that are being sought in a similarity search.

ACKNOWLEDGEMENTS

Declared None.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

ABBREVIATIONS

BCI	=	Barnard Chemical Information
ECFC	=	Extended Connectivity Fingerprint Counts
IDF	=	Inverse document frequency
IR	=	Information retrieval
MDDR	=	*MDL Drug Data Report*
MDL	=	Molecular Design Limited
SBVS	=	Similarity-based virtuals creening
TF	=	Term frequency
WOMBAT	=	*World of Molecular Bioactivity*

REFERENCES

[1] Alvarez, J.; Shoichet, B. *Virtual Screening in Drug Discovery*. CRC Press: Boca Raton, **2005**.

[2] Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010,** *50*, 205-216.

[3] McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504-1519.

[4] Rippenhausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010,** *53*, 8461-8467.

[5] Schneider, G. Virtual screening: an endless staircase? *Nature Rev. Drug Discov.* **2010,** *9*, 273-276.

[6] Gasteiger, J. *Handbook of Chemoinformatics*. Wiley-VCH: Weinheim, **2003**.

[7] Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*. 2nd edition ed.; Kluwer: Dordrecht, **2007**.

[8] Bender, A. How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin. Drug Discov.* **2010,** *5*, 1141-1151.

[9] Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W. Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **2010,** *29*, 157-170.

[10] Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitation and novel approaches. *Drug Discov. Today* **2007,** *12*, 225-233.

[11] Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J. Chem. Inf. Model.* **2010,** *50*, 771-748.

[12] Stumpfe, D.; Bajorath, J. Similarity searching *Wiley Interdisc. Rev.: Comp. Mol. Sci.* **2011,** *1*, 260-282.

[13] Willett, P. Similarity methods in chemoinformatics. *Ann. Rev. Inf. Sci. Technol.* **2009,** *43*, 3-71.

[14] Goldman, B. B.; Walters, W. P. Machine learning in computational chemistry. *Ann. Report. Comp. Chem.* **2006,** *2*, 127-140.

[15] Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* **1974,** *17*, 533-535.

[16] Arif, S. M.; Holliday, J. D.; Willett, P. Analysis and use of fragment occurrence data in similarity-based virtual screening. *J. Comp.-Aid. Mol. Design* **2009,** *23*, 655-668.

[17] Arif, S. M.; Holliday, J. D.; Willett, P. Inverse frequency weighting of fragments for similarity-based virtual screening. *J. Chem. Inf. Model.* **2010,** *50*, 1340-9.

[18] Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*. 2nd edition ed.; Addison-Wesley: Harlow, **2010**.

[19] Manning, C. D.; Raghavan, P.; Schütze, H., *Introduction to Information Retrieval*. Cambridge University Press: Cambridge, **2007**.

[20] Willett, P., Textual and chemical information retrieval: different applications but similar algorithms. In *Inf. Res.* **2000,** *5*, at http:/InformationR.net/ir/5-2/infres52.html.

[21] Spärck Jones, K. A statistical interpretation of term specificity and its application in retrieval. *J. Docum.* **1972,** *28*, 11-21.

[22] Spärck Jones, K. Index term weighting. *Inf. Stor. Ret.* **1973,** *9*, 616-633.

[23] Robertson, S. E.; Spärck Jones, K. Relevance weighting of search terms. *J. Amer. Soc. Inf. Sci.* **1976,** *27*, 129-146.

[24] Spärck Jones, K.; Walker, S.; Robertson, S. E. A probabilistic model of retrieval: development and comparative experiments. *Inf. Proc. Manag.* **2000,** *36*, 779-840.

[25] Salton, G.; Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Proc. Manag.* **1988,** *24*, 513-523.

[26] Salton, G., *Automatic Text Processing*. Addison-Wesley: Reading, MA, **1989**.

[27] Adamson, G. W.; Bush, J. A. A comparison of the performance of some similarity and dissimilarity measures in the automatic classification of chemical structures. *J. Chem. Inf. Comp. Sci.* **1975,** *15*, 55-58.

[28] Willett, P.; Winterman, V. A comparison of some measures of inter-molecular structural similarity. *Quant. Struct.-Activ. Relat.* **1986,** *5*, 18-25.

[29] Moock, T. E.; Grier, D. L.; Hounshell, W. D.; Grethe, G.; Cronin, K.; Nourse, J. G.; Theodosiou, J. Similarity searching in the organic reaction domain. *Tetrahedron Comp. Methodol.* **1988,** *1*, 117-128.

[30] Abdo, A.; Salim, N. Similarity-based virtual screening with a Bayesian inference network. *Chem. Med. Chem.* **2009,** *4*, 210-218.

[31] Downs, G. M.; Poirrette, A. R.; Walsh, P.T.; Willett, P. Evaluation of similarity searching methods using activity and toxicity data. In: *Chemical Structures 2, The International Language of Chemistry.*, Warr, W. A., Ed. Springer-Verlag: Berlin, **1993**; pp 409-421.

[32] Brown, R. D.; Martin, Y. C., Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comp. Sci.* **1996,** *36*, 572-584.

[33] Olah, M.; Bologa, C.; Oprea, T. I. An automated PLS search for biologically relevant QSAR descriptors. *J. Comp.-Aid. Mol. Design* **2004,** *18*, 437-449.

[34] Ewing, T. J. A.; Baber, J. C.; Feher, F. Novel 2D fingerprints for ligand-based virtual screening. *J. Chem. Inf. Model.* **2006,** *46*, 2423-2431.

[35] Azencott, C.-A.; Ksikes, A.; Swamidass, S. J.; Chen, J. H.; Ralaivola, L.; Baldi, P. One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical and biological properties. *J. Chem. Inf. Model.* **2007,** *47*, 965-974.

[36] Chen, X.; Reynolds, C. H. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comp. Sci.* **2002,** *42*, 1407-1414.

[37] Fechner, U.; Paetz, J.; Schneider, G. Comparison of three holographic fingerprint descriptors and their binary counterparts. *QSAR Combin. Sci.* **2005,** *24*, 961-967.

[38] Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A., ErG: 2D pharmacophore descriptions for scaffold hopping. *J. Chem. Inf. Model.* **2006,** *46*, 208-220.

[39] Brown, N.; Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini-Rev. Med. Chem.* **2006,** *6*, 1217-1229.

[40] Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *J. Med. Chem.* **2010,** *53*, 5707-5715.

[41] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010,** *50*, 742-754.

[42] Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2,* 3256-3266.

[43] Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comp. Sci.* **1998,** *38*, 983-996.

[44] Maggiora, G. M.; Shanmugasundaram, V. Molecular similarity measures. *Methods Mol. Biol.* **2010,** *672*, 39-100.

[45] Willett, P., Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **2006,** *11*, 1046-1053.

[46] Siegel, S.; Castellan, N. J., *Nonparametric Statistics for the Behavioural Sciences*. Second ed.; McGraw-Hill: New York, **1988**.

CHAPTER 6

MOLGEN 5.0, A Molecular Structure Generator

Ralf Gugisch[1], Adalbert Kerber[1,*], Axel Kohnert[1], Reinhard Laue[1], Markus Meringer[2], Christoph Rücker[3] and Alfred Wassermann[1]

[1]Department of Mathematics, University of Bayreuth, Bayreuth, Germany; [2]German Aerospace Center (DLR), Oberpfaffenhofen, Wessling, Germany and [3]Institute of Sustainable and Environmental Chemistry, Leuphana University Lüneburg, Lüneburg, Germany

Abstract: MOLGEN 5.x combines the efficiency of the molecular generator MOLGEN 3.5 and the flexibility of MOLGEN 4.x. To achieve this, the software was reimplemented based on a totally new concept. The most visible new features are fuzzy molecular formula input and explicit use of atom state patterns. We describe the version MOLGEN 5.0 of this new series.

Keywords: MOLGEN, structure generation, fuzzy molecular formula, atom state pattern, molecular graph, goodlist, badlist, backtracking, diophantine equation, orderly generation, molecular libraries, connectivity isomers, constitutions, molecular structure elucidation, substructure restriction, aromaticity detection.

INTRODUCTION

The program system MOLGEN is devoted to generating all structures (connectivity isomers, constitutions) that correspond to a given molecular formula, with optional further restrictions, *e.g.* presence or absence of particular substructures.

MOLGEN arose from the idea to provide an efficient and portable tool for molecular structure elucidation in chemical industry, research, and education. Historically, up to version MOLGEN 3.5, the main intention was to generate structures as fast as possible. The result is one of the fastest generators for

***Corresponding author Adalbert Kerber:** Department of Mathematics, University of Bayreuth, D-95440 Bayreuth, Germany; Tel: 0049 921 68009; Fax: 0049 921 55 3385; E-mails: kerber@uni-bayreuth.de; adalbert-kerber@t-online.de

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50006-0

molecular structures. However, applications showed that generator efficiency is not the only important topic for molecular structure elucidation. Thus, in the development of series MOLGEN 4.x [1, 2] the interface was organized in a much more flexible way. Now advanced restrictions can be passed to the generator that are obtained from spectroscopy. MOLGEN–MS and MOLGEN–QSPR [3, 4] are special versions that arose from these efforts. In generating huge libraries without advanced restrictions, the performance of MOLGEN 4.x is not comparable to that of MOLGEN 3.5. Series MOLGEN 5.x is now intended to combine the advantages of both approaches, *i.e.* the efficiency of MOLGEN 3.5 and the flexibility of MOLGEN 4.x.

All MOLGEN versions provide the mathematical heart of a program system for structure elucidation, rendering all mathematically possible candidates that correspond to a given set of structural constraints. MOLGEN allows computing the complete set of structures corresponding to a given molecular formula or a set of molecular formulas. Often the molecular formula is sufficient as input, the generator will then use default values for the valences of all atoms included. Of course, it is possible to override defaults, by *e.g.* specifying particular atom valences.

The generation is *free of redundance*, *i.e.* no structure is generated twice within a single run. Moreover, the construction is *complete*, which means that the full set of all possible structures is obtained that correspond to a given molecular formula and, optionally, further restrictions. For example, given the input

$$C_8H_{16}O_2$$

each MOLGEN version will construct exactly 13,190 pairwise different structures. This example already shows that, in general, the number of structures corresponding to a given molecular formula is very large. Therefore it is often desirable to reduce the output by imposing additional restrictions. For this purpose, together with a molecular formula, substructures may be specified that must be contained in each isomer constructed, or that on the contrary are not allowed. For example, if together with molecular formula $C_8H_{16}O_2$ a carboxyl group is prescribed, exactly 39 structures will be generated. If additionally the isopropyl group is excluded, then out of the 39 structures just 27 will remain.

Sometimes, compounds of interest are not described by a single molecular formula. For example, we may be interested in all chlorinated biphenyls, or even in all halogenated small alkanes with up to four carbon atoms. The present version MOLGEN 5.0 was developed to solve such problems. Solutions for these examples are presented in the applications section.

An important issue is, of course, how far MOLGEN 5.0 will reach. The only noteworthy limitations are those of time and hardware, *i.e.* due to an astronomical number of solutions, the program may not be able to generate the complete set of structures for a molecular formula within a reasonable time or to store all structures on the given hard disk.

MOLGEN 5.0 runs under Microsoft Windows (XP, Vista, 7, 8) and Linux operating systems. Generated structures are written in MDL SDfile (.sdf) or in the MOLGEN MB4 (.mb4) file format. Details on installation and hardware requirements can be found in the manual, to be obtained from

http://www.molgen.de

where the interested reader can also play with a restricted online version of MOLGEN 5.0 and can download further publications related to the MOLGEN series.

MOLGEN is unique in that it serves purposes different from those of other software packages, in particular from those of traditional combinatorial chemistry software. Both input to and output from MOLGEN differ from those of the latter software, a comparison with respect to performance, speed *etc.* is therefore impossible. From the mathematical point of view, MOLGEN's salient feature is its use of sophisticated algebraic methods, in particular of group theory, in order to avoid the combinatorial explosion as far as possible.

Methods

In describing molecular structure we distinguish several levels of detail:

Fuzzy Molecular Formula

Instead of prescribing exact occurrence numbers for each chemical element (or more exactly for each atom type, cf. atom types subsection), for broader coverage numerical intervals are allowed here. On the other hand, for each atom its state may be partially prescribed (valence, charge, hybridization, *etc.*, see atom states subsection) in a fuzzy as in an exact molecular formula.

(Exact) Molecular Formula

For each element symbol with optionally restricted state, its exact occurrence number is given.

Atom State Pattern

For each non–H atom in the molecular formula, its state is fully defined, including the numbers of bonds of various types and the number of hydrogens attached to it.

Molecular Graph

The connections between atoms are described as covalent bonds. In mathematical terms, a molecular structure can be understood as a graph, not only with single bonds, but possibly with double, triple or aromatic bonds.

The generation can be started from any of the levels, with a (set of) formula(s) provided by the user. Then, *via* backtracking, all corresponding molecular graphs are generated.

By choice of the user, the generation can be interrupted on any level, *e.g.* in order to manually select atom state patterns before generating molecular graphs.

Structures

Fuzzy and Exact Molecular Formulas

A molecular formula such as $C_5H_{10}SO_2$ is entered as a string, *e.g.*

C5H10SO2.

The string contains the following information:

- **Atom types**, which are chemical element symbols,

- Optional **atom states**, describing the environment of an atom within the molecular structure (*e.g.* its valence). For example, the formula above could be entered explicitly specifying the valence of S:

  C5H10S[val=2]O2,

- **Atom occurrences**, *i.e.* the number of atoms of given type and state occurring in a structure.

  For a fuzzy molecular formula, each atom occurrence number may be replaced by an interval of numbers, *e.g.* $C_5H_{10}SO_{0\text{-}2}$ could be specified by

  C5H10S[val=2]O0-2.

Note that an element symbol may occur more than once as input for a formula, *i.e.* in different atom states, *e.g.*

C2H4N[val=3]0-1N[val=5]0-1.

Exercise. The interested reader is invited to enter these formulas in MOLGEN–online *via* internet and the address

http://www.molgen.de/?src=documents/molgenonline

For example, enter C5H10SO2, click 'Submit' and after a few seconds you will see that this reduced version of MOLGEN 5.0 produced 4,560 structural formulas. Have a few of them displayed.

After that you may enter C5H10S[val=2]O2 and find out that the same number of isomers is produced, and on inspection you will recognize that the default valence of sulfur used in MOLGEN 5.0 is 2.

Then you may submit C5H10S[val=2]O0-2 or C5H10SO0-2, allowing 0, 1 or 2 oxygen atoms, in which case the online version produces 5,371 molecular graphs.

Atom types (element symbols): An element symbol is one or two letters. Usually an atom type is an element symbol from the Periodic Table of Elements. However, the user may define atom types not yet known to the system. Initially, MOLGEN does not know anything about a user–defined atom type, therefore one has to specify at least its valence as an atom state (see below). As an example, C4H8Qs[val=2]3O will produce structures of formula $C_4H_8Qs_3O$, where the user–defined atom type Qs has valence 2.

Atom states: Atom states describe the environment of an atom within the molecular structure. The following properties may be described:

- The valence of an atom in the structure. This is the total number of covalent bonds that connect the atom to its neighbors (including bonds to H; a double bond is counted twice, *etc.*). Default valences are according to the octet rule.
- The charge of an atom in the structure.
- Specification of an atom as a radical center.
- Isotope specification.
- Hybridization (sp3, sp2, sp), where sp2 is further distinguished for atoms in nonaromatic (sp2_n) and aromatic neighborhood (sp2_a), and sp is further distinguished for atoms bearing a single and a triple (sp_st) *versus* atoms bearing two double bonds (sp_dd).
- Number of H atoms adjacent to an atom.
- Number of single bonds (to non–H atoms) adjacent to an atom.
- Number of double bonds adjacent to an atom.
- Number of triple bonds adjacent to an atom.
- Number of aromatic bonds adjacent to an atom.

Atom State Patterns

A state pattern describes a molecular structure by listing the fully defined state of each atom as described in atom states subsection, including the number of attached hydrogens.

Each atom is listed separately. For coding atom states the following symbols are used:

Hn the number of attached hydrogens,

=n the number of adjacent double bonds,

#n the number of adjacent triple bonds,

˜n the number of adjacent aromatic bonds.

If *n* = 0, the symbol H, =, #, or ˜ is omitted; if *n* = 1, the numeral 1 is omitted. This information together with an atom's valence defines the number of adjacent single bonds. For example,

CH#C#CH=CH=CH2CH

is the state pattern corresponding to 3-ethynylcyclobutene, where

CH# codes a C atom bearing one H and a triple bond,

C# is a C atom bearing a triple bond and a single bond to a non–H atom,

CH= is a C atom bearing one H, one double bond and one single bond to a non–H atom,

CH2 is a C atom bearing two H and two single bonds to non–H atoms,

CH is a C atom bearing one H and three single bonds to non–H atoms.

For a chemist reader, the notions of atom states and atom state patterns may be new. In earlier versions of MOLGEN they were used internally. In MOLGEN 5.0, they are open to manipulation by the user. This is an advantage in certain situations, providing the opportunity to better specify very large runs or to avoid generation of unwanted isomers stemming from unrequested atom state patterns.

Examples:

- mgen -sp CH#C#CH=CH=CH2CH

  generates two structures, 3-ethynylcyclobutene and 3-(2-propynyl)-cyclopropene, while

- mgen -sp CH#C#C=CH=CH2CH2

  leads to three structures, 1-ethynylcyclobutene, 1-(2-propynyl)cyclopropene, and (2-propyn-1-ylene)cyclopropane.

Molecular Graphs

MOLGEN 5.0 is based on a graphical interaction model of a molecule. Graph nodes represent atoms, lines represent covalent bonds. Element symbol and atom state are stored as node labels, the kind of interaction (single, double, triple, aromatic bond) is stored as bond label.

We interpret bond labels as bond multiplicities. An atom's valence is the sum of its bond multiplicities. However, for an aromatic atom, its valence is composed of the number of single bonds and the number of aromatic bonds plus one. For example, in naphthalene, $C_{10}H_8$, each peripheral C atom bears one hydrogen and is involved in two aromatic bonds, while each of the two central atoms has no hydrogen and is involved in three aromatic bonds.

In a graphical representation, there is no explicit order of atoms specified. In order to handle structures without being restricted to a particular atom numbering, a massive use of group theory is necessary. Details can be found in Ref. [5, 6].

Aromaticity. MOLGEN 5.0 has a special bond type 'aromatic' for aromatic bonds. Consequently, cyclically conjugated double bonds forming an aromatic system are not generated. Rather, the corresponding structure is generated with the aromatic ring made of aromatic bonds.

Therefore MOLGEN has a built–in aromaticity detector plus filter that is based on the famous $4n+2$ π-electrons rule (Hückel rule). In the current version cyclically conjugated rings of 6, 10, 14, *etc.* members are considered aromatic. In a future version, additional rings such as pyrrol, furan, thiophen, tropylium, cyclopentadienide *etc.* will be recognized as aromatic.

For example,

```
mgen C[sp2_n]10H8 -ringsize 6-10
```

results in six molecular graphs, none of which corresponds to naphthalene, whereas

```
mgen C[sp2_a]10H8
```

produces four structures, among them naphthalene and azulene. Atom states sp2_n and sp2_a therein denote sp2 atoms in nonaromatic or aromatic systems, respectively (see atoms states subsection).

If desired, aromaticity handling may be deactivated. Then, benzene is generated with single and double bonds instead of aromatic bonds. Thus, 1,2-dimethylbenzene (o-xylene) will be generated twice, having either a single or a double bond connecting the substituted ring atoms.

Restrictions

For each level of generation, several restrictions may be formulated on the set of generated structures.

Restrictions on Exact Molecular Formulas

The following restrictions may be imposed on molecular formulas to be generated from a fuzzy molecular formula. Each number may be restricted by a minimal and maximal allowed value:

- The total number of atoms in a molecular structure (including hydrogens).
- The sum of valences over all atoms. This is double the number of bonds (bonds to H included, double and triple bonds counted as two and three bonds, respectively; aromatic bonds counted as described above).

- The mass of the molecular structure, *i.e.* the sum over atom masses.
- Charge of the molecular structure, *i.e.* the sum over all atom charges.
- Sum over all isotopic mass differences.
- Total number of unpaired electrons in the molecular structure.
- Atom sums, *i.e.* sums of occurrence numbers of atom types/states.

The usage and strength of these restrictions is demonstrated by the following examples.

Examples:

- mgen C2H0-6F0-6Cl0-6Br0-6I0-6 -atoms 8

  generates ethane and all halogenated ethanes;

- mgen C6H0-6Cl0-6 -sum H+Cl=6

  generates all C_6H_6 hydrocarbons and their chlorinated analogs;

- mgen C1-10H4-22 -mass 70-80

  generates all hydrocarbons with a mass between 70 and 80;

- mgen C1-10H4-22 -sum H–2C=2

  generates all alkanes up to the decanes;

- mgen C1-10H4-22 -sum H–2C=0-2

  generates all alkanes plus monounsaturated alkenes plus saturated monocyclic hydrocarbons of up to ten carbon atoms.

- The atom sum restriction can be used to allow alternative atom states for an element. In the following example generation is restricted to structures containing at most two nitrogen atoms of valence 3 or 5:

mgen C2H4N[val=3]0-2N[val=5]0-2 -sum N=0-2.

Restrictions on Atom State Patterns

The following restrictions influence the number and type of generated atom state patterns. Again each number may be restricted by a minimal and maximal allowed value:

- Maximal allowed bond multiplicity (*i.e.* 1, 2, or 3).
- Total number of single bonds (including bonds to hydrogens).
- Total number of double bonds.
- Total number of triple bonds.
- Total number of aromatic bonds.
- Number of bonds between atoms without counting bond multiplicity (including bonds to hydrogens).
- Number of cycles in the molecular structure. This is the number of bonds that have to be broken in order to obtain an acyclic structure, *e.g.* naphthalene has two, not three cycles, cubane has five cycles.
- Number of connected components of the molecular graph. By default connected graphs only are generated.

Restrictions on Molecular Graphs

In order to reduce the number of isomers generated, the following restriction is useful:

-ringsize n[-m] Specify the allowed ring sizes.

Any closed path in the molecular graph is considered a ring. For example, naphthalene contains rings of sizes 6 and 10, cubane has 4-, 6- and 8-membered rings. If a user allows 4-membered rings only, cubane will be missed.

Both power and limitations of the options described hitherto are easily seen in the following example, where we try to restrict the molecular formula $C_6H_5NO_2$ to nitrobenzene.

Examples:

- mgen C6H5NO2

  results in 444,199 structures, nitrobenzene not among them;

- mgen C6H5N[val=5]O2

  gives 1,038,793 structures, among them nitrobenzene;

- mgen C6H5N[val=5,d=2]O2

  renders 122,699 structures;

- mgen C6H5N[val=5,d=2,h=0]O2

  results in 98,687 structures;

- mgen C6H5N[val=5,d=2,h=0]O[d=1]2

  results in 3,893 structures;

- mgen C6H5N[val=5,d=2,h=0]O[d=1]2 -cycles 1

  renders 1,436 structures;

- mgen C6H5N[val=5,d=2,h=0]O[d=1]2 -ringsize 6-9

  gives 452 structures;

- mgen C6H5N[val=5,d=2,h=0]O[d=1]2 -cycles 1 -ringsize 6-9

  results in 140 structures;

- mgen C6H5N[val=5,d=2,h=0]O[d=1]2 -cycles 1 -ringsize 6 produces still 110 structures;

-
  ```
  mgen C[sp2_n]6H5N[val=5,d=2,h=0]O[d=1]2 -cycles 1 -ringsize 6
  ```

  results in 10 structures, nitrobenzene not among them;

-
  ```
  mgen C[sp2_n]0-6C[sp2_a]0-6H5N[val=5,d=2,h=0]O[d=1]2 -cycles 1 -ringsize 6 -sum C=6
  ```

  results in 11 structures;

-
  ```
  mgen C[sp2_a]6H5N[val=5,d=2]O2
  ```

  produces exactly one structure, nitrobenzene.

The example demonstrates the demand for more powerful restrictions, *i.e.* for substructure restrictions.

Structural Restrictions

You can specify substructures as restrictions to MOLGEN.

MOLGEN substructures support ‘Any’ atom type (element symbol A) and extended bond types like ‘single or aromatic’, ‘double or aromatic’, ‘single or double’, or ‘any bond’. For creating and editing substructures, any standard molecule editor supporting MOL files is suitable, for example Accelrys Draw or ACD Chemsketch.

MOLGEN distinguishes ‘open’ and ‘induced’ substructures. In the **induced** case, if free valences on different atoms in a given substructure get connected to each other, this is considered a non–match. Thus, additional zero–length bridges within a substructure, or higher bond multiplicities, will cause a non–match. In the **open** case, however, such variations are recognized as a match. In mathematical terms, an induced substructure is an induced subgraph of the molecular graph, while an open substructure is a subgraph in general.

Consider for example a substructure ‘general_cyclohexane.mol’ consisting of a 6–membered ring of A atoms (‘Any’ type), all bonds are single.

Using general_cyclohexane.mol as open substructure, *e.g.* cyclohexane, cyclohexene, cyclohexa-1,3-diene, cyclohexa-1,4-diene, benzene, benzyne, piperidine, pyridine, bicyclo[2.2.0]hexane substructures, *etc.*, will be considered matches of the substructure.

Using general_cyclohexane.mol as induced substructure, *e.g.* cyclohexene, cyclohexa-1,3-diene, cyclohexa-1,4-diene, benzene, benzyne, pyridine, bicyclo[2.2.0]hexane substructures will be considered as non–matches. Piperidine and of course cyclohexane are recognized as matches.

Given a substructure, you can restrict its occurrence number in the generated molecular graphs to a specific range.

Examples:

- mgen C8H11N -cycles 1-4 -ringsize 5-9

  results in 11,586 compounds, among them being substituted pyridines, dihydro- and tetrahydropyridines, piperidines, benzenes, cyclohexadienes, cyclohexenes, and cyclohexanes;

- mgen C8H11N -cycles 1-4 -ringsize 5-9 -substr open 0 general_cyclohexane.mol

  generates 6,290 compounds, none of which contains any 6–membered ring;

- mgen C8H11N -cycles 1-4 -ringsize 5-9 -substr induced 0 general_cyclohexane.mol

  leads to 10,857 compounds, among them pyridines, dihydro- and tetrahydropyridines, benzenes, cyclohexadienes and cyclohexenes, but no piperidines or cyclohexanes. So the piperidines and cyclohexanes filtered out amount to 729;

- mgen C8H11N -cycles 1-4 -ringsize 5-9 -substr induced 1-4 general_cyclohexane.mol produces exactly 729 substituted piperidines and cyclohexanes, and this set is identical to the set filtered out above.

Having another substructure ‘benzene.mol’ consisting of a 6–membered ring of carbon atoms, all bonds specified as aromatic, we can use it to restrict our generation to structures having at least one benzene substructure.

However, using benzene.mol as induced substructure, dehydrobenzene (benzyne) or a zero–bridged benzene ring will not be considered a match, and consequently structures containing a benzyne but not a benzene will not be generated. Of course, structures containing both a benzene and a benzyne may occur.

Using benzene.mol as open substructure, benzyne or a zero–bridged benzene ring will be considered a match, and consequently structures containing a benzyne but no benzene substructure will be generated.

- mgen C6H5N[val=5]O2 -substr induced 1 benzene.mol

  results in 143 structures, each containing a benzene substructure, and nitrobenzene being among them;

- mgen C6H5N[val=5]O2 -substr open 1 benzene.mol

  results in 312 structures, many of which contain a (presumably undesired) zero–bridged benzene ring;

- mgen C6H5N[val=5,h=0]O2 -substr induced 1 benzene.mol

  renders 7 structures;

- mgen C6H5N[val=5,d=2]O2 -substr induced 1 benzene.mol

  generates nitrobenzene as the only structure;

- mgen C6H5N[val=5]O2 -substr induced 1 nitro.mol

  results in 685 structures, among them nitrobenzene;

- mgen C6H5N[val=5]O2 -substr induced 1 nitro.mol -cycles 1

  gives 197 structures;

- mgen C6H5N[val=5]O2 -substr induced 1 nitro.mol -cycles 1 -ringsize 6

  renders 14 structures;

- mgen C6H5N[val=5]O2 -substr induced 1 nitro.mol -substr induced 1 benzene.mol

  of course delivers nitrobenzene as the only structure.

Recall that for the examples to work appropriately it is important that the bonds in ‘benzene.mol’ are of type ‘aromatic’ and that the nitrogen in ‘nitro.mol’ has valence 5.

Two SDfiles of ‘bad’ open substructures are shipped together with MOLGEN, named badlist.sdf and badlist2.sdf. The former contains 39 highly strained saturated and unsaturated small mono-, bi-, and polycyclic structures that we consider ‘not viable’ (Fig. **1**). The latter is a collection of 14 ‘not viable’ bridged aromatic structures, shown in Fig. **2**. Though such lists are, of course, somewhat arbitrary, they are useful for removing obviously unwanted structures, as demonstrated in the following examples.

Examples:

- mgen C6H6

  generates all 217 mathematically possible benzene isomers;

- mgen C6H6 -badlist badlist.sdf

  results in no more than 66 isomers.

Though 151 isomers are removed thereby, the remaining set still contains those isomers that are known compounds either themselves or as more or less substituted derivatives, such as prismane, Dewar benzene, benzvalene, fulvene, bi-cyclopropenyl, *etc.*

- mgen C6H5N[val=5]O2 -substr open 1 benzene.mol

  generates 312 structures (see above);

- mgen C6H5N[val=5]O2 -substr open 1 benzene.mol -badlist badlist2.sdf

  results in nitrobenzene as the only product.

Obviously, the user may edit these badlists or create one her/himself.

Required and forbidden substructures are used in other structure generators as well, see for example [7].

The Backtracking Algorithm

Restriction Sharpening

Given, for example, a fuzzy molecular formula, a couple of restrictions are induced by simple logic. For example, the number of atoms may not get larger than the sum of maximal occurrence numbers of each element symbol, and it may not get less than the sum of minimal occurrence numbers. Or, if a substructure is prescribed to occur at least once, several minimal bounds are induced, *e.g.* on the number of single bonds, *etc.* in the molecule. Before starting the generation, such induced restrictions are automatically added to the set of restrictions.

Further, the restrictions are highly intercorrelated. For example, the following formula holds for any molecular graph.

$$\text{atoms} + \text{cycles} = \text{bonds} + \text{connected components}$$

Thus, if two of the three quantities number of atoms, of bonds, and of cycles are prescribed *e.g.* for a connected molecular graph, there is no choice for the third. If there are minimal and/or maximal bounds on the numbers, some of the other bounds may be sharpened by applying this formula.

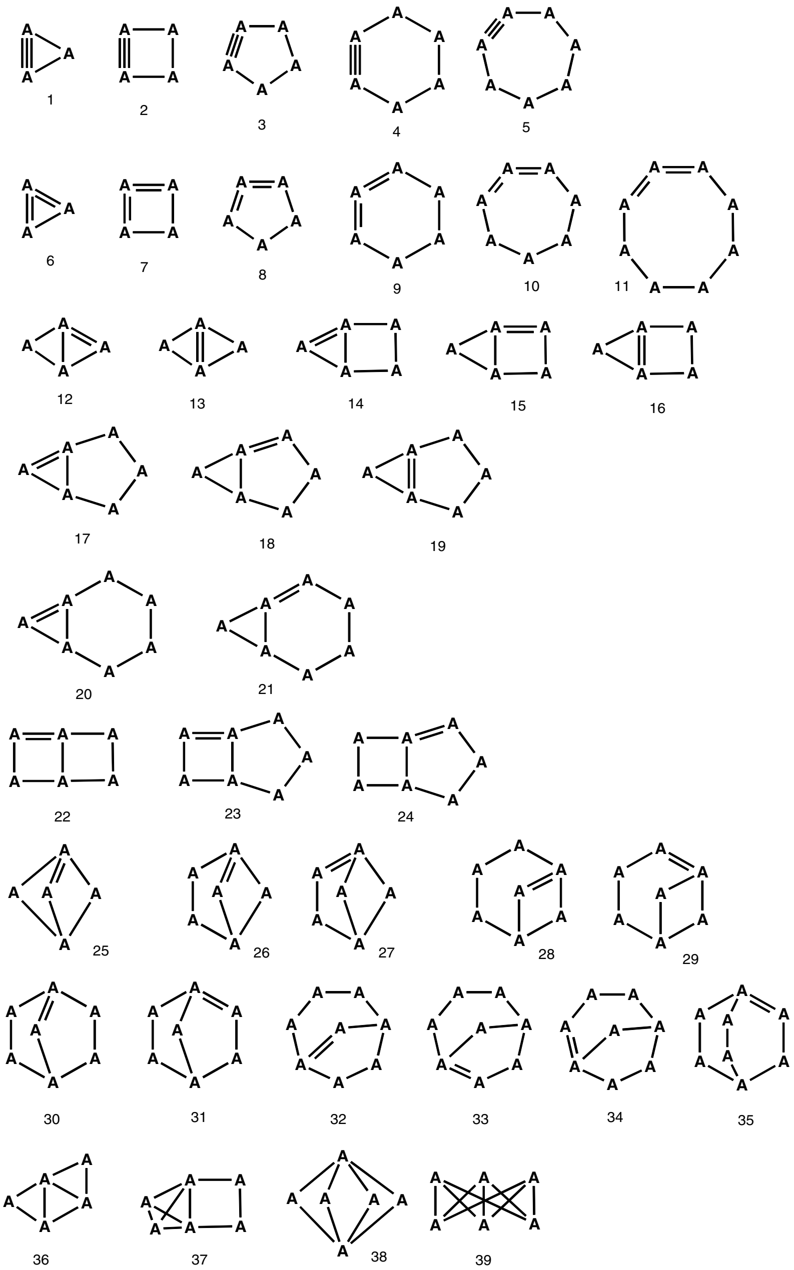


Figure 1: 'Bad' cyclic and unsaturated substructures contained in badlist.sdf.

A couple of graph–theoretic intercorrelations are checked by MOLGEN 5.0 at several stages during the generation in order to keep the restrictions as sharp as possible.

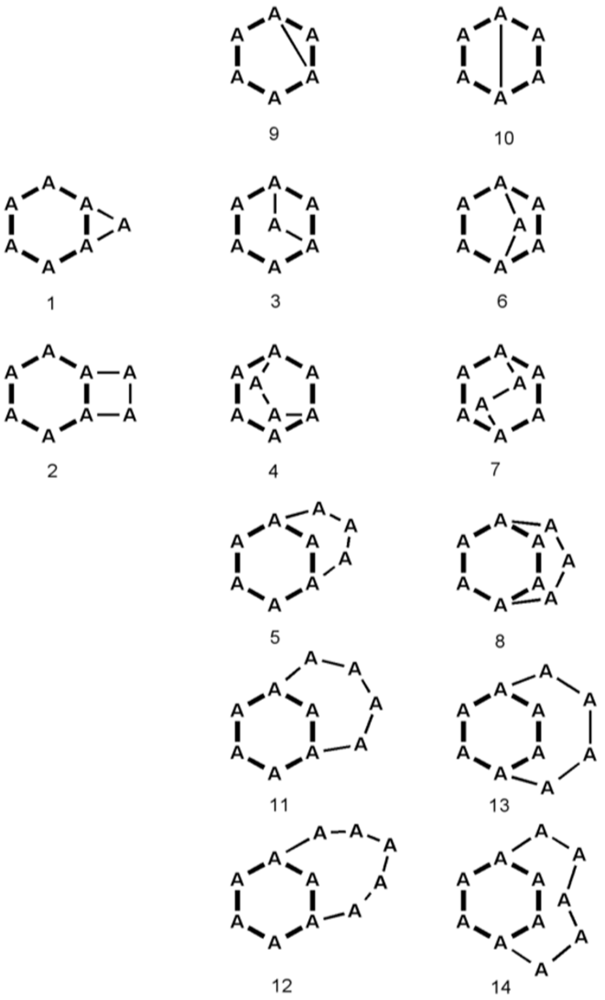


Figure 2: **'**Bad' bridged aromatic substructures contained in badlist2.sdf. Aromatic bonds are symbolized here by thick lines.

During each level of backtracking, a couple of new properties get fixed. For example, when an exact molecular formula was generated starting from a fuzzy molecular formula, the number of atoms gets fixed. Each time after some properties of the molecule get fixed, the graph–theoretic intercorellations are checked again in order to sharpen the remaining restrictions.

Whenever an inconsistency is recognized, for example if a lower bound gets larger than its corresponding upper bound, the current backtrack subtree is pruned.

From Fuzzy Formula to Exact Formulas

For a given fuzzy formula the generator runs through all corresponding exact formulas and the restrictions are tested.

Generating exact formulas implements the following mathematical problem: Generate all partitions of n, which is the maximum allowed nominal molecular mass, into m+1 blocks, where m equals the number of different atom types in the fuzzy formula. Blocks correspond to the atom types, weighted by the corresponding nominal atomic mass. An additional block is for technical purposes to allow generation of formulas not only for a fixed atom weight, but for a range of allowed atom weights.

Example: For the fuzzy formula $C_{1\text{-}10}H_{4\text{-}22}$ with molecular mass restricted to the range 70-80, all number partitions of 80 into three blocks are generated. The first block with weight 12 defines the number of carbon atoms, the second block with weight 1 defines the number of H atoms, and the third block with weight 1 fills the gap between the actual molecular weight and the maximal weight 80.

The first block is restricted to appear 1 to 10 times in the partition, the second block is restricted to appear 4 to 22 times and the third block to appear 0 to 10 times (as the difference between maximal and minimal molecular weight is 10).

The implementation is straightforward, *via* backtracking. A couple of tests are executed before a molecular formula is written to the output or passed to the next level, they follow directly from graph theory and chemistry:

- The sum of valences must be even.

Let a denote the number of atoms including H atoms and b be half of the sum of valences, *i.e.* the sum of all bond multiplicities in any graph corresponding to the formula. Then

- b must be greater than or equal to the maximum valence occurring in the formula,

- $a - b \leq c_{max}$ must be fulfilled. c_{max} is the maximal allowed number of connected components (default is 1).

Further, all user–given restrictions on molecular formulas must be fulfilled:

- All restrictions on the number of atoms.
- All restrictions on the sum of valences.
- All restrictions on charge, isotopes, unpaired electrons.
- All atom sum restrictions.

If all above tests are passed, the exact molecular formula is accepted and in turn used as input for the generation of state patterns.

From Exact Formula to Atom State Patterns

A system of linear equations is established, where the variables are restricted to nonnegative integer values. Usually, problems of this kind are hard to solve. However, MOLGEN contains its own algorithm called ‘solvediophant’ to solve these systems of equations. It is based on the mathematical concept of *lattice basis reduction* [8, 9].

Let a_i, t_i, and d_i be the numbers of aromatic, triple, and double bonds incident with non–H atom i, s_i its number of single bonds to non–H atoms, and h_i the number of H atoms attached to it. Then the number of bonds in the molecule is equal to half of

$$\sum_i (a_i + t_i + d_i + s_i + 2h_i).$$

The following restrictions are formulated as diophantine equations (all sums are over the non–H atoms):

- The numbers of aromatic, triple, double, single bonds fulfill the corresponding restrictions.
- The number of bonds, rings and connected components fulfill their restrictions.
- The sum $\sum_i (a_i + t_i + d_i + s_i + 2h_i)$ is even (as it is twice the number of bonds).

- The sums $\sum_i a_i$, $\sum_i t_i, \sum_i d_i, \sum_i s_i$ are all even (as they are twice the number of aromatic, triple, double or single bonds between non–H atoms).
- The sum $\sum_i h_i$ is equal to the number of hydrogens.
- If there are any aromatic atoms, then there are at least six aromatic atoms and six aromatic bonds. The number of aromatic atoms has to be even.[1]
- The following equation must be fulfilled:

  atoms (incl. H) + cycles = bonds + connected components.
- For each non–H atom, the sum of valences needs to be consistent with its valence v_i: Set $a_i^* = 0$ if and only if $a_i \ = 0$ and put $a_i^* = a_i \ + \ 1$ else. Then

  $$a_i^* + 3t_i + 2d_i + 1s_i + 1h_i = v_i$$
- In particular cases there are further constraints to be fulfilled.
- A system of equations ensures that each state pattern is produced only once by the diophantic solver. We allow only such state patterns in which the list of atom states is sorted in lexicographically decreasing order.

From State Pattern to Molecular Graphs

The construction of all molecular graphs corresponding to a state pattern is done mainly using the same techniques as in MOLGEN 3.5, by orderly generation [10, 11]. More details on how orderly generation is applied to molecular graphs can be found in [12] and were recently discussed in [13].

APPLICATIONS

Molecular Libraries

An interesting problem where we can sometimes take advantage of a fuzzy molecular formula is the generation of molecular libraries. The use of MOLGEN 5.0

[1] Some details on the restrictions concerning aromaticity are omitted here.

makes life easy when we want, for example, to get information on the total set of structural formulas of molecular mass 100, atoms in {C,H,N,O} and containing at least one carbon atom. Enter the fuzzy formula together with the mass constraint

```
mgen C1-8H0-16N0-6O0-4 -mass 100
```

to quickly obtain 33,537 structural formulas. In Table **1** you find numbers of structures that correspond to the various molecular formulas, for several molecular masses ≤ 100.

Table 1: Number of molecular and structural formulas for several molecular masses

mass	MF	MG	MGNAD	BS	MS
20	0	0	0	0	0
30	2	2	2	2	2
40	1	5	5	5	1
50	1	7	7	1	1
60	6	47	47	25	12
70	6	380	380	84	31
80	7	1,645	1,644	100	23
90	11	5,849	5,818	107	28
100	16	33,627	33,537	710	154

Column MF contains the number of molecular formulas corresponding to the mass and the fuzzy formula. MG means the numbers of corresponding molecular graphs, the structural formulas. The filter for aromatic duplicates was turned off when these entries were calculated, so that, for example, the total number of structures of mass 100 turned out to be 33,627. In the online version this filter is on, resulting in 33,537 structural formulas. Therefore we give in column MGNAD the number of structural formulas without aromatic duplicates. Column BS contains the number of structures that are contained in the Beilstein database, while column MS refers to the number of compounds in the NIST mass spectral library. The table is part of tables published in [14], and so these numbers found in the databases are snapshots, they may have changed in the meantime. Nevertheless they are of interest in order to show the enormous difference between the mathematically possible numbers of compounds and the numbers of existing compounds, and the number of existing compounds whose mass spectra were recorded and made publicly available.

Exercise. Refine this table by manually evaluating the molecular formulas corresponding to mass 100, and obtain the isomer numbers online. Look up these molecular formulas in a database such as SciFinder or Reaxys to find out how many corresponding compounds are contained therein. Comparing the numbers keep in mind that database compounds may include stereoisomers, isotopomers, radical ions and various other compound categories that are not included in MOLGEN counts.

Generate All Chlorinated Biphenyls

Often a search space cannot be defined by a single molecular formula, but by a range of several related molecular formulas (a fuzzy molecular formula). A typical example is the generation of congeners. In MOLGEN 5.0 the generation of all chlorinated biphenyls is solved as follows:

```
mgen C12H10 -bonds3 0 -bonds2 0 -bonds1 11 -cycles 2 -ringsize 6
```

produces a single molecule, biphenyl, within about a second on a standard PC.

```
mgen C12H0-10Cl0-10 -sum H+Cl=10 -bonds3 0 -bonds2 0
-bonds1 11 -cycles 2 -ringsize 6
```

results in 210 molecules within 3 sec, *i.e.* the non–chlorinated parent biphenyl and the fully chlorinated decachlorobiphenyl, 3 mono- and 3 nonachlorinated, 12 di- and 12 octachlorinated, 24 tri- and 24 heptachlorinated, 42 tetra- and 42 hexachlorinated, and 46 pentachlorinated biphenyls.

In this example, of course, alternatively eleven runs on an exact molecular formula each could be performed, *e.g.* in MOLGEN 3.5. In the next example, however, such a semi–manual procedure would be a tedious exercise, to say the least.

Halogenated Alkanes

Generate all halogenated (as well as nonhalogenated) alkanes C_1-C_4, where halogenated means bearing at least one F, Cl, Br, or I substituent.

```
mgen C1-4H0-10F0-10Cl0-10Br0-10I0-10 -sum H+F+Cl+Br+I–2C=2
```

generates 187,075 compounds, *i.e.* the alkanes methane, ethane, propane, butane, isobutane, and all their halogen derivatives, corresponding to altogether 1,776 molecular formulas. This takes 35 sec on a standard PC.

Molecular Structure Elucidation

An important real case use of MOLGEN is molecular structure elucidation based on mass spectra. Molecular structure generation is crucial whenever the unknown chemical compound considered is not contained in the available databases. This kind of problem is carefully discussed in all detail in a PhD thesis [15], see also [16, 17]. The role of MOLGEN–MS is described and additional software that is useful in this context is mentioned. In particular, Section 6 contains examples of tentative identification of contaminants in groundwater of Bitterfeld, Germany. Mass spectra of 150 contaminants were obtained, of which 42 could be tentatively identified using the NIST database search alone. 32 of these compounds identified using NIST were confirmed using structure generation techniques. In addition, 20 further peaks were tentatively identified using structure generation techniques alone, resulting in a total of 62 tentative identifications. In another case, an unknown spectrum had the molecular formula $C_{13}H_{10}ClNO$ that has more than 10^9 connectivity isomers, but substructures derived from the spectrum and generation using MOLGEN–MS reduced this number to just 36 candidates. Literature search on diclofenac and additional confirmation analysis further reduced this set to a known diclofenac phototransformation product that was also identified as the one responsible for the enhanced toxicity of the transformed diclofenac towards the green alga *S. vacuolatus*.

For molecular structure elucidation based mainly on NMR spectra see [18] and later papers by these authors.

ACKNOWLEDGEMENTS

Declared none.

CONFLICT OF INTEREST

All authors together are the MOLGEN team which distributes MOLGEN software at a nominal fee.

REFERENCES

[1] Kerber, A.; Laue, R.; Grüner, T.; Meringer, M. MOLGEN 4.0. *MATCH Commun. Math. Comput. Chem.*, **1998**, 37, 205–208.

[2] Laue, R.; Grüner, T.; Meringer, M.; Kerber, A. Constrained generation of molecular graphs. In: *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol 69, American Mathematical Society, **2005**, 319–332.

[3] Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. MOLGEN–QSPR, a software package for the search of quantitative structure-property relationships. *MATCH Commun. Math. Comput. Chem.*, **2004**, 51, 187–204.

[4] Kerber, A.; Laue, R.; Meringer, M.; Varmuza, K. MOLGEN–MS: Evaluation of low resolution electron impact mass spectra with MS classification and exhaustive structure generation. *Adv. Mass Spectrom.*, **2001**, 15, 939–940.

[5] Braun, J.; Gugisch, R.; Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. MOLGEN–CID, a canonizer for molecules and graphs accessible through the Internet. *J. Chem. Inf. Comput. Sci.*, **2004**, 44, 542–548.

[6] Kerber, A.; Laue, R.; Meringer, M.; Rücker, C.; Schymanski, E. *Mathematical Chemistry and Chemoinformatics - Structure Generation, Elucidation and Quantitative Structure-Property Relationships.* DeGruyter Publishers, Berlin, **2013**.

[7] Molodtsov, S. G. The generation of molecular graphs with obligatory, forbidden and desirable fragments. *MATCH Commun. Math. Comput. Chem.*, **1998**, 37, 157–162.

[8] Wassermann, A. Finding simple t-designs with enumeration techniques. *J. Comb. Designs*, **1998**, 6, 79–90.

[9] Wassermann. A. Attacking the market split problem with lattice point enumeration. *J. Comb. Optimization*, **2002**, 6, 5–16.

[10] Faradzhev, I. A. Generation of nonisomorphic graphs with a given degree sequence. *Algorithmic Studies in Combinatorics*, NAUKA, Moscow, **1978**, 11–19 (in Russian).

[11] Read, R. C. Everyone a winner. *Ann. Discr. Math.*, **1978**, 2, 107–120.

[12] Grund. R. Konstruktion molekularer Graphen mit gegebenen Hybridisierungen und überlappungsfreien Fragmenten. *Bayreuther Mathematische Schriften*, **1995**, 49, 1–113.

[13] Meringer, M. Structure enumeration and sampling. In: *Handbook of Chemoinformatics Algorithms*. Eds. J.-L. Faulon and A. Bender. Chapman & Hall/CRC, Mathematical & Computational Biology, **2010**, Chapter 8, 233–267.

[14] Kerber, A.; Laue, R.; Meringer, M.; Rücker, C. Molecules *in silico*: Potential *versus* known organic compounds. *MATCH Commun. Math. Comput. Chem.*, **2005**, 54, 301–312.

[15] Schymanski, E. L. *Integrated analytical and computer tools for toxicant identification in effect–directed analysis*. PhD thesis 07/2011, Helmholtz Centre for Environmental Reseach–UFZ.

[16] Schulze, T.; Weiss, S.; Schymanski, E.; von der Ohe, P. C.; Schmitt-Jansen, M.; Altenburger, R.; Streck, G.; Brack, W. Identification of a phytotoxic phototransformation product of diclofenac using effect-directed analysis. *Environmental Pollution*, **2010**, 158, 1461–1466.

[17] Schymanski, E. L.; Meinert, C.; Meringer, M.; Brack, W. The use of MS classifiers and structure generation to assist in the identification of unknowns in effect–directed analysis. *Analytica Chimica Acta*, **2008**, 615 (2), 136–147.

[18] Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R. A new approach to computer-aided molecular structure elucidation: the expert system Structure Elucidator. *Lab. Autom. Inf. Manag.*, **1999**, 34, 15–30.

CHAPTER 7

On Comparability Graphs: Theory and Applications

Matthias Dehmer* **and Lavanya Sivakumar**

Institute of Bioinformatics and Translational Research, UMIT, A-6060, Hall in Tyrol, Austria

Abstract: In this paper, we review classical and recent developments on comparability graphs. Also, we demonstrate that comparability graphs are useful to analyze molecular graphs by presenting classical and new results. In fact, it turns out that the underlying model is quite general and, hence, could be used to analyze any kind of network data.

Keywords: Comparability graphs, chemical graph, topological indices, molecular descriptor, structural complexity, graph entropy, Shannon information content, quantitative network analysis, information-theoretic measures, structural similarity, similarity measures, graph edit distance, similarity matrix, correlation matrix.

INTRODUCTION

During the last decades, properties of relational structures have been extensively investigated [1-5]. As a particular result, a theory for structurally investigating relational structures representing graphs has been established [3, 4, 6]. Highlights from this theory are for instance, graph colorings, graph minors and random graphs. After establishing the theoretical fundament of graph theory, it turned out that graphs are quite generic and, therefore, useful to explore complex systems in various disciplines meaningfully. In fact, modern application areas such as network biology [7], structural chemistry [8] and mathematical psychology [9] exist in which graphs and methods for their structural analysis have been proven useful.

Graph analysis as a tool to analyze biological or chemical systems turned out to be of particular interest because intriguing facts of life in molecular and cell

***Corresponding author Matthias Dehmer:** Institute for Bioinformatics and Translational Research, UMIT - The Health and Life Sciences University, Hall in Tirol, Austria; Tel: +43 50 8648 3851; Fax: +43 50 8648-673836; E-mail: matthias.dehmer@umit.at

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50007-2

biology could be explored [7, 10-12]. Also, the behavior of complex systems ranging from cell to social networks emerges from the unified activity of components interacting. This implies that a complex system can be modeled as a network, where the components are the vertices and the interactions between these components are the edges.

To tackle the problem of analyzing such complex graph-based systems, there exist two major categories: descriptive and quantitative methods. It is worth mentioning that, particularly, quantitative techniques for graphs analysis such as descriptors and comparative methods turned out to be crucial [5, 13-16]. For instance, the use of structural graph descriptors [15-19] has had a tremendous impact in structural chemistry and related areas such as drug research [20] and medicinal chemistry [21].

In this paper, we do not only focus on developing techniques to analyze graphs quantitatively. In particular, we study graph representations called *comparability graphs* and comparative techniques for their analysis. Note that comparability graphs have been proven useful when interpreting structural data sets in chemistry, see [22-25]. Comparability graphs have been defined by using partial orders. If one defines such partial orders based on the underlying data set, deeper insights when studying properties of molecular structures such as their branching could already be obtained [22-25].

The article is organized as follows. In the first part, we present definitions, properties and mathematical characterizations of comparability graphs. The second part deals with applying comparability graphs as a tool for characterizing molecular and biological structures. We also discuss some new results in Section Numerical Results and Analysis. The paper ends with a summary and conclusion.

COMPARABILITY GRAPHS

Let $G = (V; E)$ be a graph with n vertices and let $P = (V; <)$ be a poset with a partial order $(<)$ defined on the vertex set V.

Definition 1. A graph G is said to be *simple* if it does not contain multiple edges or loops.

Throughout this chapter, we consider only simple graphs unless stated otherwise.

Definition 2. A graph G is a *comparability graph* if there is a partial order $<$ on V such that $(x, y) \epsilon E$ if and only if $x < y$ or $y < x$.

Various theoretical characterizations of comparability graphs have been established in the literature [26-29]. Before reproducing some of these results, we state some definitions.

Definition 3. A *walk* in a graph is a sequence of vertices $(v_0, v_1, \ldots, v_k)$, such that any two consecutive vertices in the sequence are adjacent. In addition, if $v_0 = v_k$ then the walk is said to be *closed*. A closed walk is *odd or even* if the number of vertices in its sequence is odd or even, respectively. A closed walk is called a *cycle*, if all its vertices are distinct except that $v_0 = v_k$.

Definition 4. A *triangular chord* of a closed walk $\{v_0, v_1, \ldots, v_{k-1}, v_k = v_0\}$ is the edge (not belonging to the walk) connecting any two alternate vertices in the sequence of a walk. In other words, a walk is said to possess a triangular chord if any one of the following edges (v_j, v_{j+2}), $0 \leq j \leq k-2$ or (v_{k-1}, v_1) exist in the given graph.

Definition 5. A *chord* of a cycle is the edge connecting any two nonconsecutive vertices of the cycle.

Definition 6. A simple graph G is said to be a *chordal graph*, if every cycle (of length $\geq$ 4) in G contains a chord. A chordal graph is also known as a *triangulated graph.*

Definition 7. A graph G is *perfect,* if $\chi(H) = \omega(H)$, for every induced subgraph $H \subset G$, where χ represents the vertex chromatic number (minimum number of colors required to label the vertices such that no two adjacent vertices receive same color) of H and ω is size of the largest clique (set of pairwise adjacent vertices in a graph) in H.

Definition 8. An *interval graph* is a graph having an interval representation. That is, a family of intervals is assigned to the vertices so that vertices are adjacent if and only if the corresponding intervals intersect.

An early result has been proven by Gilmore *et al*. [27].

Theorem 1 [27]. *A graph G is a comparability graph if and only if each odd closed walk has at least one triangular chord.*

The following characterizations reveal the relation between comparability graphs, interval graphs and perfect graphs.

Theorem 2 [30]. *Any comparability graph, G, is a trivially perfect graph. That is, for all induced subgraphs H of G, the size of the maximum independent set of H is equal to the number of maximal cliques (complete subgraph) in H.*

It has been proven by Berge (1960) [28] that every comparability graph is a perfect graph. Alternatively, the same result can be arrived through one of the equivalent characterization of trivially perfect graphs which states that, trivially perfect graphs form a subclass of interval graphs and, hence, perfect graphs. Thus from the above theorem, it follows that the comparability graphs are perfect graphs.

While characterizing the complements of interval graph, Ghouila-Houri [26] proved the following result.

Theorem 3 [26]. *The complement of an interval graph is a comparability graph with the partial order being the interval order.*

The above theorem plays an important role when characterizing interval graphs which is immediate from the following result.

Theorem 4 [28]. The following conditions are equivalent for a graph G:

(a) G has an interval representation (that is, G is an interval graph).

(b) G is a chordal graph and the complement of G, $\bar{G}$, is a comparability graph.

Directed Comparability Graphs

Directed comparability graphs received considerable attention for solving problems such as registry allocation in parallel processing, scheduling problems and the analysis of flow networks and molecular networks [24, 31, 32]. By definition, a simple graph that admits a transitive *orientation* on its edges is a comparability graph. Here, an orientation of a graph G is a digraph D obtained from G by choosing an orientation (either $x \rightarrow y$ or $y \rightarrow x$ and not both) for each edge $(x, y) \epsilon E(G)$. A digraph is *transitive*, if $x \rightarrow y$ and $y \rightarrow z$ implies $x \rightarrow z$. Each transitive orientation on G defines a poset P on the vertex set. Further, the vertices of a path in a transitive digraph induce a tournament.

Another interesting property of directed comparability graphs is that they are hereditary. That means, every induced sub-digraph of a directed comparability graph is transitive. Using this hereditary property and Zorn's lemma, the following result has been proven by Wolk [29].

Theorem 5 [29]. *If every finite subgraph of an undirected graph G admits a transitive orientation, then G also possesses a transitive orientation.*

In this chapter, we concentrate on analyzing directed comparability graphs.

APPLICATIONS

An interesting branch of science is molecular engineering [20, 33, 34] where methods for designing new compounds, mixtures, *etc.*, have been combined with the evaluation of their properties by means of quantitative structure property (QSPR) and quantitative structure activity (QSAR) relationships [35-37]. In particular, chemical graph theory has served as an efficient tool to quantify a chemical structure by converting it into a number. Special tools to tackle this problem are well known as *molecular descriptors and topological indices*, see [8, 18, 19]. Numerous molecular descriptors have been proposed in the literature involving both combinatorial [38, 39] and information-theoretic techniques [19, 40, 41]. Apart from quantifying a structure, establishing relations between molecular networks is also crucial. For example, this could be achieved by using existing similarity measures for graphs [14, 42] or by ordering the networks in terms of the complexity involved. Interestingly, comparability graphs have found useful to tackle the just mentioned problems [22-25, 43-46].

Note that the concept of partial ordering has been firstly applied to arrange/order molecular structures in terms of their structural complexity [22-25]. For instance, Bonchev *et al*. [22-25] ordered alkane isomers as well as condensed benzenoid hydrocarbon isomers by using comparability graphs. The set of rules for branching and cyclicity from [44, 45] have been used to determine the underlying partial order. That is, by applying such rules to every pair of graph, a collection of directed, ordered graphs have been derived. This collection of directed graphs using certain criteria [44, 45] has been finally used to obtain the required comparability graph. Eventually, any two structures are said to be comparable if they lie on a directed path in this comparability graph, otherwise they are non-comparable. Using such an ordering of the comparable graphs, optimal correlation samples have been chosen to determine the structural complexity, based on branching and cyclicity.

Note that the comparability graph obtained by the above mentioned procedure also resembles the reaction graph introduced by Balaban [43, 47] to enumerate all intra-molecular rearrangements among a group of isomeric molecules. Randić [48, 49] proposed another structural representation called *grid graph* for ordering the molecular structures based on the values of selected structural descriptors arranged in a grid-like structure to study the regularity of molecules. However, the above mentioned techniques have only been applied to sets containing relatively small-sized isomers and, hence, the feasibility of these methods when using large arbitrary networks has not been explored so far. As another attempt, the concept of structural similarity has also been used to order molecular structures. For related work, see [46, 50-55]. In addition, the posets of directed graphs have been well studied in chemistry and have been established as an application for the study of correspondence in property values *via* the reaction networks. Here, each partial ordering of structures are represented as a Hasse diagram. As it is out of scope for our current discussion, we refer to more related work, see [56-58].

Structural Similarity of Comparability Graphs

In the following, we state a method to combine the concept of comparability graphs (derived from topological descriptors applied to a collection of graphs) and graph similarity. The purpose is to gain additional structural insights when comparing the given comparability graphs using existing graph similarity (or distance) measures. For this, we consider a set of topological descriptors $\mathfrak{D} = \{D_1, \dots, D_N\}$ being evaluated on a set of graphs $\mathcal{G} = \{G_1, \dots, G_n\}$. As a result, a collection of comparability graphs $\mathcal{CG}_{\mathcal{D}}$ is derived using the values of the descriptors as follows.

Definition 9. Let $\mathcal{CG}_{\mathcal{D}} = \{CG_D: D \in \mathcal{D}\}$. Each $CG_D = (V, E), D \in \mathcal{D}$ is a simple directed graph with a common vertex set $V = \{1, \ldots, n\}$, each vertex corresponding to a graph of G and the edge set $E \subseteq V \times V$, is such that an edge from i to j exists if and only if $D(G_i) \geq D(G_j)$, for $i \neq j$.

Note that the edge set *E* describes the relation between the graphs in terms of the descriptor.

This kind of comparability graphs satisfies the following properties.

Proposition 6. Every graph $CG_D \in \mathcal{CG}_{\mathcal{D}}$ is a directed, labeled graph with $n = |\mathcal{G}|$ vertices. Every CG_D possesses a naturally induced hierarchy and, thus, represents a hierarchical structure manifesting the following properties on the "levels" of vertices:

1. The root(s) of CG_D (at level 1) contains the graph(s) possessing the maximum value of the descriptor among other graphs.

2. Any vertex on level i has an outbound edge to all vertices on level j, $\forall j \geq i$

3. Any vertex on level i has an inbound edge from all vertices on level j, $\forall j \leq i$

4. If there exists more than one vertex on a particular level, all the vertices are connected as a bidirectional clique. That is, all the graphs having same value of the descriptor falls in one level of CG_D thereby forming a tournament.

Corollary 7. CG_D is acyclic if for every pair of graphs $G_i, G_j \epsilon \mathcal{G}, D(G_i) \neq D(G_j)\, holds\, for\; 1 \leq i \neq j \leq n$.

By the above proposition, we have characterized our comparability graphs. The next step is to find appropriate graph similarity measures for determining the structural similarity. Also, we determine the structural similarity between every pair of graphs in $\mathcal{CG}_{\mathcal{D}}$ to better understand the characteristics of the descriptors and their interrelatedness when applied to a given set of underlying graphs.

To tackle this problem concretely, we choose the well-known graph edit distance (GED) [13, 59] because it is easily interpretable. In general terms, GED is computed in two steps: Given two graphs, we firstly obtain a sequence of graph edit operations (such as adding/deleting vertices and edges) required to transform one graph into another. Second, define a cost function for each operation so that the cost for the edit operation sequence is the sum of the costs for all the operations in the sequence. Among all possible edit operation sequences, the least cost sequence is defined to be the GED between the two graphs. Details for implementing GED can be found in [60-62]. In our case, we need a special definition of GED since the two graphs have the same vertex set.

Definition 10. The *Graph Edit Distance*, *GED* between any two labeled graphs $G = (V, E_1)$ and $H = (V, E_2)$ is the symmetric difference between their edge sets E1 and E2, given by

$$GED(G, H) = |(E_1 \cup E_2) \setminus (E_1 \cap E_2)| \tag{1}$$

Numerical Results and Analysis

We start this section by stating two definitions.

Definition 11. Let *GedM*, also known as similarity matrix, denote an $N \times N$ matrix where

$$[GedM]_{ij} = GED\left(CG_{D_i}; CG_{D_j}\right), \tag{2}$$

where $GED(\cdot,\cdot)$ is the graph edit distance between graphs.

Definition 12. Let *CorM*, known as *correlation matrix*, denote an $N \times N$ matrix where

$$[CorM]_{ij} = \rho(D_i, D_j), \tag{3}$$

where $\rho(\cdot,\cdot)$ denotes the Pearson's correlation coefficient.

In this context, we firstly compute the *GED* between all pairs of comparability graphs in $\mathcal{CG}_{\mathcal{D}}$ and apply agglomerative clustering [63] to the resulting similarity matrix. This relates to finding the groups of comparability graphs in which the graphs are similar, with respect to *GED*. Secondly, we compare the resulting dendograms with those obtained by determining the correlations between the descriptors, instead of calculating *GED*.

For this purpose, we use special graph classes and concrete structural descriptors. In particular, we choose 32 graph measures where 20 of those are recently developed entropy-based measures [64] and the remaining 12 are classical measures (both entropy based and nonentropy measures). The reason why we particularly choose graph entropies is to explore differences between information-theoretic and non-information-theoretic indices, *e.g.*, how the measures in question capture structural information. As graph entropies, we choose [64]

$$\mathcal{D}_1 = \{H_{M,1}(G): M\ is\ a\ molecular\ matrix\}; \tag{4}$$

and

$$\mathcal{D}_2 = \{H_{M,2}(G): M\ is\ a\ molecular\ matrix\}; \tag{5}$$

where $H_{M,s}(G)$, for $s = 1; 2$, is the entropy measure defined by the eigenvalues of a molecular matrix M. Let $\{\lambda_1, \lambda_2, \dots, \lambda_k\}$ be the nonzero eigenvalues of M. Then,

$$H_{M,s}(G) = -\sum_{i=1}^{k} \frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^{k} |\lambda_i|^{\frac{1}{s}}} \log_2 \left(\frac{|\lambda_i|^{\frac{1}{s}}}{\sum_{j=1}^{k} |\lambda_i|^{\frac{1}{s}}} \right)$$

Note that numerous graph-theoretical matrices have been defined by using the structural properties of a molecular graph [65]. To perform our analysis, we only consider 10 different types of molecular matrices as stated in [64]. These matrices are defined using the adjacency between vertices, the degree of a vertex and/or the distance between two vertices of a graph.

As stated before, to compare the results with other indices, we also choose the following classical measures from the literature namely, the Wiener Index, W [16], Randić connectivity index, RC [15], Harary index, H [66], Compactness index, C [67], Mean Distance Deviation, *MDD* [68], Hyper-distance-path index, *HDP* [19], Zagreb index Z1 [8], Topological information content, *TIC* [69], Bonchev-Trinajstic index, *BT* [45], Bertz complexity index, B [70], Balaban index, *J* [36], Balaban-like information index, *U* [37],

$$\mathcal{D}_3 = \{W, RC, H, C, MDD, HDP, Z1, T\,IC, BT, B, J, U\} \tag{7}$$

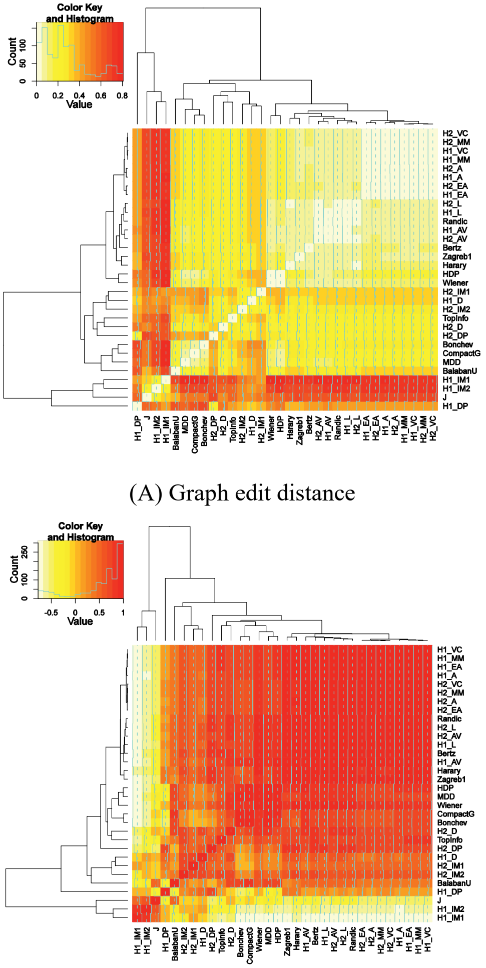
Now, we choose the datasets $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$ and $\mathcal{G}_4$, referred to as MS2265, C12*Ring*1, C12*Ring*2 and C15*Trees*, respectively [64, 71, 72]. These datasets contain both real and synthetic chemical structures. In particular, the set $\mathcal{G}_1$ contains real graphs, while each of the set $\mathcal{G}_2, \mathcal{G}_3$ and $\mathcal{G}_4$ contains only synthetic isomers. The synthetic graphs have been generated by using the software Molgen [73]. Each of the datasets only contain the skeletons of the underlying chemical structures (all bond and atom types are considered equal), isomorphic structures have been filtered out [74]. Further, we choose a random sample of 100 graphs from each of these databases and each of the 32 descriptors (from $\mathcal{D}_1 \cup \mathcal{D}_2 \cup \mathcal{D}_3$) are used to calculate the values using these graph collections. Thus, for each graph class, $\mathcal{C}\mathcal{G}_\mathcal{D}$ contains 32 comparability graphs each having 100 vertices each.

In Fig. (**1**), we present the dendograms using the similarity matrix *GedM* and the correlation matrix *CorM* for the graphs from $\mathcal{G}_1$.

From Fig. (**1a**), it is immediate that among all the entropy measures, the $GED(CG_{H_{M,1}}, CG_{H_{M,2}})$ for a given molecular matrix M, is very low since the pair of graphs enter into a cluster at the same level and, hence, they are highly similar. In addition, we infer that among the entropy-based measures from $\mathcal{D}_3$, the Balaban index J is very similar to the eigenvalue-based entropy measures $H_{M,s}$ for distance-based matrices DP(G), IM1(G) and IM2(G). Also, J is highly dissimilar to the remaining descriptors.

From the above observations, we infer that highly similar comparability graphs grouped within a cluster show that the underlying graphs from a collection $\mathcal{G}$ have a similar ordering. The comparability graphs from different clusters produce a considerably different ordering of the graphs from $\mathcal{G}$. In particular, the comparability graphs belonging to a cluster from the first level and a cluster from the last level possess an almost reverse ordering of the underlying graphs from $\mathcal{G}$.

Hence, we see that clustering technique of comparability graphs provides information about ordering the graphs.



(A) Graph edit distance

(B) Correlation matrix

Figure 1: Cluster analysis of the matrices *GedM* and *CorM* for the graphs from $\mathcal{G}_1$.

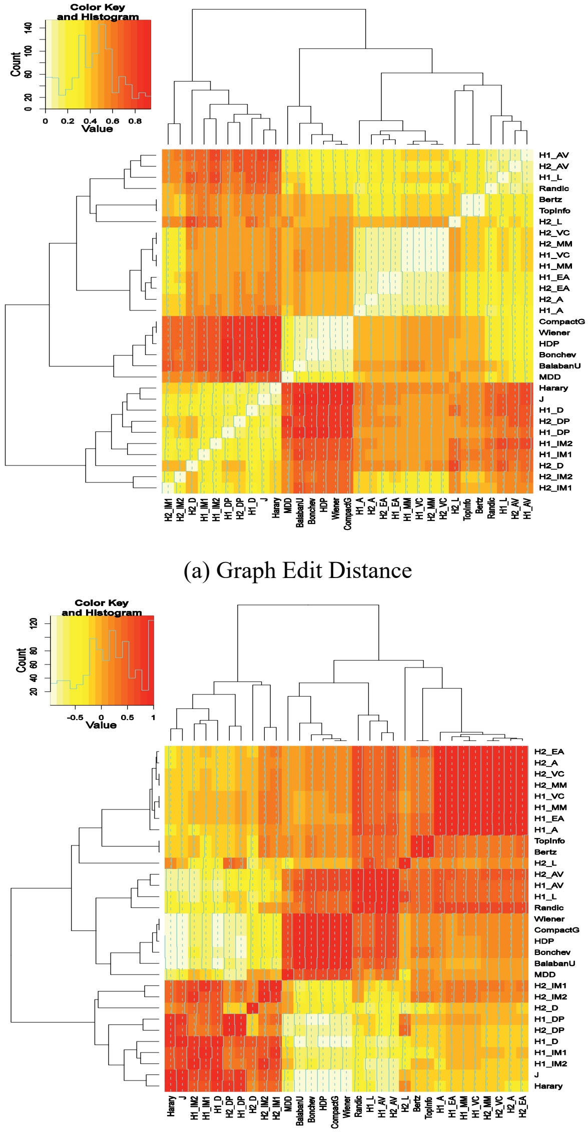
Next, we perform a comparative analysis between the similarity matrix (Fig. (**1a**)) and the correlation matrix (Fig. (**1b**)). First of all, we obtain that the comparability graphs of highly correlated descriptors are very similar. That is, if two descriptors D_1 and D_2 possess high correlation coefficient, then the graph edit distance between the corresponding comparability graphs are very low (*i.e.*, very few edit operations are required to transform a given graph into the another one).

Thus, this implies that the comparability graphs are very similar. For example, the dark red region in the Fig. (**1a**) represents the comparability graphs of the descriptors that require more number of edit operations (at least 80% of the edges are modified) to transform one into another, while the same descriptors in Fig. (**1b**) show a very low correlation (or high negative correlation, less than -0.5) with the region colored by white and shades of light yellow. That is, the descriptors with negative correlation imply that the direction of most of the edges in the corresponding comparability graphs need to be changed and, hence, the graphs are highly dissimilar.

Secondly, the entropy measure based on the matrices $MM(G)$ and $VC(G)$ have identical comparability graphs. This leads us to rediscover, in support of [64], that the descriptors possess the maximum correlation value of $+1$. Additionally, the matrices $A(G)$ and $EA(G)$ also have a very similar comparability graph. This fact is immediate from the dendograms where all these eight measures get assigned to the same cluster in the last but one level showing that these matrices are closely related to each other.

Further when analyzing the performance of the descriptors, it is immediate that the Balaban J index, the entropy measures $H_{DP,1}, H_{IM_1,1}, H_{IM_2,1}$ outperform all the other measures as the distance between them is very high and they enter into a cluster at a very later stage (level 1 an level 3 in the dendogram). In general, the descriptors that enter a cluster at a particular level (with respect to the matrices *CorM* and *GedM*) are identical for the descriptors from $\mathcal{D}_1$, while it is very similar for the descriptors from $\mathcal{D}_1 \cup \mathcal{D}_2$ and $\mathcal{D}_3$. At the outset, we note that the dendograms of both the matrix clusters are not identical. However, the elements (descriptors) that enter into a cluster in the lower levels show an identical pattern.

In Figs. (**2**), (**3**) and (**4**), we present the density plot along with the dendograms obtained by applying clustering techniques to the similarity matrix and the correlation matrix for the graphs in $\mathcal{G}_2, \mathcal{G}_3$and $\mathcal{G}_4$ respectively. As before, we deduce similar conclusions when the above analysis is performed on these collection of isomer structures. However, it is worth to note that, though the obtained dendograms are not identical, the elements that enter into a cluster in the first few levels show an identical pattern. In addition, we note from Figs. (**2a**) and (**4a**) that the distribution of dendograms and the density plot are identical for $\mathcal{G}_2$ and $\mathcal{G}_4$, and is much similar to $\mathcal{G}_3$. This can be understood and interpreted as a kind of measure of stability of the measures on various graph collections.

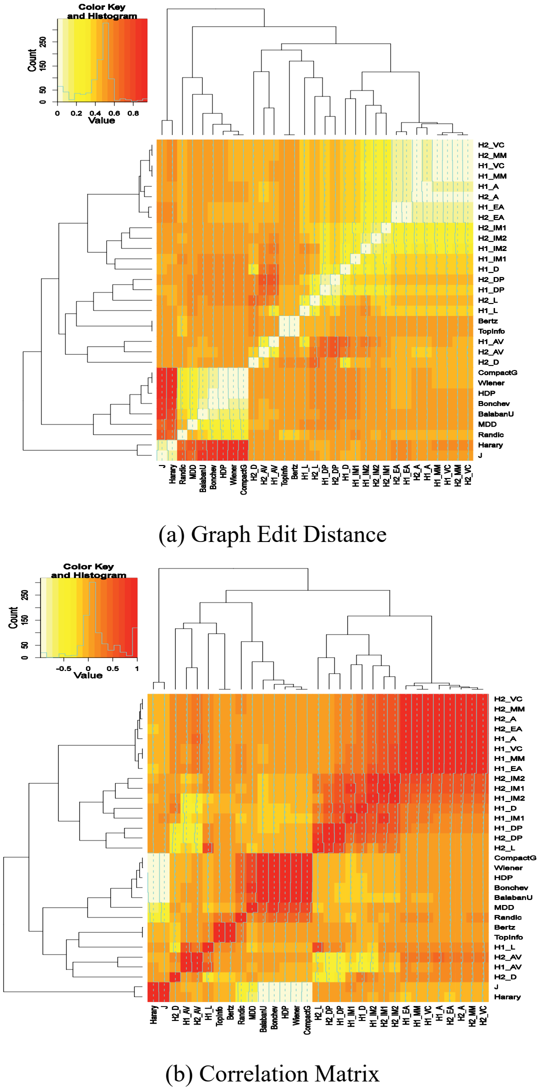


(a) Graph Edit Distance

(b) Correlation Matrix

Figure 2: Cluster analysis of the matrices *GedM* and *CorM* for the graphs from $\mathcal{G}_2$.

Thus, we have presented the similarities and dissimilarities of the comparability graphs in a more descriptive way. From the definition, it is clear that the underlying undirected structure of a comparability graph is a complete graph, since every pair of graphs from G is compared and ordered with respect to the descriptor's value. However such a representation does not reveal much information neither about the graphs nor the descriptors. Hence, the orientation (or the ordering) becomes mandatory and plays a crucial role in this analysis.



(a) Graph Edit Distance

(b) Correlation Matrix

Figure 3: Cluster analysis of the matrices $GedM$ and $CorM$ for the graphs from $\mathcal{G}_3$.

Other Comparability Graphs

In this section, we state another definition of a comparability graph for analyzing interrelations between network descriptors.

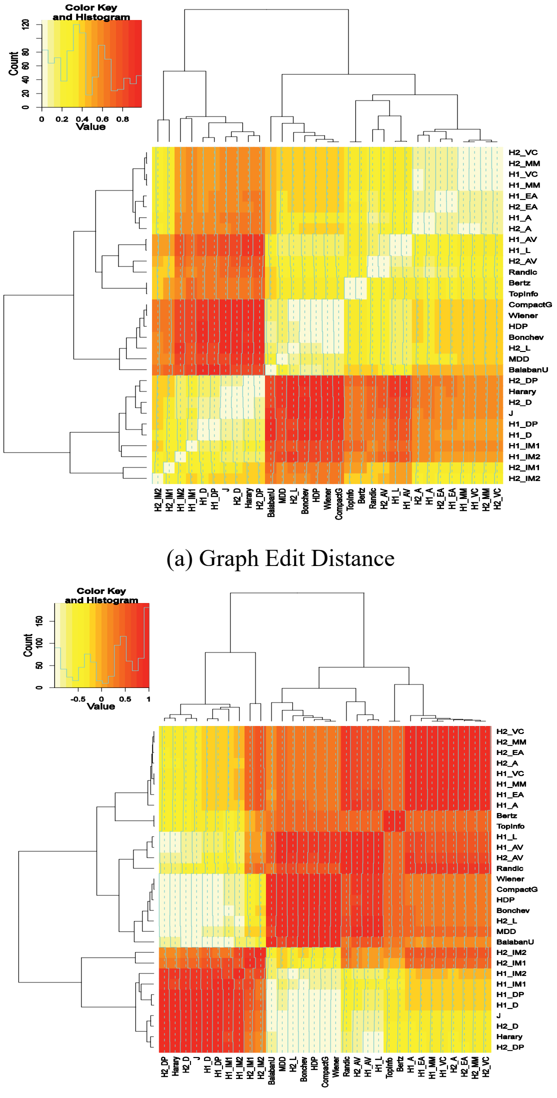
Definition 13. Given a graph class $\mathcal{G}$ and a set of descriptors $\mathcal{D} = \{D_1, \ldots, D_N\}$ computed for every graph $G \in \mathcal{G}$. Then, $CG_G = (V, E)$ where $V = \mathcal{D}$ and $E \subseteq V \times V$ such that there is an edge from D_i to D_j if and only if $D_i(G) \geq D_j(G)$, for $(i \neq j)$.

We see that the edge set E describes relations between the descriptors. Following this procedure, we obtain a collection of comparability graphs defined for each graph in $\mathcal{G}$. Clearly, such a comparability graph represents relations between the descriptors based on their values. In general, inferring such interrelations between network measures analytically is a challenging problem, see, *e.g.*, [45, 75, 76]. Note that analytical relationships by means of inequalities between information theoretic network measures have been called *information inequalities* [75-77]. Hence, a natural question arises namely how to compare such comparability graphs representing relations between network measures and analytically proven inequalities? Clearly, numerical values can be easily computed between each pair of graphs. However, it is not straightforward to prove analytical relations by means of inequalities between all pairs of given network measures. Also, note that the advantage of mathematical relations between descriptors is that they often hold for a graph class, not only for a single graph. From these arguments, it is evident that this problem needs further investigation in the future.

SUMMARY AND CONCLUSION

In this article, we reviewed the concept of comparability graphs and their properties. For this, we stated some mathematical results to characterize comparability graphs. Afterwards, we have discussed the applicability of comparability graphs in chemoinformatics and sketched existing results.

Also, we explored the relationship between the structural similarity of comparability graphs induced by using certain descriptors and the correlation of these descriptors. As a result, we found that highly correlated descriptors correspond to comparability graphs which are structurally similar (with respect to a graph similarity measure). Some of our findings also led to rediscover some of our earlier result from [64] *e.g.*, that highly non-correlated descriptors possess highly dissimilar comparability graphs. Apart from the presented results, we also studied various descriptors calculated by using the software DRAGON [78] and arrived at similar conclusions.



(a) Graph Edit Distance

(b) Correlation Matrix

Figure 4: Cluster analysis of the matrices *GedM* and *CorM* for the graphs from $\mathcal{G}_4$.

In the future, we would like to explore this kind of comparability graph in depth. This would involve using other graph similarity measures and studying their impact when clustering the data. Also, we already sketched (see Section Other Comparability graphs) that this concept could be useful to study the relatedness between graph measures. A study to explore this problem is already in development.

ACKNOWLEDGEMENTS

Matthias Dehmer and Lavanya Sivakumar thank the Austrian Science Fund for supporting this work (project P22029-N13). Also, we thank Kurt Varmuza for providing the data sets of molecular structures and the descriptors calculated using DRAGON.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

ABBREVIATIONS AND NOTATIONS

We have used standard definitions and notations from graph theory [28] and for other definitions of terms mentioned below, we refer to [64].

QSAR	=	Quantitative Structure Activity Relationships
QSPR	=	Quantitative Structure Property Relationships
χ	=	Vertex chromatic number of a graph
ω	=	Size of the largest clique in a graph
ρ	=	Pearson's Correlation coefficient
GED	=	Graph Edit Distance
GedM	=	Similarity matrix
CorM	=	Correlation matrix
J	=	Balaban index
U	=	Balaban-like information index

B	=	Bertz complexity index
BT	=	Bonchev-Trinajstic index
C	=	Compactness index
H	=	Harary index
HDP	=	Hyper-distance-path index
MDD	=	Mean Distance deviation
RC	=	Randic Connectivity index ′
$T\,IC$	=	Topological information content
W	=	Wiener index
$Z1$	=	Zagreb index
$A(G)$	=	Adjacency matrix of a graph G
$D(G)$	=	Distance matrix of a graph G
$EA(G)$	=	Extended-adjacency matrix of a graph G
$DP(G)$	=	Distance-Path Matrix of a graph G
$MM(G)$	=	Random-walk Markov matrix of a graph G
$VC(G)$	=	Vertex-connectivity matrix of a graph G
$IM_1(G)$, $IM_2(G)$	=	Weighted structure function matrices of a graph G

REFERENCES

[1] Cvetkovic, D. M.; Doob, M. and H. Sachs. *Spectra of Graphs. Theory and Application.* Academic Press, **1997**.

[2] Godsil, C. and Royle, G. *Algebraic Graph Theory*. Graduate Texts in Mathematics. Academic Press, **2001**.

[3] Halin, R. *Graphentheorie*. Akademie Verlag, **1989**. Berlin, Germany.

[4] Harary, F. *Graph Theory*. Addison Wesley Publishing Company, **1969**. Reading, MA, USA.

[5] Sobik, F. Modellierung von Vergleichsprozessen auf der Grundlage von Ähnlichkeitsmaßen für Graphen. *ZKI-Informationen, Akad. Wiss. DDR,* **1986**, *4*:104–144,.

[6] Gross, J. L. and Yellen, J. Graph *Theory and Its Applications, Second Edition.* Discrete Mathematics and Its Applications. Chapman & Hall, Boca Raton, **2006.**

[7] Emmert-Streib, F. and Dehmer, M. Networks for systems biology: Conceptual connection of data and function. *IET Systems Biology*, **2011,** *5*(3):185–207.

[8] Diudea, M. V.; Gutman, I. and Jäntschi, L. *Molecular Topology*. Nova Publishing, **2001**. New York, NY, USA

[9] Sommerfeld, E. and Sobik, F. Operations on cognitive structures - their modeling on the basis of graph theory. In D. Albert, editor, *Knowledge Structures*, pages 146–190. Springer**, 1994**.

[10] Emmert-Streib, F. The chronic fatigue syndrome: A comparative pathway analysis. *J. Comput. Biol*, **2007,** *14*(7).

[11] Morowitz, H. Some order-disorder considerations in living systems. *Bull. Math. Biophys.*, **1953,** *17*:81–86,.

[12] Olken, F. Graph data management for molecular biology. OMICS: A *Journal of Integrative Biology*, **2003,** *7*(1):75–78,.

[13] Bunke, H. What is the distance between graphs? Bulletin of the EATCS, **1983**, *20*, 35–39.

[14] Dehmer, M. and Mehler, A. A new method of measuring similarity for a special class of directed graphs. *Tatra Mountains Mathematical Publications*, **2007**, *36*, 39– 59.

[15] Randić, M. On characterization of molecular branching. ´ *J. Amer. Chem. Soc.*, **1975,** *97*:6609–6615.

[16] Wiener, H. Structural determination of paraffin boiling points. *J. Amer. Chem. Soc*, **1947,** *69*(17):17–20.

[17] Basak, S. C. Information-theoretic indices of neighborhood complexity and their applications. In J. Devillers and A. T. Balaban, editors, *Topological Indices and Related Descriptors in QSAR and QSPAR*, pages 563–595. Gordon and Breach Science Publishers, **1999**. Amsterdam, The Netherlands.

[18] Bonchev, D. Information *Theoretic Indices for Characterization of Chemical Structures*. Research Studies Press, Chichester, **1983**.

[19] R. Todeschini, V. Consonni, and R. Mannhold. *Handbook of Molecular Descriptors*. Wiley-VCH, **2002**. Weinheim, Germany.

[20] Kier , L. B. and Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*. Research Studies Press, **1986.** Letchworth, UK.

[21] Mukherjee, P.; Desai, P.; Ross, L.; White, and Averya, M. A. Structure-based virtual screening against sars-3clpro to identify novel non-peptidic hits. *Bioorganic & Medicinal Chemistry*, **2008,** *16*:4138–4149**.**

[22] Bonchev, D.; Kamenska, V.; and Mekenyan, O. Comparability graphs and molecular properties III.C9 andC10 alkanes. *Int J. Quantum Chem,* **1990,** *37(2)*, 135–153.

[23] Bonchev, D.; Kamenska, V. and Comparability graphs and molecular properties: IV Generalizations and Applications. *J. Math. Chem*., **1990**, *5*, 43–72.

[24] Bonchev, D.; and Mekenyan, O. Comparability graphs and electronic spectra of condensed benzenoid hydrocarbons. *Chemical Physics Letters*, **1983**, *98*,134–138.

[25] Bonchev D. and Mekenyan, O. Comparability graphs and molecular properties. a novel approach to the ordering of isomers. *J. Chem. Soc. Faraday Trans*., **1984**, *2*, 695–712.

[26] Ghouila-Houri, A. Caractérisation des graphes non orientés dont on peut orienter les aretes de manière à obtenir le graphe d'une relation d'ordre. ˇ *C. R. Acad. Sci. Paris*, **1962,** *254*:1370–1371.

[27] Gilmore, P. C. and Hoffman , A. J. A characterization of comparability graphs and of interval graphs. *Canad. J. Math*., **1964,** *16*:539–548.

[28] West, D.B. *Introduction to Graph Theory.* Prentice Hall, USA, **1996.**

[29] Wolk, E.S. A note on the comparability graph of a tree. *Proc. Amer. Math. Soc.*, **1965,** *16*:17–20.

[30] Golumbic, M.C. Trivially perfect graphs. *Discrete Mathematics*, **1978,** *24*:105–107.

[31] Olariu , S. On sources in comparability graphs, with applications. *Discrete Mathematics,* **1992,** *110*(1-3):289 – 292.

[32] Yang, X.; Wang, L.; Xue, J. Deng, Y. and Zhang, Y. Comparability graph coloring for optimizing utilization of stream register files in stream processors. *In Proceedings of the 14th ACM SIGPLAN symposium on Principles and practice of parallel programming, PPoPP '0*9, pages 111–120, New York, NY, USA, **2009**. ACM

[33] Diudea, M. V. QSPR / QSAR *Studies by Molecular Descriptors. Nova Publishing*, **2001.**

[34] K. Eric Drexler. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proceedings of the National Academy of Sciences*, **1981**, *78(9)*, 5275–5278.

[35] Bajorath, J. *Chemoinformatics: Concepts, Methods, and Tools for Drug Discovery*. Methods in Molecular Biology. Humana Press, **2004**. Totowa, NJ, USA.

[36] Balaban, A.T. Highly discriminating distance-based topological index. *Chemical Physics Letters*, **1982,** *89(5)*, 399 – 404.

[37] Balaban, A.T. and Balaban, T.S. New vertex invariants and topological indices of chemical graphs based on information on distances. *J. Math. Chem.,* **1991**, 8, 383–397.

[38] Gutman, I. The energy of a graph. *Ber. Math. Stat. Sekt. Forschungszentrum Graz*., **1978,** *103*:1–22**.**

[39] Minoli, D. Combinatorial graph complexity. *Atti. Accad. Naz. Lincei, VIII Ser., Rend., Cl. Sci. Fis. Mat. Nat*., **1975,** *59*:651–661**.**

[40] Bonchev, D. Information Theoretic Indices for Characterization of Chemical Structures. Research Studies Press, **1983**.

[41] Dehmer, M. and Mowshowitz, A. A history of graph entropy measures. *Information Sciences*, **2011**, *181(1)*, 57 – 78.

[42] Bunke, H. Recent developments in graph matching. In 15-th *International Conference on Pattern Recognition*, volume 2, pages 117–124, **2000**.

[43] Balaban, A.T.; Mekenyan, O.; and Bonchev, D. Unique description of chemical structures based on hierarchically ordered extended connectivities (HOC procedures). I. Algorithms for finding graph orbits and canonical numbering of atoms. *J. Comput. Chem*, **1985**, *6(6)*, 538–551.

[44] Bonchev, D.; Mekenyan, O. and Trinajstic, N. Isomer discrimination by topo- ´logical information approach. *J. Comp. Chem*., **1981,** *2(2)*, 127–148.

[45] Bonchev, D. and Trinajstic, N. Information theory, distance matrix, and molecular branching. *The Journal of Chemical Physics*, **1977,** *67(10)*, 4517–4533.

[46] Mekenyan, O.; Bonchev, D. and Balaban , A. T. Unique description of chemical structures based on hierarchically ordered extended connectivities. v. new topological indices, ordering of graphs, and recognition of graph similarity. *J. Comput. Chem*., **1984,** *5*(6):629–639.

[47] Balaban , A. T. Chemical graphs. I. *Rev. Roum. Chim*., **1966,** *11*, 1097–1116.

[48] Randić, M. and Wilkins, C.L. Graph theoretical ordering of structures as a basis for systematic searches for regularities in molecular data. The *Journal of Physical Chemistry*, **1979,** *83*(11):1525–1540**.**

[49] Randić, M. and Wilkins, C.L. On a graph theoretical basis for ordering of structures. *Chemical Physics Letters*, **1979,** *63*:332–336.

[50] Basak, S. C.; Magnuson, V. R.; Niemi, G. J. and Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math*., **1988,** *19*, 17–44.

[51] Randić, M. Similarity based on extended basis descriptors. *´ J. Chem. Inf. Comput. Sci.,* **1992,** *32*:686–692**.**

[52] Randić, M. and Wilkins, C.L. Graph theoretical approach to recognition of structural similarity in molecules. *J. Chem. Inf. Comput. Sci*., **1979,** *19*:31–37.

[53] Santini, S. and Jain, R. Similarity measures. *IEEE Trans. Pattern Anal. Mach. Intell.*, **1999,** *21*(9):871–883.

[54] Skvortsova, M. I.; Baskin, I. I.; Stankevich, I. V.; Palyulin, V. A. and Ze- firov, N. S. Molecular similarity in structure-property relationship studies. Analytical description of the complete set of graph similarity measures. In International *symposium CACR-96. Book of Abstracts*, page 16, **1996**.

[55] Skvortsova, M. I.; Baskin, I. I.; Stankevich, I. V.; Palyulin, V. A. and Ze- firov, N. S. Molecular similarity. 1. analytical description of the set of graph similarity measures*. J. Chem. Inf. Comput. Sci*., **1998,** *38*:785–790.

[56] Ivanciuc, T.; Ivanciuc, O. and Klein, D. J. Prediction of environmental properties for chlorophenols with posetic quantitative super-structure/property relationships (QSSPR). *International Journal of Molecular Sciences*, 7(9):358–374**, 2006**.

[57] Klein, D. J. Prolegomenon on partial orderings in chemistry. *MATCH Commun. Math. Comput. Chem*., **2000,** *42*:7 –21**.**

[58] Restrepo, G.; Brüggemann, R. and Klein, D. J. Partially ordered sets: ranking and prediction of substances' properties. *Curr Comput Aided Drug Des*., **2011,** *7*:133–145.

[59] Bunke, H. and Allermann, G. A Metric on Graphs for Structural Pattern Recognition. In EUSIPCO, editor*, Proc. 2nd European Signal Processing Conference EUSIPCO*, pages 257–260, **1983**.

[60] Gao, X.; Xiao, B.; Tao, D. and Li, X. A survey of graph edit distance. *Pattern Analysis & Applications*, **2010,** *13*:113–129. 10.1007/s10044-008-0141-y.

[61] Robles-Kelly, A. and Hancock, R. Edit distance from graph spectra. *In Proceedings of the IEEE International Conference on Computer Vision,* pages 234–241, **2003**.

[62] Robles-Kelly, A. and Hancock, R. Graph edit distance from spectral seriation. *IEEE Trans. Pattern Anal. Mach. Intell.,* **2005,** *27*(3):365–378**.**

[63] Hastie, T.; Tibshirani, R.; and Friedman, J. H. *The elements of statistical learning*. Springer, Berlin, New York, **2001**.

[64] Dehmer , M. and Sivakumar, L. Uniquely discriminating molecular structures using novel eigenvalue-based descriptors. *MATCH Commun. Math. Comput. Chem*., **2012**, *67(1)*, 147–172.

[65] Janežic, D.; Miličevič, A.; Nikolić, S. and Trinajstić, N. *Graph Theoretical Matrices in Chemistry*, volume 3. Mathematical Chemistry Monographs, Kragujevac, **2007.**

[66] Balaban, A. T.; and Ivanciuc, O. Historical development of topological indices. In J. Devillers and A. T. Balaban, editors, *Topological Indices and Related Descriptors in QSAR*

and QSPAR, pages 21–57. Gordon and Breach Science Publishers, **1999**. Amsterdam, The Netherlands.

[67] Doyle, J. K.; and Garver, J. E. Mean distance in a graph. *Discrete Mathematics*, **1977,** *17*, 147–154.

[68] Skorobogatov, V. A. and Dobrynin, A. A. Metrical analysis of graphs. *Commun. Math. Comp. Chem*., 23:105–155, **1988**.

[69] Mowshowitz, A. Entropy and the complexity of graphs: I. an index of the relative complexity of a graph. *Bull. Math. Biophys*, **1968,** *30*:175– 204.

[70] Bertz, S. H. The first general index of molecular complexity. *J Am Chem Soc*, **1981**, *103*, 3241–3243.

[71] Dehmer , M.; Barbarini, N.; Varmuza, K. and Graber, A. A large scale analysis of information-theoretic network complexity measures using chemical structures. *PLoS ONE*, **2009**, *4(12)*, 1–13, 12.

[72] Dehmer, M.; Laurin A.; Mueller, J. and Graber, A. New polynomial-based molecular descriptors with low degeneracy. *PLoS ONE*, **2010**, *5(7)*, e11393, 07.

[73] Molgen. Molgen isomer generator software. www.molgen.de, **2000**. Institute of Mathematics II, University of Bayreuth, Germany.

[74] Dehmer, M.; Varmuza, K.; Borgert, S. and Emmert-Streib, F. On entropy-based molecular descriptors: Statistical analysis of real and synthetic chemical structures. *J. Chem. Inf. Model*., **2009**, *49*, 1655–1663.

[75] Dehmer, M.; Borgert, S. and Emmert-Streib, F. Entropy Bounds for Molecular Hierarchical Networks. *PLoS ONE*, **2008**, *3(8)*, e3079.

[76] Dehmer, M.; and Mowshowitz, A. Inequalities for entropy-based measures of network information content. *Applied Mathematics and Computation,* **2010**, *215(12)*, 4263 – 4271.

[77] Dehmer, M.; and Lavanya Sivakumar. Recent developments in quantitative graph theory: Information inequalities for networks. PLoS ONE, **2012**, *7*(2): e31395. doi:10.1371/journal.pone.0031395**.**

[78] Todeschini, R.; Consonni, V.; Mauri, A. and Pavan, M. Dragon, software for calculation of molecular descriptors. www.talete.mi.it, **2004**. Talete srl, Milano, Italy.

CHAPTER 8

Basic Concepts and Applications of Molecular Topology to Drug Design

Jorge Gálvez*, María Gálvez-Llompart and Ramón García-Domenech

Molecular Connectivity and Drug Design Research Unit, Faculty of Pharmacy, Department of Physical Chemistry, University of Valencia Avd, V.A. Estellés, s/n 46100-Burjassot, Valencia, Spain

Abstract: This chapter deals with the use of molecular topology (MT) in the selection and design of new drugs. After an introduction of the actual methods used for drug design, the basic concepts of MT are defined, including examples of calculation of topological indices, which are numerical descriptors of molecular structures. The goal is making this calculation familiar to the potential students and allowing a straightforward comprehension of the topic. Finally, the achievements obtained in this field are detailed, so that the reader can figure out the great interest of this approach.

Keywords: Molecular topology, drugs, drug design, topological indices, molecular structure, computer-aided drug discovery and development, virtual screening, chemical libraries, quantitative structure-activity relationships, quantitative structure-property relationships, molecular descriptors, connectivity indices, modeling, molecular design, molecular connectivity.

INTRODUCTION

It was Corvin Hansch [1] who introduced, at early sixties in the past century, an equation linking some experimental properties of molecules with physicochemical parameters taking into account their electronic and steric characteristics. This is generally considered as the birth of the so called quantitative structure-activity relationship (QSAR) methods. Since then, the development of these methods, which require the use of computer (*in silico*), has been extraordinary, so much so that today they are in regular use worldwide.

***Corresponding author Jorge Gálvez:** Molecular Connectivity & Drug Design Research Unit, Faculty of Pharmacy, Department of Physical Chemistry, University of Valencia, 46100 Burjassot, Valencia, Spain; Tel: 34-6-3544891; Fax: 34-6-354 48 92; E-mail: jorge.galvez@uv.es

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50008-4

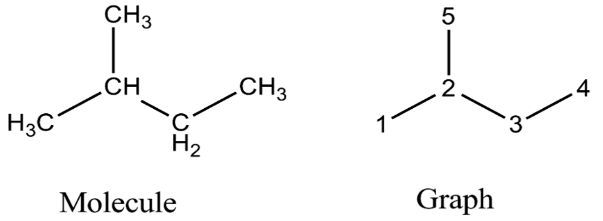
In the work of Hansch, the basic assumption was the existence of an intrinsic relationship between the experimental properties and the structure of the chemical compound. Of course, such a qualitative relationship was well known long time before Hansch, but his merit was to provide a quantitative measure of it (QSAR). This quantitative approach has resulted in the use of powerful computers capable of predicting the properties of compounds even before they are obtained in the laboratory or to design new ones with the desired properties.

A significant advance in the field of QSAR was the introduction by Cramer in 1988 of the Comparative Molecular Field Analysis (CoMFA). This method, based on the three-dimensional structure of the molecules, enabled the development of 3D-QSAR [2] that started a new era both in the conceptual and practical viewpoints. In the case of drugs, the effects from different conformers, stereoisomers or enantiomers in 3D-QSAR models, permitted the comparison of molecular structures so that it was possible to disclose the structural arrangement of atoms responsible for the activity, known as the pharmacophore [3]. However, a common criticism to Cramer's approach is that, given that the molecular fields are compared within a grid, the outcome can be closely dependent on the size of the grid; though there are solutions for this problem.

Other 3D-QSAR approaches, such as Comparative Molecular Similarity Indices Analysis (CoMSIA) [4], Self Organizing Molecular Field Analysis (SomFA) [5], or GRID/GOLPE have also demonstrated significant efficacy in drug design/discovery. Some of these methods may be used to compare different sets of molecular descriptors.

Along with the rapid development of computational science, a pull of new techniques based on formalisms such as molecular mechanics, molecular dynamics, docking, scoring and pharmacophore analysis, are now widely used in the area of drug discovery. These computational techniques have been proven to assist in the design of novel, more potent and specific drugs as they can visualize the mechanisms of ligand-receptor interactions.

Molecular topology (MT), a discipline typically related to the QSAR methods, has demonstrated to be an excellent tool for a quick and accurate prediction of many physicochemical and biological properties [6-8]. One of the most interesting advantages of MT is the straightforward calculation of molecular descriptors to work with. Within this mathematical formalism, a molecule is assimilated to a graph, where each vertex represents one atom and each edge one bond. Starting from the interconnections between the vertices, an adjacency topological matrix can be built up, whose *ij* elements take the values either one or zero, depending if the vertex *i* is connected or unconnected to the vertex *j*, respectively (Fig. (**1**)). The valence or degree of each vertex δ_i is the number of edges converging on it, which is equal to the sum of the terms that are in the row (or column) corresponding to that vertex. The manipulation of this matrix gives origin to a set of topological indices or topological descriptors which characterize each graph and allow the developments of quantitative structure–property relationships (QSPR) [9–11] and QSAR [12–17] analysis as well.



$$A=\begin{bmatrix}0&1&0&0&0\\1&0&1&0&1\\0&1&0&1&0\\0&0&1&0&0\\0&1&0&0&0\end{bmatrix}\overset{\delta}{\begin{bmatrix}1\\3\\2\\1\\1\end{bmatrix}}$$

Adjacency matrix

Figure 1: The chemical graph and adjacency matrix of isopentane.

The use of MT within the framework of QSAR has grown in the last years in an exponential way. Altogether the topological scope covers over 20% of the overall papers on QSAR. A recent search, carried out with the Scifinder Scholar database, disclosed that about 3,466 papers out of 17,664 dealing on QSAR, were devoted to topological descriptors. Today there is an increasing number of authors applying MT to drug discovery and design, particularly in the field of anticancer compounds [18], D1 Dopaminergic antagonists [19], anti-convulsants [20], anti-HIV compounds [21], tyrosinase inhibitors [22], MAO-A inhibitors [23], antimalarials [24], and immunosuppressive compounds [25]. In the current work, we focus on the contribution of molecular topology to QSAR studies obtained by our research group in the last years and its application to drug design/discovery.

METHODOLOGY AND APPLICATIONS

The first to show that a molecule could be represented by a set of points (atoms) linked by edges (bonds), was the British mathematician James J. Sylvester in 1874; these representations were called graphs. It is interesting that the word 'graph' rose in a completely interdisciplinary context. In 1878 Sylvester published in *Nature* a paper entitled 'Chemistry and Algebra', where the word graph was used for the first time as a derivation of the term used by chemists of the end of the 19th century to refer to the molecular structure.

Molecular Descriptors

It is well known that a key issue to ensure the success of a QSAR approach is the selection of the adequate descriptors. Basak *et al*. [26] classified these descriptors into four categories:

a. Topostructural indices (TS), which quantify information regarding the connectivity, adjacency, and distances between atoms or vertices according to graph theoretical nomenclature—ignoring their distinct chemical nature.

b. Topochemical indices (TC), which are sensitive to both the pattern of connectedness of the atoms and their chemical and bonding characteristics.

c. 3D or geometrical parameters (3D).

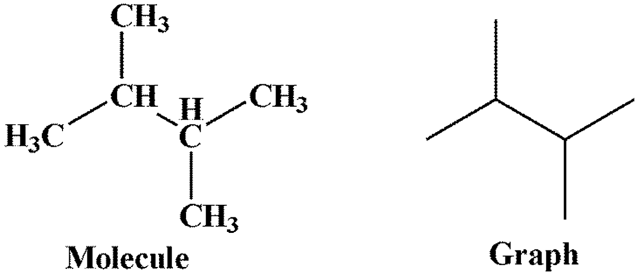
d. Quantum chemical descriptors (QC), which encode electronic aspects of chemical structure.

Today, it is known that topostructural and topochemical information can explain the main part of the predicted properties, and that the inclusion of three-dimensional features results in slightly improved predictive models in many cases [27], which is a surprising feature.

A complete review of topological descriptors is almost impossible due to the great quantity of such indices that are published in the literature and the number of them that are introduced every year, which continues to grow more and more. In the following we discuss the most important indices used by our research group, described in increasing order of complexity.

Discrete Invariants

They are real (most natural) numbers calculated from what the chemists understand qualitatively as the chemical structure. *N* is the number of non-hydrogen atoms, this is, the number of molecular graph vertices [28, 29] represented by V_k, where *k* represents the degree of the vertex, taking values of 3 or 4, which applies to atoms having *k* bonds (σ or π), except hydrogen atoms [29]. *R* is the branching number of number of single structural branches in the graph [29]. *L* is the length, *i.e.*, the maximal distance between non-hydrogen atoms in terms of bonds, that is, the diameter of the molecular graph defined as max(d_{ij}) [29]. In other words, the graph diameter is the number of edges between the two most separated vertices in the graph, by the shortest path (topological distance). PR_k, are pairs of ramifications at distance *k*, where *k* is between 0 and 3, *i.e.*, they are the number of pairs of single branches at distance *k* in terms of bonds [29]. *E* = shape factor = $\Sigma n_i d_i / L$, were n_i is the number of vertices situated at a d_i distance from the main path, which is the path linking the two vertices farther apart each other in the graph. The shape factor, *E*, evaluates the graph shape, *i.e.* the lower the *E* value, the more "linear" and elongated the graph is [30]. Fig. (**2**) shows the values of these descriptors for the molecule of 2,3-dimethylbutane.



Descriptors:

N=6	**R=2**	**L=3**	**E=2.67**	**PR0=0**
PR1=1	**PR2=0**	**PR3=0**	**V3=2**	**V4=0**

Figure 2: Discrete invariants of 2,3-dimethylbutane.

Some biological properties showed a curious dependence with these indices. For example, the relation between the plasma protein binding, PPB, and E for a group of cephalosporins was clearly nonlinear (see Fig. (**3**)).

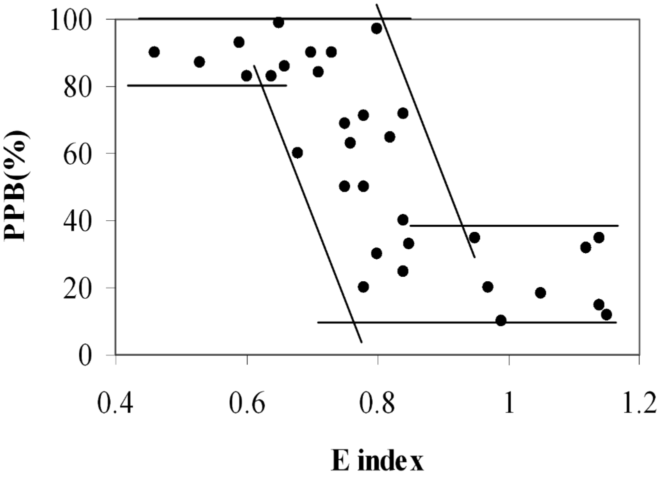


Figure 3: Relationship between plasma protein binding (PPB) *vs.* the shape factor E for a group of cephalosporins.

From Fig. (**3**), it is clear that there are no molecules with small PPB rate and *E* values between 0.4 and 0.7, and the degree of PPB is below 35% for *E* > 0.86. Although PPB is a complex property in which other factors such as lipid solubility play an important role, these results point to a notable influence of molecular shape on the PPB rate for these drugs [30]. It is remarkable, however, that such a 2D parameter as simple as *E*, can account for such a complex property.

Connectivity Indices

The first connectivity index, the branching index or Randić index [31], χ , was introduced by professor Milan Randić at Drake University in 1975. This descriptor is defined as the sum of the reciprocals of the square roots of products of the valences of the two vertices adjacent to each edge, extended to all edges of the graph. Fig. (**4**) shows the detailed calculation of the Randić index for isopentane.

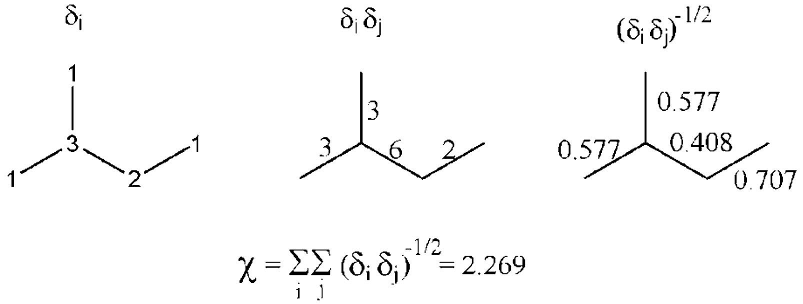


Figure 4: Randić index, χ, for the isopentane.

An example that illustrates the role of Randić index in describing structural features influencing non-specific drug action, is the relation observed between the log minimum blocking concentration (logMBC), related to 90% non-specific local anesthetic activity, and χ [6]. The simple χ term was found to correlate closely with logMBC value, as follows:

$$\text{logMBC} = 3.60 - 0.779\chi \qquad (1)$$

$r = 0.982$, $s = 0.409$, $N = 36$

Table **1** shows the results of prediction for each analysed compound.

In 1976 Kier and Hall extended the Randić index, introducing the connectivity indices of order *k* [32] and type *t*. They are the first example of "family of indices" and the entire set can be calculated from the adjacency matrix. They are normally written as, $^{k}\chi_t$, where *k* varies between 0 and *n*. Here the order is the number of connected non-hydrogen atoms that appear in a given sub-structure; in other words, the number of edges in the connected subgraph. Hence, the connectivity indices are defined as [6]:

$$^{k}\chi_t = \sum_{j=1}^{^{k}n_t}\left(\prod_{i\in S_j}\delta_i\right)^{-1/2} \qquad (2)$$

where δ_i is the number of simple bonds (σ bonds only) of the atom *i* to non-hydrogen atoms, S_j represents the j^{th} sub-structure of order *k* and type *t*, $^{k}n_t$ is the total number of sub-graphs of order *k* and type *t* that can be identified in the molecular structure. Types used are path (*p*), cluster (*c*) and path-cluster (*pc*). Upon the concepts defined in the Introduction, a sub-graph of type *p* is formed by a path, a sub-graph of type *c* is formed by a star (a graph in which all vertices are attached to a central one), while a *pc* sub-graph can be defined as every tree which is neither a path nor a star. As an example, Table **2** displays all the *p*, *c* and *pc* sub-graphs found in a simple molecular structure.

Table 1: Local Anesthetic activity and χ

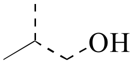
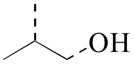
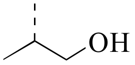
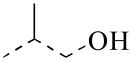
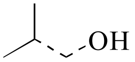
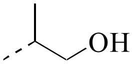
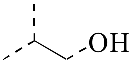
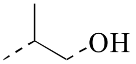
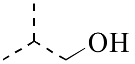
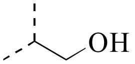
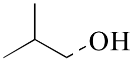
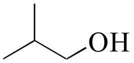
Anesthetic	logMBCexp	logMBCcalc	Anesthetic	logMBCexp	logMBCcalc
Methanol	3.090	2.790	Quinoline	0.300	0.528
Ethanol	2.750	2.470	8-Hydroxyquinoline	0.300	0.174
Acetone	2.600	2.230	Heptanol	0.200	0.567
2-Propanol	2.550	2.230	2-Naphthol	0.000	0.228
Propanol	2.400	2.090	Methylanthranilate	0.000	-0.072
Urethane	2.000	1.440	Octanol	-0.160	0.186
Ether	1.930	1.710	Thymol	-0.520	0.052
Butanol	1.780	1.710	*o*-Phenanthroline	-0.800	-0.602
Pyridine	1.770	1.650	Ephedrine Procaine	-2.470	-2.743
Hydroquinone	1.400	1.050	Lidocaine	-1.960	-2.310
Aniline	1.300	1.350	Diphenhydramine	-2.800	-2.750
Benzyl Alcohol	1.300	0.935	Tetracaine	-2.900	-3.030
Pentanol	1.200	1.330	Phenyltoloxamine	-3.200	-2.740
Phenol	1.000	1.350	Quinine	-3.600	-3.850
Toluene	1.000	1.350	Physostigmine	-3.660	-2.520
Benzimidazole	0.810	0.901	Caramiphen	-4.000	-3.480
Hexanol	0.560	0.949	Dibucaine	-4.200	-4.970
Nitrobenzene	0.470	0.651			

To take into account the presence of heteroatoms (atoms other than carbon) in the molecule, δ^{v} is used instead of δ, what allows encoding the influence of π and lone-pair electrons [6].

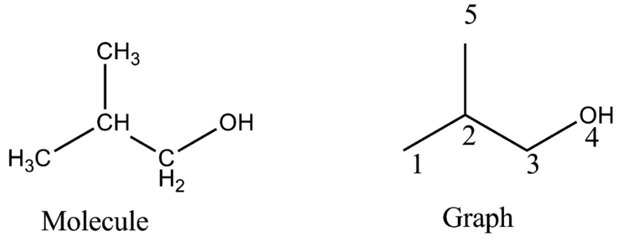
$$^{k}\chi_{t}^{\mathrm{v}} = \sum_{j=1}^{^{k}n_{t}} \left(\prod_{i \in \mathrm{S}_{j}} \delta_{i}^{\mathrm{v}} \right)^{-1/2} \tag{3}$$

Here $\delta_i^v = Z^v - H$, where Z^v is the number of valence electrons and H the number of hydrogen atoms bonded to the heteroatom.

Table 2: Subgraphs within the 2-methylpropanol structure

Type	Order 1	Order 2	Order 3	Order 4
Path				
				
				
				
Cluster				
Path-Cluster				

To illustrate the calculation of the connectivity indices up to order four, Fig. (**5**) shows the example of 2 methyl propanol.



$$A = \begin{array}{|ccccc|c} & & & & & \delta \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{array} \qquad A^v = \begin{array}{|ccccc|c} & & & & & \delta^v \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 3 \\ 0 & 1 & 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 & 0 & 5 \\ 0 & 1 & 0 & 0 & 0 & 1 \end{array}$$

Figure 5: Kier and Hall connectivity indices for 2 methyl propanol.

$$^{0}\chi=\sum\left(\delta_{i}\right)^{-1/2} = 3\cdot(1)^{-1/2}+(3)^{-1/2}+(2)^{-1/2}=4.285$$

$$^{0}\chi^{v}=\sum\left(\delta_{i}^{v}\right)^{-1/2}=2\cdot(1)^{-1/2}+(3)^{-1/2}+(2)^{-1/2}+(5)^{-1/2}=3.732$$

$$^{1}\chi=\sum_{s=1}^{n}\left(\delta_{i}\delta_{j}\right)_{s}^{-1/2}=2\cdot(1x3)^{-1/2}+(2x3)^{-1/2}+(2x1)^{-1/2}=2.27$$

$$^{1}\chi^{v}=\sum_{s=1}^{n}\left(\delta_{i}^{v}\delta_{j}^{v}\right)_{s}^{-1/2}=2\cdot(1x3)^{-1/2}+(3x2)^{-1/2}+(2x5)^{-1/2}=1.879$$

$$^{2}\chi=\sum_{s=1}^{n}\left(\delta_{i}\delta_{j}\delta_{k}\right)_{s}^{-1/2}=(1x3x1)^{-1/2}+(1x3x2)^{-1/2}+(1x3x2)^{-1/2}+(2x3x1)^{-1/2}=1.802$$

$$^{2}\chi^{v}=\sum_{s=1}^{n}\left(\delta_{i}^{v}\delta_{j}^{v}\delta_{k}^{v}\right)_{s}^{-1/2}=(3)^{-1/2}+(6)^{-1/2}+(6)^{-1/2}+(30)^{-1/2}=1.576$$

$$^{3}\chi_{p}=\sum_{s=1}^{n}\left(\delta_{i}\delta_{j}\delta_{k}\delta_{l}\right)_{s}^{-1/2}=(6)^{-1/2}+(6)^{-1/2}=0.816$$

$$^{3}\chi_{p}^{v}=\sum_{s=1}^{n}\left(\delta_{i}^{v}\delta_{j}^{v}\delta_{k}^{v}\delta_{l}^{v}\right)_{s}^{-1/2}=(30)^{-1/2}+(30)^{-1/2}=0.365$$

$$^{3}\chi_{c}=\sum_{s=1}^{n}\left(\delta_{i}\delta_{j}\delta_{k}\delta_{l}\right)_{s}^{-1/2}=(6)^{-1/2}=0.408$$

$$^{3}\chi_{c}^{v}=\sum_{s=1}^{n}\left(\delta_{i}^{v}\delta_{j}^{v}\delta_{k}^{v}\delta_{l}^{v}\right)_{s}^{-1/2}=(6)^{-1/2}=0.408$$

$$^{4}\chi_{c}=\sum_{s=1}^{n}\left(\delta_{i}\delta_{j}\delta_{k}\delta_{l}\delta_{m}\right)_{s}^{-1/2}=(6)^{-1/2}=0.408$$

$$^{4}\chi_{c}^{v}=\sum_{s=1}^{n}\left(\delta_{i}^{v}\delta_{j}^{v}\delta_{k}^{v}\delta_{l}^{v}\delta_{m}^{v}\right)_{s}^{-1/2}=(30)^{-1/2}=0.183$$

It is useful the use of combinations of connectivity indices, as for instance the differences and quotients between valence and non-valence indices: ${}^{k}\mathrm{D}_{t} = {}^{k}\chi_{t} - {}^{k}\chi_{t}^{\mathrm{v}}$ and ${}^{k}\mathrm{C}_{t} = \dfrac{{}^{k}\chi_{t}}{{}^{k}\chi_{t}^{\mathrm{v}}}$ [29,33]

The connectivity indices are among the most widely used in QSAR [6, 34].

A good example of their excellent predictive capability is a work from Kier in which he applied the connectivity indices in predicting the sweet or bitter taste of a group of aldoximes [35]. He applied discriminant analysis including a training (molecules used to get the discriminant function) and test (external molecules) sets. The best linear discriminant function was the following two-variable equation: DF=1.21 ${}^{1}\chi - 3.88\ {}^{4}\chi_{\mathrm{p}} - 3.27$.

Fig. **(6)** shows the results obtained for the training and test sets.

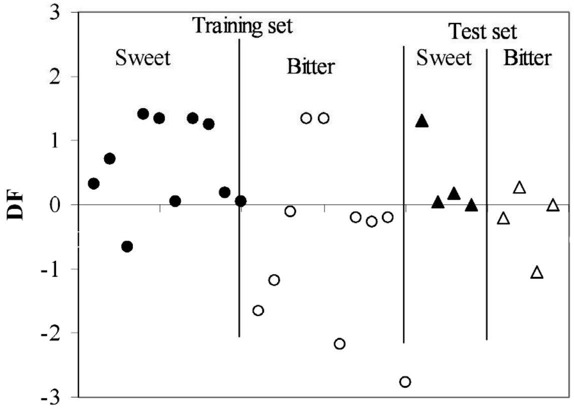


Figure 6: Prediction of taste potency relative to sucrose of an aldoximes group by molecular connectivity from Kier's results.

Topological Charge Indices (TCI)

In 1994 our team introduced the Topological Charge Indices (TCI), namely G_k and J_k, of order k for a given graph, where k ranges from 1 to 5, which are defined as [28]:

$$\mathrm{G}_k = \sum_{i=1}^{\mathrm{N}\text{-}1} \sum_{j=i+1}^{\mathrm{N}} \left| \mathrm{c}_{ji} \right| \delta(k, \mathrm{d}_{ij}) \text{ and } \mathrm{J}_k = \frac{\mathrm{G}_k}{\mathrm{N}-1} \tag{4-5}$$

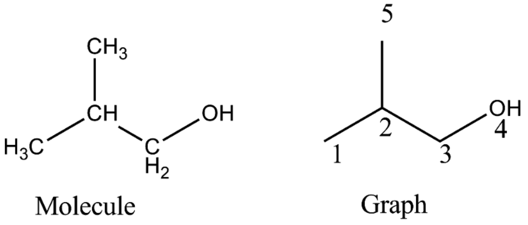
Here N is the previously defined number of vertices in the graph-molecule, c_{ij} is the charge term between vertices i and j, which is defined as $\mathrm{c}_{ij} = \mathrm{m}_{ij} - \mathrm{m}_{ji}$. δ_{ij} represents the *Krönecker* delta symbol ($\delta_{ij} = 1$ if $i = j$ and 0 otherwise); d_{ij} is the topological distance between the vertices i and j.

The variables m_{ij} and m_{ji} are the elements of the square $N\times N$ matrix $\boldsymbol{M}$ obtained as the product of two matrices $\boldsymbol{A}$ and $\boldsymbol{Q}$, *i.e.* $\boldsymbol{M} = \boldsymbol{A}\cdot\boldsymbol{Q}$. Consequently:

$$\mathrm{m}_{ij} = \sum_{h=1}^{\mathrm{N}} a_{ih}\, \mathrm{q}_{hj} \tag{6}$$

A is the *adjacency* matrix in which elements a_{ih} are 0 if $i = h$ or one of the following values if $i \neq h$: 1 if i is bonded to h *via* a single bond; 1.5 if the bond is aromatic; 2 if it is a double bond; and 3 if it is a triple one. $\boldsymbol{Q}$ is the inverse squared distance or *Coulombian* matrix. Its elements, q_{hj}, are 0 if $h = j$; otherwise, $q_{hj} = 1/d_{hj}^{2}$, where d_{hj} is the topological distance between vertices h and j. Thus, G_k represents the overall sum of the c_{ij} charge terms for every pair of vertices i and j at a topological distance k. The valence TCIs, G_k^{v} and J_k^{v}, are defined in a similar fashion, by substituting the matrix $\boldsymbol{A}$ by A^{v}, the electronegativity-modified adjacency matrix. The elements of both matrices are identical except for the main diagonal of A^{v}, which are obtained by replacing the zeroes in the main diagonal by the corresponding Pauling electronegativity values *EN*, normalized for chlorine electronegativity = 2. Hence, any other heteroatom will have the proportional value according to the Pauling's scale. The *G* charge indices are obtained as the algebraic sum of the differences between the terms m_{ij} and m_{ji}, whereas the *J* indices are just *J*/*N*. It is interesting to realize that the G_k index represent the average charge transferred at a distance k between all pairs of atoms in the molecule, whilst the J_k index is the mean G_k value per atom. What is remarkable is that these charge transfers have been evaluated within a pure mathematical framework.

To illustrate the calculation of the TCIs, Fig. (**7**) exhibits the example of 2-methyl propanol.



$$A = \begin{vmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{vmatrix}$$

Adjacency matrix

$$Q = \begin{vmatrix} 0 & 1 & 1/4 & 1/9 & 1/4 \\ 1 & 0 & 1 & 1/4 & 1 \\ 1/4 & 1 & 0 & 1 & 1/4 \\ 1/9 & 1/4 & 1 & 0 & 1/9 \\ 1/4 & 1 & 1/4 & 1/9 & 0 \end{vmatrix}$$

Coulombian matrix

$$M = \begin{vmatrix} 1 & 0 & 1 & 1/4 & 1 \\ 1/2 & 3 & 2/4 & 11/9 & 1/2 \\ 10/9 & 1/4 & 2 & 1/4 & 10/9 \\ 1/4 & 1 & 0 & 1 & 1/4 \\ 1 & 0 & 1 & 1/4 & 1 \end{vmatrix}$$

Matrix M=AxQ

Charge indices

$G1 = 1/2+1/4+1/2+1/4 = 1.5$ $J1 = G1/4 = 0.375$
$G2 = 1/9+0+2/9+1/9 = 0.44$ $J2 = G2/4 = 0.111$
$G3 = 0$ $J3 = G3/4 = 0$

Figure 7: Non valence charge indices for 2-methyl propanol.

As an example illustrating the high performance of TCIs, we show a study on xanthine-oxidase inhibition by 22 flavonoids, including flavones, flavonols, flavanones and chalcones, in which TCIs were employed to establish the structure-activity relationship model. Flavonoids were classified into four groups according to their activity on xanthine-oxidase (inactive, low, significant or high), and linear discriminant analysis (LDA) was used to classify each compound within a group. The results led to a very good one-index model, which was able to classify correctly as xanthine oxidase inhibitors not only the molecules in the training set but also those of an external test set of very heterogeneous compounds, such as allopurinol, caffeic acid, esculetin, and alloxantin [36].

Table **3** shows the classification functions obtained from LDA. The topological charge index *J2* takes into account the average value of the charge transferred between atoms placed at a topological distance = 2. This means that the intramolecular charge transfers between atoms located at such distance play an important role in this property. A possible explanation is related to the presence of hydroxyl groups on the positions 5, 7, and 4’, which enhance the inhibitory effect, whereas the presence of methoxy groups clearly weaken such an effect.

Table 3: Classification functions obtained from linear discriminant analysis

		Groups		
	Inactive	**Low**	**Significant**	**High**
J2	116.31	137.37	150.1	193.2
constant	-24.85	-34.11	-40.46	-66.13

The key influence of the *J2* index is clearly outlined in Table **4**. Indeed, it is noteworthy that those compounds showing *J2* values lower than 0.48 are either inactive or little active. Most of the compounds with *J2* values between 0.48 and 0.58 are significantly active, while compounds showing values above 0.60 are highly active.

Table 4: Classification for each one of the compounds studied

Flavonoid	***J2***	**PIG (%)**	**Class$_{exp}$**	**Class$_{calc}$**
Training Set				
chalcone	0.341	0	inact.	inact.
4F-chalcone	0.417	0	inact.	inact.
flavone	0.486	3	inact.	low
flavanone	0.375	3.6	inact.	inact.
4(OCH$_3$)-chalcone	0.399	4.8	inact.	inact.
2’(OH),4(OCH$_3$)-chalcone	0.475	16.2	low	low
2’,4’(OH)$_2$ -3’(OCH$_3$)-chalcone	0.532	17.1	low	signif.
2’(OH),4’(OCH$_3$)-chalcone	0.463	17.2	low	low
5,7(OH)$_2$-flavanone	0.481	20.1	low	low
2’(OH)-chalcone	0.431	21.5	low	inact.
5,7,4’(OH)$_3$-flavanone	0.538	28.7	signif.	signif.
2’(OH),4F-chalcone	0.497	30.3	signif.	low
2’,4’(OH)$_2$-chalcone	0.484	39.8	signif.	low
4(OH)-chalcone	0.417	45.5	signif.	inact.
7(OH)-flavone	0.536	49.7	signif	signif
5,7(OH)$_2$ -6,8,4’(OCH$_3$)$_3$-flavone	0.653	61.4	signif.	high
3,5,7,2’,4’(OH)$_5$-flavone	0.72	70.1	high	high
5,7,4’(OH)$_3$ -6,8(OCH$_3$)$_2$-flavone	0.676	72.7	high	high
5,7(OH)$_2$ -6,4’(OCH$_3$)$_2$-flavone	0.636	77.8	high	high
5,7(OH)$_2$-flavone	0.58	91.1	high	signif
3,5,7,3’,4’(OH)$_5$-flavone	0.72	91.4	high	high

Table 4: contd….

3,7,3',4'$(OH)_4$-flavone	0.689	91.8	high	high
Test Set				
allopurinol	0.648	-	active	high
probenecid	0.327	-	inact.	inact.
sulfinpyrazone	0.294	-	inact.	inact.
caffeic acid	0.648	-	active	high
esculetin	0.741	-	active	high
TEI-6720	0.696	-	active	high
alloxantin	0.667	-	active	high

To test the efficacy of the discriminant function, a validation test with a set of compounds was carried out, including both, highly heterogeneous structures and significant inhibitory activity. All of them, namely allopurinol, caffeic acid, esculetin, TEI-6720 (2-(3-cyano-4-isobutoxyphenyl)-4-methyl-5-thiazolecarboxylic acid), and alloxantin, were correctly classified within the group of "high activity". Likewise, other uricosuric but not inhibitor compounds, such as probenecid or sulfinpyrazone, were also correctly classified as such.

The percent inhibition degree (PIG) on xanthine-oxidase, along with the *J2* values and the classification obtained for each flavonoid studied, are shown in Table **4**.

Other Molecular Descriptors

Among other widely used topological descriptors stand the ***Wiener path number W*** [37], ***kappa indices*** of molecular shape and ***flexibility index*** [38, 39], ***Balaban J index*** [40], ***electrotopological state indices*** [41], ***spectral moments, μ***, [42], *etc.* The number of topological descriptors is actually very large and potentially much larger. Fortunately, there are a number of software programs that are commercially available to calculate them. Among these programs stand MOLCONNZ [43], DRAGON [44], POLLY [45], CODESSA [46], *etc.*

Statistic Techniques

Statistic tools are essential to get a good outcome in the QSAR equations. Although there are many choices available, two types of analysis are commonly used: The first is to predict quantitative properties (multilinear regression analysis, MLRA) and the second to recognize the category to which the compound belongs to (linear discriminant analysis, LDA).

Multilinear Regression Analysis (MLRA)

Once calculated the indices, they are correlated to the biological/pharmacological experimental values to get a multilinear regression equation:

$$P_i = A_o + \sum A_i X_i \tag{7}$$

where P_i is the experimental property, X_i are the topological indices, and A_o and A_i are the regression coefficients of the equation obtained.

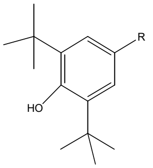
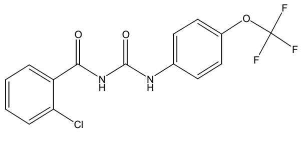
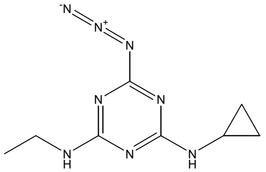
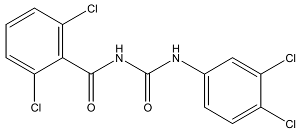
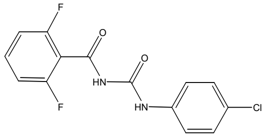
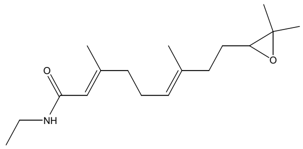
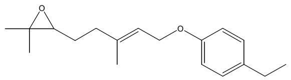
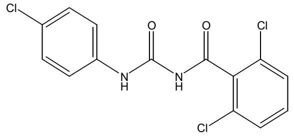
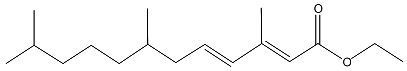
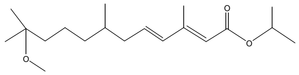
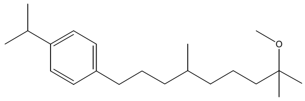
The predictability, quality and robustness of the model can be verified by means of different types of criteria. Usually three strategies are adopted [47]:

a) Internal validation or cross-validation with leave-one-out, LOO. To do this, one compound of the set is extracted, and the model is recalculated using as training set the remaining $N - 1$ compounds. The property is then predicted for the removed element. This process is repeated for all the compounds of the set, obtaining a prediction for everyone. From the residual values obtained, the standard error of estimates for the cross-validation, *SEE(CV)* and prediction coefficient, r^2_{cv}, (Q^2) are determined. A more robust stability validation method is the leave-some-out, LSO [48] in which we proceed the same way but leaving out not one but several compounds.

b) External validation. The model's predictive capability is tested by its application over an external set of molecules.

c) Data randomization or Y-scrambling. In order to evidence the possible existence of fortuitous correlations, a randomization test can be performed [49]. To do this, the values of the property of each compound are randomly permuted and linearly correlated with the topological descriptors. The process is repeated, as many times as compounds there are in the set.

Table 5: Chemical structures of the IGRs studied

Butyl substituted phenols 	**1** COMP01 (R=Cl) **2** COMP02 (R=Br) **3** COMP04 (R=NO_2) **4** COMP05 (R=cyano) **5** COMP10 (R=CH_3)	**6** COMP11 (R=OCH_3) **7** COMP13 (R=n-butyl) **8** COMP31 (R=morpholinocarbonyl) **9** MON585 (R=α□α’-dimethylbenzyl)
10 BAYSIR8514 	**11** CGA19255 	**12** DU19111 
13 DIFLUBENZURON 	**14** CRD-9499 	**15** R20458 
16 TH6038 	**17** HYDROPRENE 	**18** METHOPRENE 
19 MV678 		

Example: Prediction of Potency of Insecticides Against Malaria Vectors [50].

MT was employed to predict the potency of insecticides active against malaria vector mosquito (Culex). The insect growth regulators are substances that alter and interfere with development processes and insect growth. The group was composed of a representative sample of the various classes of IGRs, as for instance the juvenile hormone active (JHAs) mimetics; among them stand methoprene 18, hydroprene 17 and others, several butyl phenols, ureas (diflubenzuron 13 among others) and a triazine (CGA 19255 11). Table **5** shows the respective chemical structures.

Insecticidal activity was expressed as LC_{50}, which is the lethal dose in ppm causing 50% inhibition of adult emergence for larvae of *Culex pipiens quinquefasciatus*. The regression equation selected was:

$$\text{Log } LC_{50} = -2.632 + 14.991\ J_3^v - 0.239\ V4 \qquad (8)$$

N= 19 r^2=0.843 SEE=0.467 F=43.1

J_3^v takes into account the average charge transferred at a topological distance three per atom in the molecule. The fact that the index is weighted by the valence unveils the influence of heteroatoms such as N, O and Cl in the insecticidal activity. The second index, V4, is the simple sum of vertices (atoms different from hydrogen) with degree four. It includes quaternary carbons, carbonyl groups and carbons substituted on aromatic rings. These results are all the most consistent since the activity seems to depend on both, the charge transfers between donor and acceptor groups as well as of structural features such as steric hindrance or enhancement, encoded by the V4 index.

Table **6** compares the experimental and calculated values for each compound. As can be seen, it is worth pointing out the good concordance between them.

Linear Discriminant Analysis

As expressed before, the goal of the linear discriminant analysis, LDA, is to find a linear combination of variables allowing the discrimination between two or more categories or objects.

In our case, the "objects" are molecules. The final equation has the form:

$$DF = A_o + \sum A_i X_i \qquad (9)$$

Where *DF* is the value of the discriminant function related to a particular activity, X_i are the topological indices, and A_o and A_i are the regression coefficients relating one and others.

Although there are several choices, the simplest approach is the disjunctive, in which two sets of compounds are considered: One with proven pharmacological activity which constitute the “active” set and another one comprised of inactive compounds. The selection of the descriptors is based on the Fisher-Snedecor parameter, and the classification criterion is the shortest Mahalanobis distance (*i.e.* the distance of each case from the mean of all cases used in the regression equation). The quality of the discriminant function is evaluated by Wilks’ λ [49].

Table 6: Results obtained by multilinear regression analysis with IGRs

Compound	Log LC_{50exp}	J_3^v	V4	Log LC_{50calc}
COMP01	-0.032	0.2515	6	-0.292
COMP10	0.326	0.2812	6	0.152
COMP02	-0.222	0.2534	6	-0.264
COMP04	-0.444	0.2272	6	-0.657
COMP05	0.218	0.3020	7	0.227
COMP11	-0.244	0.2587	6	-0.185
COMP13	-0.620	0.2551	6	-0.238
COMP31	0.456	0.2737	7	-0.198
MON585	-1.699	0.2540	8	-0.732
BAYSIR8514	-2.699	0.1111	7	-2.636
CGA19255	-0.456	0.1915	4	-0.715
DU19111	-2.699	0.1059	8	-2.953
DIFLUBENZURON	-3.301	0.1133	7	-2.603
CRD-9499	-1.523	0.1538	4	-1.281
R20458	-1.398	0.1260	4	-1.697
TH6038	-1.886	0.1076	7	-2.689
HYDROPRENE	-1.000	0.1246	2	-1.242
METHOPRENE	-2.000	0.1272	3	-1.440
MV678	-1.699	0.1246	3	-1.479

The discriminant ability of the selected function is evaluated by:

a) The Classification matrix, in which each case is classified into a group according to the classification function. The number of cases classified into each group and the percentage of correct classifications are shown.

b) The Jack-knifed classification matrix, in which each case is classified into a group according to the classification functions computed from all the data except the case being classified.

c) The use of an External test set, which entails the use of an external compound set to check the validity of the selected discriminant functions.

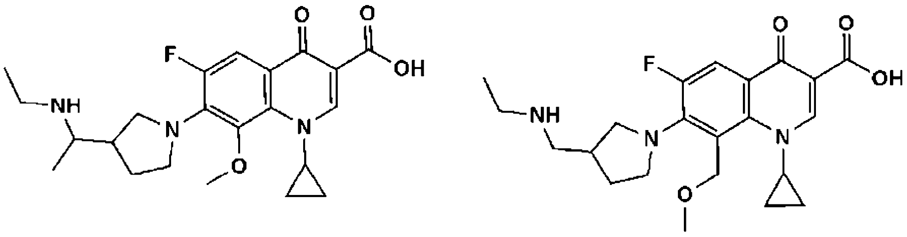
Example: Prediction of Quinolone Activity against Mycobacterium Avium [51].

In this example, a QSAR study using a database of 158 quinolones previously tested against *Mycobacterium avium-M. intracellular* (MAV) complex, was carried out. The goal was to find new active compounds against the MAV complex. Topological indices were used as structural descriptors and LDA was employed as statistical technique. Using a MIC cut-off of 32 mg/mL, the following equation was obtained in such a way that the compound was classified as active if DF >0 (MIC below 6 mg/mL), inactive if DF<0 (MIC above 32) or uncertain (MIC between 6 and 32):

$$DF= -2.6+20.1\,{}^{3}\chi_{ch} -12.9\,{}^{4}\chi_{c} +42.5\,{}^{4}\chi_{c}^{v} +25.6\,{}^{6}\chi_{ch} -2.2G_{3}^{v} +2.4G_{4}^{v} \qquad (10)$$

Statistical parameters were as follows: n=5 114, F=5 30.79, Wilk's $\lambda = 0.37$. The indices ${}^{4}\chi_{c}$ and ${}^{4}\chi_{c}^{v}$ represent the quaternary ramifications, ${}^{3}\chi_{ch}$ and ${}^{6}\chi_{ch}$ reflect the presence of cycles of three and six atoms, respectively, and G_3^v and G_4^v furnish information about the transfer of intramolecular charges between atoms separated by distances of 3 and 4, respectively. The ${}^{3}\chi_{ch}$ index made a marked contribution to the positivity of the equation, reflecting the role of the cyclopropyl substituent on nitrogen N-1 to anti-MAV activity. Sixty-one out of 77 quinolones with cyclopropyl substitutions were active *in vitro*, and all of them showed positive *DF* values.

A good example of the discriminating capacity of the model was the result obtained with two quinolones that had the same molecular weight and large structural similarities but very different anti-MAV complex activities. The DF function value was 0.9765 for PD139586, which is active *in vitro*, and DF = -1.2546 for PD138362, which is inactive (see Fig. (**8**)).



PD139586 **PD138362**

Figure 8: Structures of quinolones PD139586 and PD138362.

Later on LDA was applied to 24 commercial quinolones that had not been used to define the model and whose MICs were subsequently determined *in vitro*. From them, seven quinolones were classified as active, nine as inactive and eight as uncertain, what, as can be seen in Table 7, fits very well with the MIC experimental results. It is to emphasize the correct prediction of seven active quinolones which had low *in vitro* MICs, and specially the correct prediction of three of them (moxifloxacin, sparfloxacin and gatifloxacin), which exhibited MICs below 1 μg/mL (see Table 7).

Table 7: Comparison of predictions of activity by molecular topology and LDA analysis *versus* experimental MICs

LDA Analysis Results							
Quinolone	**M value**	**Class**	**MICexp**	**Quinolone**	**M value**	**Class**	**MICexp**
Moxifloxacin	3.9	Active	0.2	Pefloxacin	1.13	N.C.	10
Sparfloxacin	5.06	Active	0.4	Norfloxacin	-0.95	N.C.	11.4
Gatifloxacin	4.54	Active	0.9	Enoxacin	-1.78	Inactive	13.7
Temafloxacin	-0.21	N.C.	1	Acrosoxacin	-0.68	N.C.	23.5
Levofloxacin	0.77	N.C.	2.1	Rufloxacin	0.51	N.C.	31
Ofloxacin	0.77	N.C.	2.5	Irloxacin	-3.63	Inactive	47.2
Trovafloxacin	1.69	Active	2.7	Pipemidic acid	-1.97	Inactive	>250
Ciprofloxacin	7.2	Active	2.8	Flumequine	-1.96	Inactive	>250
Lomefloxacin	-1.28	Inactive	4.5	Piromidic acid	-4.43	Inactive	>250
Clinafloxacin	2.16	Active	5	Nalidixic acid	-3.65	Inactive	>250
Grepafloxacin	3.85	Active	5.4	Cinoxacin	-2.24	Inactive	>250
Fleroxacin	-0.18	N.C.	8.1	Oxolinic acid	-3.25	Inactive	>250

Pharmacological-Activity Distribution Diagrams (PDDs)

A very practical tool for better visualizing the discrimination between the active and inactive compounds is the pharmacological distribution diagram (PDDs) [52]. These diagrams are histogram-like plots in which the compounds are grouped into intervals of the predicted value of the property under analysis (P). The diagrams are arranged so that the number of compounds in each interval of P is determined for each group. The Expectancy (E) of finding a molecule with a desired value of P is obtained so that for each arbitrary interval of whatever function, it is defined an expectancy of activity as: $Ea = a / (i + 1)$, where a is the quotient between the number of active compounds in this interval and the overall number of active compounds; likewise, i represents the ratio of inactive compounds. The expectancy of inactivity is then obtained as: $Ei = i / (a + 1)$. For a given equation, it is straightforward to see the zones in which the overlapping between *Ea* and *Ei* is minimal, and thereby deciding if the equation studied can be useful or not for the selection and molecular design. This also permits to determine the intervals of the property where the probability of finding new active compounds is maximal and those regions in which the probability of inactivity is minimal.

Example: Topological virtual screening to find out new active compounds in ulcerative colitis by inhibiting NF-κB [53].

In this case study, MT was used to find out new compounds active in ulcerative colitis by inhibiting nuclear factor kappa beta (NF-κB), one of the standard mechanisms of action related to the disease. Different topological indices were used as structural descriptors, and their relation to biological activity was determined by using LDA.

A topological model consisting of two discriminant functions was built up. The first function was mechanistic, *i.e.* focused on the discrimination between NF-κB active and inactive compounds, and the second one was not mechanistic, *i.e.* just distinguishing between compounds active and inactive on ulcerative colitis in general.

The model was then applied sequentially to a large database of compounds with unknown activity. 28 of such compounds were predicted to be active and selected for *in vitro* and *in vivo* testing.

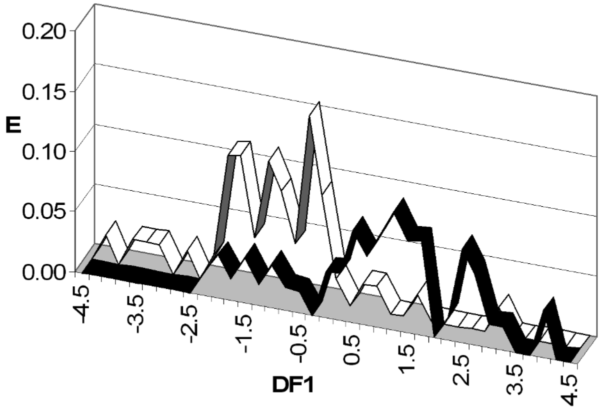
The first equation, corresponding to DF1, distinguished compounds that were predicted to have NF-κB inhibitory activity. The five-variable equation was:

$$DF1 = 0.633\ G_1 + 24.29\ {}^{0}C - 3.62\ {}^{2}D + 2.09\ {}^{4}Dp + 0.288\ V_3 - 30.94$$

$$N=95\ F= 9.5\ \lambda=0.354. \qquad (11)$$

where DF1 is the discriminant function, G1 is the first order TCI, C and D represent quotients and differences between connectivity indices and V3 the number of vertices with valence three.

According to Eq. 11, a compound is classified as active if DF1>0, otherwise it is considered inactive. By applying this criterion to the DF1 training set (95 compounds), 40 out of 51 experimentally active compounds were correctly classified as such (78% accuracy), whereas 40 out of 44 experimentally inactive compounds were also well classified (91% accuracy).



Figrue 9: Pharmacological distribution diagram for NF-*κ*B inhibitors obtained using the discriminant function DF1. (The *black* color represents the compounds with inhibitor activity and the *white* color, the compounds without it).

To establish the adequate range of activity, the PDD obtained with DF1 was built up. Observing Fig. (**9**), one can see that all compounds show DF1 values within the range 4.5>DF1>-3.8. Moreover, there is little overlapping between compounds with probability of activity above 40% and below 60%. These percentages match the DF1 values in Eq. 11, in the range 0.4>DF1>-0.4.

Therefore, a compound will be selected as NF-κB inhibitor if it stands in the range 4.5>DF1> 0.4 and as inactive if it is in the range -0.4>DF1>-3.8. Outside these intervals, the classification is uncertain and the compound is considered as "not-classified" (outlier, NC).

As pointed before, the second equation, DF2, was employed to discriminate compounds showing a general profile of activity in ulcerative colitis. The selected four-variable equation was:

$$DF2 = 3.15\ SEige - 10.28\ GATS6p - 261.66\ X3A - 0.06\ D/Dr05 + 52.73$$

$$N=31\ F=13.3\ \lambda= 0.328. \qquad (12)$$

where SEige is defined as eigenvalue sum from the electronegativity weighted distance matrix; GATS6p as Geary autocorrelation - lag 6/weighted by atomic polarizabilities; X3A as the average connectivity index chi-3 and finally, D/Dr05 as the distance/detour ring index of order 5.

The PDD for this equation (Fig. (**10**)), shows that all the compounds present DF2 values in the range 8>DF2>-10. Hence, a compound is selected as active if it lies in the range: 8>DF2> 0.89 and as inactive if the range is -0.49>DF2>-10. Outside these ranges any compound is taken as uncertain (NC).

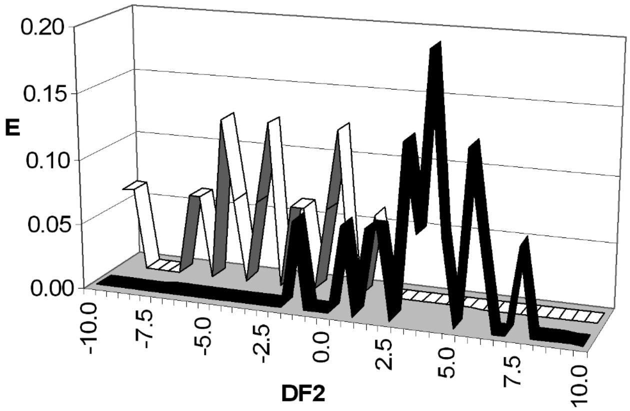


Figure 10: Pharmacological distribution diagram for ulcerative colitis active drugs obtained using the discriminant function DF2. (*Black* color represents the compounds with anti-ulcerative colitis activity and the *white* color, the compounds without it).

Based on the models described above, a virtual screening study was carried out on a database of heterogeneous drug molecules. A library (MicroSource Pure Natural Products Collection) consisting of 800 natural products and the Merck index database (about 12,000 compounds) were screened for that purpose. The composition of the library is at the MicroSource Discovery Systems website (http://www.msdiscovery.com).

Based upon these models, it was expected that some 28 compounds might be active against UC by NF-κB inhibition. Almost all of them were commercially available and hence were selected for future *in vitro* and *in vivo* tests which would strengthen the model's predictive capability. Table **8** illustrates the DF1 and DF2 values, as well as the classification for each compound from the PDDs. Five compounds (cromolyn sodium, rotenone, 10-hydroxycamptothecin, methylorsellinate and 2-methoxyresorcinol) were classified as active by DF2 but not by DF1; two compounds (rosmarinic acid and Ro 41-0960) were classified as active by DF1 but inactive by DF2 and two more (calcein and (+)-dibenzyl L-tartrate) were classified as active by DF1 but as outliers by DF2. The compounds selected were those passing at least one of the filters, either the NF-κB inhibition or the general profile, because both represent a good choice.

As shown in Table **8**, most of the compounds selected had been described previously as anti-inflammatory in the literature (10-hydroxycamptothecin, purpurin, physcion, methyl orsenillate, aconitic acid, genkwanin, uvaol, cromolyn sodium, hesperidin and rosmarinic acid) and one of them, namely ursolic acid, is described to show also anti-ulcerative properties.

Those not previously reported as active –or at least not found as such- were selected for testing. Our recent results, (work in press) confirm that several of the selected compounds were significantly active at *in vivo* and/or *in vitro* tests related UC.

The results described here for ulcerative colitis demonstrate, once more, that MT is an excellent way to search for new drugs.

Table 8: Values of DF1, DF2, probability of activity and classification of the potential anti UC compounds, selected from the Merck Index database and the MicroSource Pure Natural Products Collection. Compounds' therapeutic profile from the literature are also included

Compound	DF1	Prob (activ.) %	Class.	DF2	Prob (activ.) %	Class.	Activity/Therapeutics category
Aconitic acid	2.90	95	A	2.97	95	A	Anti-inflammatory
Ajmalinediacetate	1.16	77	A	4.86	99	A	-
Alizarin -3-methylininodiacetic acid	1.86	87	A	4.33	99	A	Red staining
Apigenin	1.64	84	A	2.6	94	A	Anti-inflammatory
Calcein	2.21	90	A	13.9	100	NC	Fluorescent dye
Carapin	2.51	93	A	3.04	96	A	-
Cromolyn sodium	0.33	58	NC	1.19	78	A	Anti-inflammatory
(+)-Dibenzyl L-tartrate	0.62	65	A	-12.7	0	NC	-
Emodic acid	2.69	94	A	4.44	99	A	-
Evernic acid	1.42	81	A	7.09	99	A	-
Fissinolide	2.38	92	A	4.39	99	A	-
Folicacid	1.65	84	A	3.9	98	A	-
Genkwanin	0.57	64	A	2.59	93	A	Anti-inflammatory
Haematommic acid	2.22	90	A	5.62	100	A	Antioxidant
Hesperidin	2.20	90	A	6.08	100	A	Anti-inflammatory
10-Hydroxycamptothecin	0.37	59	NC	2.57	93	A	Anti-inflammatory
Lonchocarpic acid	1.06	75	A	2.49	93	A	Antimicrobial
3-Methylorsellinic acid	0.99	73	A	6.65	100	A	-
Methylorsellinate	0.14	54	NC	6.38	100	A	Anti-inflammatory
2-Methoxyresorcinol	0.31	58	NC	1.94	88	A	-
Physcion	1.17	76	A	2.91	95	A	-
Purpurin	1.79	86	A	1.59	84	A	Anti-inflammatory
Pyrocatechuic acid	2.21	90	A	3.24	97	A	-
Ro 41-0960	1.75	85	A	-0.46	40	I	-
Rosmarinic acid	1.99	88	A	-8.03	0	I	Anti-inflammatory
Rotenone	0.26	57	NC	1.53	83	A	-
Ursolic acid	2.66	94	A	3.33	97	A	Anti-ulcer
Uvaol	2.98	95	A	3	96	A	Anti-inflammatory

Molecular Selection and Drug Design

Our strategy for searching new drugs consists of different approaches:

Molecular Selection by Virtual Screening on Databases

A mathematical model, constituted by one or more equations with their corresponding thresholds and intervals of effectiveness is used to screen a structural database, and the selected structures are searched in the literature to check their predicted activity. The compounds described to be active, stand for the model validation as a proof of concept. The compounds not reported as active, are proposed for lab assays. Compounds selected as active but not showing activity *in vitro*, *i.e.* false positives, as well as those actually active, *i.e.* true positives, are used to refine the model. For more details see Refs. [54, 55].

Virtual Combinatorial Syntheses and Computational Screening

In this case, the model is used to track a virtual library consisting of molecular structures resulting from combinatorial chemistry, so that the structures selected are synthesized and tested. For specific details consult references [56, 57].

Molecular Design of New Structures

In 1985 and 1988 our group presented two doctoral thesis dealing on the use of topological descriptors in drug design/discovery [58]. The results were published in a follow-up paper [59]. The principal idea therein was the possibility to use topological indices in a reverse way as compared to the conventional: *i.e.* obtaining “tailor-made” molecular structures from topological indices. This goal was supported by the fact that topological indices are not simply structure-related descriptors, but they are rather a pure algebraic description of the structure itself. The method enables for molecular construction from the scratch or, alternatively, the use of a scaffold (referred to as *base structure*) to which the carbon-carbon substructures and functional groups, can be attached.

The substructural fragments were acyclic and their bond orders were between one and three. The fragments and functional groups were computationally assembled to the base structure on the previously defined attachment sites. These could be attached by each one of their available atoms, in such a way that the formation of multiple bonds and cyclic structures was possible. For each new compound designed, the models’ outcome decided whether it was potentially active or not. The models were arranged according to a previous QSAR study based on Randić-Kier-Hall type indices. For more detail see Ref. [60].

Table **9** displays the results of the search of new biological/pharmacological activities for different compounds, most of them can be considered as new *hits* or *leads*.

Table 9: New biological activities discovered through virtual screening. For details see the references in the last column

Found Activity	Selected Drugs	Refs.
Cytostatic	6-azuridine, quinine	[61]
Antibacterial	1-Chloro-2,4-dinitrobenzene, 3-Chloro-5-nitroindazole, 1-Phenyl-3-methyl-2-pyrazolin-5-one, neohesperidin, amaranth, mordant brown 24, hesperidin, morine, niflumic acid, silymarine, fraxine	[62]
Antifungal	Neotetrazolium chloride, benzotropine mesilate, 3-(2-Bromethyl)-indole, 1-Chloro-2,4-dinitrobenzene	[63]
Hypoglycaemic	3-Hydroxybutyl acetate 4-(3-Methyl-5-oxo-2-pyrazolin-1-yl) benzoicacid 1-(Mesitylene-2-sulfonyl) 1H-1,2,3-triazole	[64]
Antivirals (anti-Herpes)	3,5-dimethyl-4-nitroisoxazole, nitrofurantoin, 1-(pyrrolidinocarbonylmethyl)piperazine, nebularine, cordycepin, adipicacid, thymidine, α□thymidine, inosine, 2,4-diamino-6-(hydroxymethyl)pteridine, 7-(carboxymethoxy)-4-methylcoumarin, 5-methylcytidine	[56]
Antineoplastic	Carminic acid, tetracycline, piromidic acid, doxycycline	[65]
Antimalarial	Monensin, nigericin, vinblastine, vincristine, vindesine, ethylhydrocupreine, quinacrine, salinomycin	[66, 67]
Antitoxoplasma	Cefamandolenafate Prazosin Andrographolide Dibenzothiophenesulfone 2-Acetamido-4-methyl-5 thiazolesulfonylchloride	[68]
Antihystaminic	Benzydamine 4-(1-Butylpentyl)pyridine N-(3-Bromopropyl)phtalimide N-(3-Chloropropyl)phtalimide N-(3-Chloropropyl)piperidine hydrochloride 5-Bromoindole	[69]
Bronchodilator	Griseofulvin, anthrarobin, 9,10-Dihydro-2-methyl-4H-benzo [5,6]cyclohept[1,2-*d*] oxazol-4-ol, 2-Aminothiazole, Maltol, esculetin, fisetin, hesperetin, 4-methyl-umbellipheryl-4-guanidine benzoate	[70]
Analgesics	2-(1-propenyl)phenol, 2',4' dimethylacetophenone, p- chlorobenzohydrazide, 1-(p-chlorophenyl) propanol, 4-benzoyl-3-methyl-1-phenyl-2-pyrazolin-5-one	[60,71]
NSAIDs	1,3-bis(benzyloxycarbonyl)-2-methyl-2-thiopseudourea, 4,6-dichloro-2-methylthio-5-phenylpyrimidine, 2-chloro-2',6'-acetoxylidide, trans-1,3-diphenyl-2-propen-1-ol	[55]

Particularly relevant is the discovery of a novel anticancer lead compound, namely MT477 (see Fig. (**11**)), which showed very potent activity *in vitro* and *in vivo* against human cell carcinoma [72]. Moreover, MT477 is a novel thiopyrano [2,3-c] quinoline with a high activity against protein kinase C (PKC) isoforms. MT477 interfered with PKC activity as well as phosphorylation of Ras and ERK1/2 in H226 human lung carcinoma cells. It also induced poly-caspase-dependent apoptosis.

Another antineoplastic compound obtained by molecular topology was MT103, (Fig. **11**), which is an isoborneol derivative, with a promising profile predicted to slow tumor growth through pro-apoptotic signaling and protein kinase C inhibition. It was found that MT103 inhibited the growth of a wide variety of cancer cell types as verified by the NCI-60 cancer cell line panel. MTT cell viability assay showed that MT103 inhibited 50% of the growth of HOP-92, ACHN, NCI-H226, MCF-7, and A549 cancer cell lines at much lower concentrations than that required for HUVECs and human fibroblasts [73].

In the field of malaria our research group have found two ionophores (monensin and nigericin) that inhibited completely the parasite development at the liver stage. The liver stage of *Plasmodium* is a very interesting drug target because it precedes the emergence of blood stages that cause the symptoms and complications of malaria. Drugs that inhibit parasite maturation within hepatocytes could be used for short-term prophylaxis in areas of endemicity (refugees, travellers, *etc.*). For more details see the ref. [67].

About the Alzheimer's disease, MT has also demonstrated to be very efficient in the search of novel drug treatments. In a study carried out by Medisyn Technologies in collaboration with Mount Sinai School of Medicine, eight compounds were patented as very efficient anti-beta amyloid and as anti-oligomeric [74]. The chemical structures of some of these compounds are shown in Fig. **11**.

CONCLUSION

The results outlined in this review, clearly demonstrate that the QSAR approach based on molecular topology is a powerful tool for the prediction of properties and the design and selection of new drugs. Moreover, MT is based strictly on a mathematical layout of molecular structure, thereby bypassing any geometrical or physical profile, what is also an important asset of the approach and makes of it a new paradigm. Another important advantage of MT is that, contrary to most drug design methods, it does not need a previous knowledge of the mechanism of action of the target drugs, which, considering the extremely huge number of possible structures potentially active as drugs, is a very important asset.

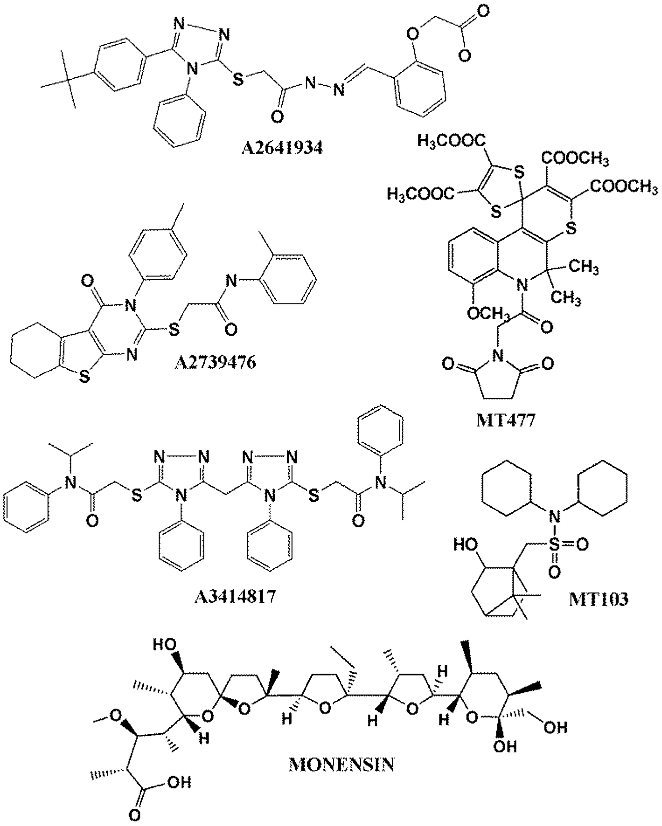


Figure 11: Chemical structure of anticancer, antimalarial and anti-Alzheimer drugs designed by MT.

The reason why MT works so well is unknown up to date; it remains as an open question and it is probably a good challenge to take over in the future.

ACKNOWLEDGEMENTS

We thank the Ministerio de Educación e Innovación, Spain (project SAF2009-13059-C03-02) for support of this work.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

ABBREVIATIONS

CoMFA	=	Comparative molecular field analysis
CoMSIA	=	Comparative molecular similarity indices analysis
CV	=	Cross-validation
E	=	Expectancy
HIV	=	Human immunodeficiency virus
JHA	=	Juvenile hormone active
LOO	=	Leave-one-out
LSO	=	Leave-some-out
LDA	=	Linear discriminant analysis
MBC	=	Minimum blocking concentration
MIC	=	Minimum inhibitory concentration
MT	=	Molecular topology
MAO	=	Monoamine oxidase
MLRA	=	Multilinear regression analysis
MAV	=	Mycobacterium avium
NC	=	not-classified
NF-κB	=	Nuclear factor kappa beta
PIG	=	Percentage inhibition degree
PDD	=	Pharmacological distribution diagram
PPB	=	Plasma protein binding

PKC	=	Protein kinase C
QSAR	=	Quantitative structure-activity relationships
QSPR	=	Quantitative structure–property relationships
QC	=	Quantum chemical
SomFA	=	Self Organizing Molecular Field Analysis
SEE	=	Standard error of estimates
TCI	=	Topological Charge Indices

REFERENCES

[1] Hansch, C.; Fujita, T. p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc*., **1964**, 86, 1616-1626.

[2] Cramer III, R. D.; Patterson, 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*., **1988**, 110, 5959-5967.

[3] Guner, O. History and evolution of the pharmacophore concept in computer-aided drug design. *Curr. Top. Med. Chem.,* **2002**, 2, 1321-1332.

[4] Klebe, G.; Abraham, U.; Mietzner, T. Molecular similarity indices in a comparative analysis [CoMSIA] of drug molecules to correlate and predict their biological activity. *J. Med. Chem*., **1994**, 37, 4130-4146.

[5] Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-organizing molecular field analysis: A tool for structure-activity studies. *J. Med. Chem*., **1999**, 42, 573-583.

[6] Kier, L. B.; Hall, L. H. In Molecular connectivity in chemistry and drug research; Academic Press New York: **1976**; Vol. 2.

[7] Devillers, J. New trends in QSAR modeling with topological indices. *Curr. Opin. Drug. Discov. Devel*., **2000**, 3, 275-279.

[8] Diudea, M.; Florescu, M.; Khadikar, P. Molecular Topology and Its Applications, EFICON, Bucharest. **2006**.

[9] Pogliani, L. From molecular connectivity indices to semiempirical connectivity terms: Recent trends in graph theoretical descriptors. *Chem. Rev*., **2000**, 100, 3827-3858.

[10] Ivanciuc, O.; Ivanciuc, T.; Balaban, A. T. Quantitative structure-property relationship study of normal boiling points for halogen-/oxygen-/sulfur-containing organic compounds using the CODESSA program. *Tetrahedron*, **1998**, 54, 9129-9142.

[11] Hosoya, H.; Gotoh, M.; Murakami, M.; Ikeda, S. Topological Index and Thermodynamic Properties. 5. How Can We Explain the Topological Dependency of Thermodynamic Properties of Alkanes with the Topology of Graphs? *J. Chem. Inf. Comput. Sci*., **1999**, 39, 192-196.

[12] García-Domenech, R.; Gálvez, J.; de Julián-Ortiz, J. V.; Pogliani, L. Some new trends in chemical graph theory. *Chem. Rev*., **2008**, 108, 1127-1169.

[13] Basak, S. C.; Mills, D. R.; Gute, B. D.; Natarajan, R. Predicting pharmacological and toxicological activity of heterocyclic compounds using QSAR and molecular modeling. QSAR and Molecular Modeling Studies in Heterocyclic Drugs I, **2006**, 39-80.

[14] Balaban, A. T.; Motoc, I.; Bonchev, D.; Mekenyan, O. Topological indices for structure-activity correlations. *Top. Curr. Chem*., **1983**, 114, 21-55.

[15] Estrada, E.; Uriarte, E. Recent advances on the role of topological indices in drug discovery research. *Curr. Med. Chem*., **2001**, 8, 1573-1588.

[16] Marrero Ponce, Y. Total and local (atom and atom type) molecular quadratic indices: significance interpretation, comparison to other molecular descriptors, and QSPR/QSAR applications. *Bioorg. Med. Chem*., **2004**, 12, 6351-6369.

[17] Dudek, A. Z.; Arodz, T.; Gálvez, J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb. Chem. High. T. Scr*., **2006**, 9, 213-228.

[18] Gonzales-Diaz, H.; Gia, O.; Uriarte, E.; Hernadez, I.; Ramos, R.; Chaviano, M.; Seijo, S.; Castillo, J. A.; Morales, L.; Santana, L. Markovian chemicals" *in silico*" design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds. *J. Mol. Mod*., **2003**, 9, 395-407.

[19] Oloff, S.; Mailman, R. B.; Tropsha, A. Application of validated QSAR models of D1 dopaminergic antagonists for database mining. *J. Med. Chem*., **2005**, 48, 7322-7332.

[20] Estrada, E.; Peña, A. *In silico* studies for the rational discovery of anticonvulsant compounds. *Bioorg. Med. Chem*., **2000**, 8, 2755-2770.

[21] Estrada, E.; Vilar, S.; Uriarte, E.; Gutierrez, Y. *In silico* studies toward the discovery of new anti-HIV nucleoside compounds with the use of TOPS-MODE and 2D/3D connectivity indices. 1. Pyrimidyl derivatives. *J. Chem. Inf. Comput. Sci*., **2002**, 42, 1194-1203.

[22] Le-Thi-Thu, H.; Casañola-Martín, G. M.; Marrero-Ponce, Y.; Rescigno, A.; Saso, L.; Parmar, V. S.; Torrens, F.; Abad, C. Novel coumarin-based tyrosinase inhibitors discovered by OECD principles-validated QSAR approach from an enlarged, balanced database. *Mol. Divers.,* **2011**, 1-14.

[23] Kumar, V.; Bansal, H. QSAR studies on estimation of monoamine oxidase-A inhibitory activity using topological descriptors. *Med. Chem. Res*., **2011**, 20, 168-174.

[24] Basak, S. C.; Mills, D. R.; Hawkins, D.; Bhattacharjee, A. Quantitative structure–activity relationship studies of antimalarial compounds from their calculated mathematical descriptors. SAR QSAR *Environ. Res*., **2010**, 21, 103-125.

[25] 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 compounds. *Nat. Biotechnol*., **1998**, 16, 748-752.

[26] Natarajan, R.; Basak, S. C.; Mills, D. R.; Kraker, J. J.; Hawkins, D. M. Quantitative structure-activity relationship modeling of mosquito repellents using calculated descriptors. *Croat. Chem. Acta*, **2008**, 81, 333–340.

[27] Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: a hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci*., **2001**, 41, 671-678.

[28] Gálvez, J.; Garcia, R.; Salabert, M. T.; Soler, R. Charge Indexes. New Topological Descriptors. *J. Chem. Inf. Comput. Sci*., **1994**, 34, 520-525.

[29] Gálvez, J.; Garcia-Domenech, R.; de Julián-Ortiz, J.; Soler, R. Topological approach to drug design. *J. Chem. Inf. Comput. Sci*., **1995**, 35, 272-284.

[30] García-Domenech, R.; Gálvez, J.; Moliner, R.; García-March, F. Prediction and interpretation of some pharmacological properties of cephalosporins using molecular connectivity. *Drug Invest*., **1991**, 3, 344–350.

[31] Randić, M. Characterization of molecular branching. *J. Am. Chem. Soc*., **1975**, 97, 6609-6615.

[32] Kier, L. B.; Murray, W. J.; Randić, M.; Hall, L. H. Molecular connectivity V: connectivity series concept applied to density. *J. Pharm. Sci*., **1976**, 65, 1226-1230.

[33] Kier, L. B.; Hall, L. H. Differential molecular connectivity in data-base fragment searching. *Pharm. Res*., **1989**, 6, 497-500.

[34] Kier, L. B.; Hall, L. H. In Molecular connectivity in structure-activity analysis; Research Studies Press Letchworth, Hertfordshire, UK: **1986**; Vol. 9.

[35] Kier, L. B. Molecular structure influencing either a sweet or bitter taste among aldoximes. *J. Pharm. Sci.,* **1980**, 69, 416-419.

[36] Ponce, A.; Blanco, S.; Molina, A.; García-Domenech, R.; Gálvez, J. Study of the action of flavonoids on xanthine-oxidase by molecular topology. *J. Chem. Inf. Comput. Sci*., **2000**, 40, 1039-1045.

[37] Wiener, H. Structural determination of paraffin boiling points. *J. Am. Chem. Soc*., **1947**, 69, 17-20.

[38] Kier, L. B. Shape indexes of orders one and three from molecular graphs. *QSAR*, **1986**, 5, 1-7.

[39] Kier LB. An index of molecular flexibility from kappa shape attributes. *QSAR,* **1989**; 8, 221-4.

[40] Balaban, A. T. Highly discriminating distance-based topological index. *Chem. Phys*. *Lett*., **1982**, 89, 399-404.

[41] Kier, L. B.; Hall, L. H. An electrotopological-state index for atoms in molecules. *Pharm. Res*., **1990**, 7, 801-807.

[42] Estrada, E. Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. *J. Chem. Inf. Comput. Sci*., **1996**, 36, 844-849.

[43] EduSoft, L. MolconnZ version 4. 05, **2003**.

[44] Todeschini, R.; Consonni, V. DRAGON software (version 1. 11-2001). Milano, Italy **2003**.

[45] Basak, S.; Harriss, D.; Magnuson, V. POLLY 2. 3. Copyright of the University of Minnesota, **1988**.

[46] Katritzky, A.; Lobanov, V.; Karelson, M. CODESSA software. University of Florida, SemiChem, Shawnee, KS, **1994**.

[47] Roy, P. P.; Leonard, J. T.; Roy, K. Exploring the impact of size of training sets for the development of predictive QSAR models. Chemometrics Intellig. *Lab. Syst*., **2008**, 90, 31-42.

[48] Besalú, E. Fast computation of cross-validated properties in full linear leave-many-out procedures. *J. Math. Chem*., **2001**, 29, 191-204.

[49] Wold, S.; Eriksson, L.; Clementi, S. Statistical validation of QSAR results. Chemometric methods in molecular design, **1995**, 309-338.

[50] Gálvez, J.; de Julian-Ortiz, J.; Garcia-Domenech, R. Application of molecular topology to the prediction of potency and selection of novel insecticides active against malaria vectors. *Journal of Molecular Structure: THEOCHEM*, **2005**, 727, 107-113.

[51] Gozalbes, R.; Brun-Pascaud, M.; García-Domenech, R.; Gálvez, J.; Girard, P. M.; Doucet, J. P.; Derouin, F. Prediction of quinolone activity against Mycobacterium avium by molecular topology and virtual computational screening. *Antimicrob. Agents Chemother*., **2000**, 44, 2764-2770.

[52] Gálvez, J.; García-Domenech, R.; de GregorioAlapont, C.; de Julián-Ortiz, J.; Popa, L. Pharmacologicaldistributiondiagrams: a tool for *de novo* drug design. *J. Mol. Graph.,* **1996**, 14, 272-276.

[53] Gálvez-Llompart, M.; Recio, M. C.; García-Domenech, R. Topological virtual screening: a way to find new compounds active in ulcerative colitis by inhibiting NF-κB. *Mol. Divers*., **2011**, 1-10.

[54] Bruno-Blanch, L.; Gálvez, J.; Garcia-Domenech, R. Topological virtual screening: a way to find new anticonvulsant drugs from chemical diversity. *Bioorg. Med. Chem. Lett*., **2003**, 13, 2749-2754.

[55] Gálvez-Llompart, M.; Giner, M.; Recio, C.; Candeletti, S.; Garcia-Domenech, R. Application of molecular topology to the search of novel NSAIDs: Experimental validation of activity. *Lett. Drug Des. Discov.,* **2010**, 7, 438-445.

[56] de Julián-Ortiz, J. V.; Gálvez, J.; Munoz-Collado, C.; Garcia-Domenech, R.; Gimeno-Cardona, C. Virtual Combinatorial Syntheses and Computational Screening of New Potential Anti-Herpes Compounds 1. *J. Med. Chem.,* **1999**, 42, 3308-3314.

[57] Duart, M. J.; Antón-Fos, G. M.; Alemán, P. A.; Gay-Roig, J. B.; González-Rosende, M. E.; Gálvez, J.; García-Domenech, R. New potential antihistaminic compounds. Virtual combinatorial chemistry, computational screening, real synthesis, and pharmacological evaluation. *J. Med. Chem*., **2005**, 48, 1260-1264.

[58] Arviza, M. Predicción e interpretación de algunas propiedades fisicoquímicas y biológicas de un grupo de barbitśricos y sufonamidas por el método de conectividad molecular. PhdThesis, Universitat de Valencia, Spain, **1985**.

[59] Gálvez, J.; García-Domenech, R.; Bernal, J.; García-March, F. Desarrollo de un nuevo método de diseño de fármacos por topología molecular. Su aplicación a analgésicos no narcóticos. *An. Real. Acad. Farm*., **1991**, 57, 533-546.

[60] García-Domenech, R.; García-March, F.; Soler, R.; Gálvez, J.; Antón-Fos, G.; De Julián-Ortiz, *J. New analgesicsdesignedby molecular topology. QSAR*, **1996**, 15, 201-207.

[61] Gálvez, J.; Garcia-Domenech, R.; Gomez-Lechon, M.; Castell, J. Use of molecular topology in the selection of new cytostatic drugs. *J. Mol. Struct. : THEOCHEM*, **2000**, 504, 241-248.

[62] de Gregorio Alapont, C.; Garcia-Domenech, R.; Gálvez, J.; Ros, M.; Wolski, S.; García, M. Molecular topology: a useful tool for the search of new antibacterials. *Bioorg. Med. Chem. Lett*., **2000**, 10, 2033-2036.

[63] Pastor, L.; García-Domenech, R. New antifungals selected by molecular topology. *Bioorg. Med. Chem. Lett*., **1998**, 8, 2577-2582.

[64] Antón-Fos, G.; García-Domenech, R.; Perez-Gimenez, F.; Peris-Ribera, J.; García-March, F.; Salabert-Salvador, M. Pharmacological Studies of the Two New Hypoglycaemic Compounds 4-(3-Methyl-5-oxo-2-pyrazolin-1-yl) benzoic Acid and 1-(Mesitylen-2-sulfonyl)-1H-1, 2, 4-triazole. *Arzneim*., **1994**, 44, 821-826.

[65] Gálvez, J.; Gomez-Lechón, M.; García-Domenech, R.; Castell, J. New cytostatic agents obtained by molecular topology. *Bioorg. Med. Chem. Lett*., **1996**, 6, 2301-2306.

[66] Mahmoudi, N.; de Julián-Ortiz, J. V.; Ciceron, L.; Gálvez, J.; Mazier, D.; Danis, M.; Derouin, F.; García-Domenech, R. Identification of new antimalarial drugs by linear discriminant analysis and topological virtual screening. *J. Antimicrob. Chemother*., **2006**, 57, 489.

[67] Mahmoudi, N.; Garcia-Domenech, R.; Gálvez, J.; Farhati, K.; Franetich, J. F.; Sauerwein, R.; Hannoun, L.; Derouin, F.; Danis, M.; Mazier, D. New active drugs against liver stages of Plasmodium predicted by molecular topology. *Antimicrob. Agents Chemother.,* **2008**, 52, 1215.

[68] Gozalbes, R.; Gálvez, J.; Garcia-Domenech, R.; Derouin, F. Molecular search of new active drugs against Toxoplasma gondii. *SAR QSAR Environ. Res.,* **1999**, 10, 47-60.

[69] Casabán-Ros, E.; Antón-Fos, G.; Gálvez, J.; Duart, M.; García-Doménech, R. Search for new antihistaminic compounds by molecular connectivity. *QSAR*, **1999**, 18, 35-42.

[70] Rios-Santamarina, I.; Garcia-Domenech, R.; Gálvez, J. New bronchodilators selected by molecular topology. *Bioorg. Med. Chem. Lett*., **1998**, 8, 477-482.

[71] Gálvez, J.; García-Domenech, R.; De Julian-Ortiz, V.; Soler, R. Topological approach to analgesia. *J. Chem. Inf. Comput. Sci*., **1994**, 34, 1198-1203.

[72] Jasinski, P.; Welsh, B.; Gálvez, J.; Land, D.; Zwolak, P.; Ghandi, L.; Terai, K.; Dudek, A. Z. A novel quinoline, MT477: suppresses cell signaling through Ras molecular pathway, inhibits PKC activity, and demonstrates *in vivo* anti-tumor activity against human carcinoma cell lines. *Invest. New Drugs*, **2008**, 26, 223-232.

[73] Jasinski, P.; Zwolak, P.; Isaksson Vogel, R.; Bodempudi, V.; Terai, K.; Gálvez, J.; Land, D.; Dudek, A. Z. MT103 inhibits tumor growth with minimal toxicity in murine model of lung carcinoma *via* induction of apoptosis. *Invest. New Drugs*, **2011**, 29, 1-7.

[74] Gálvez, J.; Llompart, J.; Land, D.; Pasinetti, G. Patent Application Country: Application: WO; WO; Priority Application Country: US Patent WO2010114636, **2010**.

CHAPTER 9

Conceptual Density Functional Theory of Chemical Reactivity

Pratim K. Chattaraj[1,*] **and Debesh R. Roy**[1,2]

[1]Department of Chemistry and Center for Theoretical Studies, Indian Institute of Technology, Kharagpur 721302, India and [2]Department of Applied Physics, S. V. National Institute of Technology, Surat 395007, India

Abstract: A rudimentary treatment of density functional theory (DFT) is presented in this article. Various global and local reactivity descriptors are defined within the broad framework of conceptual DFT. A theory of chemical reactivity is developed in terms of these descriptors and the associated electronic structure principles.

Keywords: Density Functional Theory (DFT), chemical reactivity, electronegativity, chemical potential, chemical hardness, chemical softness, polarizability, electrophilicity index, Fukui function, local softness, local hardness, philicity, electronegativity equalization principle, HSAB principle, maximum hardness principle, minimum polarizability principle, minimum electrophilicity principle, minimum magnetizability principle, electrophilicity equalization principle, Quantum Fluid Dynamics (QFD), Time Dependent Density Functional Theory (TDDFT), Quantum Fluid Density Functional Theory (QFDFT).

INTRODUCTION

In classical mechanics all the properties are functions of 3N coordinates (q_i, i=1–3N) and 3N canonically conjugate momenta (p_i, i=1–3N), for an N – particle system. A knowledge of the $\{q_i\}$ and $\{p_i\}$ at t=0 will provide the same for later time by solving a classical equation of motion, as an initial value problem and hence the future behavior of the system can be analyzed. In the corresponding quantum version the properties are obtained as the expectation values of the associated linear hermitian operators over the wave functions, $\Psi(q_1, q_2, ..., q_{3N})$.

[*] **Corresponding author Pratim K. Chattaraj:** Department of Chemistry, Indian Institute of Technology, Kharagpur – 721 302, India; Tel: +91-3222-283304; Fax: +91-3222-255303;
E-mail: pkc@chem.iitkgp.ernet.in

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50009-6

These wave functions are well-behaved functions of $\{q_i\}$ and are obtained through the solution of the pertinent Schrödinger equation as a boundary value problem.

Problems associated with the solution of the Schrödinger equation and the interpretation of the wave functions of the many-particle systems prompted researchers to explore classical interpretation of quantum mechanics since we live and perceive in a 3-D classical world. The most famous approach in this direction is the density functional theory (DFT) [1-4].

For an N-electron system the density $\rho(\vec{r})$ is defined as

$$\rho(\vec{r}) = N\int \Psi^*(q_{1,}q_2...,q_{3N})\Psi(q_{1,}q_2...,q_{3N})\,dq_4....dq_{3N} \tag{1}$$

This probability density function is connected to the ordinary electron density function measured by crystallographers. Being a 3-D quantity even for an N-particle system, it allows us to visualize it as well as various models developed using it.

A functional is a correspondence which assigns a definite (real) number to each function (or curve) belonging to some class (*i.e.* a function of a function). Let $F(x, y(x), y'(x))$ be a continuous function. Then, the expression

$$J[y] = \int_a^b F(x, y(x), y'(x))dx \tag{2}$$

where y(x) ranges over the set of all continuously differentiable functions defined, on the interval [a,b], defines a functional $J[y]$.

Let us consider J[y] as a functional with the form $\int_a^b F(x, y, y')dx$ and is defined on the set of functions $y(x)$ whose first derivatives are continuous in [a, b] and satisfy the boundary conditions y(a)=A, y(b)=B. Then $J[y]$ will have an extremum for a given function $y(x)$, with a necessary condition that $y(x)$ satisfies Euler's equation:

$$\frac{\delta J[y]}{\delta y} = F_y - \frac{d}{dx}F_{y'} = 0 \tag{3}$$

In DFT $\rho(\vec{r})$ is considered as the basic variable. For an N- electron system, external potential $v(\vec{r})$ completely fixes the Hamiltonian $\hat{H}$ and hence N and $v(\vec{r})$ determine all properties of the ground state.

Hohenberg and Kohn proved two theorems which show that $\rho(\vec{r})$ contains all information [1].

(1) The external potential $v(\vec{r})$ is determined, within a trivial additive constant, by the electron density $\rho(\vec{r})$ and also

$$\int \rho(\vec{r})d\vec{r} = N \tag{4a}$$

Therefore, $\rho(\vec{r})$ provides $\hat{H}$ and hence Ψ (ground state) which in turn gives all electronic properties.

Proof. Reductio ad absurdum

Let there are two external potentials $v(\vec{r})$ and $v'(\vec{r})$ differing by more than a constant, each giving the same $\rho(\vec{r})$ for its ground state. Then we would have two Hamiltonians $\hat{H}$ and whose ground state densities were the same although the normalized wavefunctions Ψ and Ψ' are different.

Take Ψ' to be a trial function for the $\hat{H}$ problem,

Then

$$E_0 \prec \left\langle \Psi' | \hat{H} | \Psi' \right\rangle$$

$$= \left\langle \Psi' | \hat{H}' | \Psi' \right\rangle + \left\langle \Psi' | \hat{H} - \hat{H}' | \Psi' \right\rangle = E_0' + \int \rho(\vec{r})[v(\vec{r}) - v'(\vec{r})]d\vec{r} \tag{4b}$$

where E_0 and $E_0^{'}$ are ground state energies for the H and H' problems respectively.

Now, take Ψ as a trial function for the $\hat{H}'$ problem,

Therefore,

$$E_0' \prec \left\langle \Psi \,|\, \hat{H}' \,|\, \Psi \right\rangle = \left\langle \Psi \,|\, \hat{H} \,|\, \Psi \right\rangle + \left\langle \Psi \,|\, \hat{H}' - \hat{H} \,|\, \Psi \right\rangle$$

$$= E_0 - \int \rho(\vec{r})[v(\vec{r}) - v'(\vec{r})]d\vec{r} \ldots \tag{4c}$$

Adding Eqs. (3) and (4) we have, $E_0 + E_0' \prec E_0' + E_0$

a contradiction, hence there cannot be two different v that give the same ρ for the ground state.

$\therefore$ ρ gives N and $v(\vec{r})$ and hence Ψ and all ground state properties.

Energy functional $E_v[\rho] = T[\rho] + V_{ne}[\rho] + V_{ee}[\rho]$

$$= \int \rho(\vec{r})v(\vec{r})d\vec{r} + F_{HK}[\rho]$$

where F_{HK}[ρ] is a universal functional of $\tilde{\rho}$ [as it does not depend on v]

$$F_{HK}[\rho] = T[\rho] + V_{ee}[\rho] = T[\rho] + \frac{1}{2}\int\int \frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r}d\vec{r}' + E_{XC}[\rho] \tag{4d}$$

(2) For a trial density $\tilde{\rho}(\vec{r})$ [$\tilde{\rho}(\vec{r}) \geq 0 \ \forall \vec{r}$ and $\int \tilde{\rho}(\vec{r})d\vec{r} = N$] $E_0 \leq E_v[\tilde{\rho}]$ from the usual variational principle.

where E_0 is the ground state energy for the respective H and $E_v[\tilde{\rho}]$ is the energy functional for the trial density $\tilde{\rho}$. Now, $\tilde{\rho}$ determines its own $\tilde{v}$, $\tilde{H}$, $\tilde{\Psi}$ (can be taken as trial function for the problem with external potential v).

$$\therefore \left\langle \tilde{\Psi} \,|\, \hat{H} \,|\, \tilde{\Psi} \right\rangle = \int \tilde{\rho}(\vec{r})v(\vec{r})d\vec{r} + F_{HK}[\tilde{\rho}] = E_v[\tilde{\rho}] \geq E_v[\rho] \tag{5}$$

where $F_{HK}[\tilde{\rho}]$ is a universal functional (known as the Hohenberg-Kohn functional) of $\tilde{\rho}$ (as it does not depend on $\tilde{v}$) and is given by

$$F_{HK}[\tilde{\rho}] = T[\tilde{\rho}] + V_{ee}[\tilde{\rho}] = T[\tilde{\rho}] + \frac{1}{2}\int\int \frac{\tilde{\rho}(\vec{r})\tilde{\rho}(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r}d\vec{r}' + E_{XC}[\tilde{\rho}] \tag{6}$$

Here, $T[\tilde{\rho}]$, $V_{ee}[\tilde{\rho}]$ and $E_{XC}[\tilde{\rho}]$ are kinetic energy functional, electron-electron interaction and exchange-correlation energy functional respectively.

Now, the electron density ρ must satisfy the stationary principle

$$\delta\{E_v[\rho]-\mu[\int\rho(\vec{r})d\vec{r}-N]\}=0 \tag{7}$$

to provide the Euler-Lagrange Equation (ELE):

$$\mu=\frac{\delta E_v[\rho]}{\delta\rho(\vec{r})}=v(\vec{r})+\frac{\delta F_{HK}[\rho]}{\delta\rho(\vec{r})} \tag{8}$$

where μ is the chemical potential.

The exact form for $F_{HK}[\rho]$ is not known. In the Kohn-Sham picture it is written as [2]

$$F_{HK}[\rho]=T_s[\rho]+J[\rho]+E_{XC}[\rho] \tag{9a}$$

where $T_s[\rho]$ and $J[\rho]$ are the kinetic energy functional of the reference system and classical Coulombic interaction energy respectively, given by

$$T_s[\rho]=\sum_i^N\left\langle\Psi_i\,|-\frac{1}{2}\nabla^2\,|\Psi_i\right\rangle \tag{9b}$$

$$J[\rho]=\frac{1}{2}\int\int\frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}d\vec{r}' \tag{9c}$$

The associated ELE is

$$\mu=v_{eff}+\frac{\delta T_s[\rho]}{\delta\rho} \tag{10}$$

where v_{eff} is the KS effective potential which is given by

$$v_{eff}(\vec{r})=v(\vec{r})+\frac{\delta J[\rho]}{\delta\rho}+\frac{\delta E_{XC}[\rho]}{\delta\rho}\quad=v(\vec{r})+\int\frac{\rho(\vec{r}')}{|\vec{r}-\vec{r}'|}d\vec{r}+v_{xc}(\vec{r}) \tag{11}$$

There are various approximations for the exchange correlation potential $v_{xc}(\vec{r})$, *e.g.*, local density approximation (*e.g.*, VWN [5] *etc.*) in which it v_{xc} depends only on the electron density ρ, generalized gradient approximation (*e.g.*, PW91 [6], PBE [7] *etc.*) where v_{xc} depends on both ρ and its gradient $\nabla\rho$. Various hybrid exchange correlation potentials are also developed, *e.g.*, B3LYP [8], PBE0 [9] *etc.* which are essentially a mixing of exchange and local/ non-local correlation functionals.

This ELE takes the following form (Kohn-Sham Equation):

$$[-\frac{1}{2}\nabla^2 + v_{eff}]\Psi_i = \varepsilon_i \Psi_i \tag{12}$$

where the density and the energy are given by

$$\rho(\vec{r}) = \sum_i^N \sum_s |\Psi_i(\vec{r}, s)|^2 \tag{13}$$

$$E = \sum_i^N \varepsilon_i - \frac{1}{2}\int\int \frac{\rho(\vec{r})\rho(\vec{r}')}{|\vec{r} - \vec{r}'|} d\vec{r} d\vec{r}' + E_{XC}[\rho] - \int v_{xc}(\vec{r})\rho(\vec{r}) d\vec{r} \tag{14}$$

Now let us consider the change in energy from one ground state to another. Then we have

$$E = E[N, v]$$

$$dE = \left(\frac{\partial E}{\partial N}\right)_v dN + \int \left[\frac{\delta E}{\delta v(\vec{r})}\right]_N dv(\vec{r}) d\vec{r}$$

$$= \mu\, dN + \int \rho(\vec{r}) dv(\vec{r}) d\vec{r} \tag{15}$$

$$E = E[\rho] \;\; dE = \int \left(\frac{\delta E}{\delta \rho(\vec{r})}\right)_v d\rho(\vec{r})\, d\vec{r} + \int \left[\frac{\delta E}{\delta v(\vec{r})}\right]_\rho dv(\vec{r}) d\vec{r} \tag{16}$$

$$\therefore \left[\frac{\delta E}{\delta \rho(\vec{r})}\right]_v = \mu = \text{Constant} \tag{17}$$

$$\left[\frac{\delta E}{\delta v(\vec{r})}\right]_\rho = \left[\frac{\delta E}{\delta v(\vec{r})}\right]_N = \rho(\vec{r}) \tag{18}$$

Similarly,

$$d\mu = \left(\frac{\partial \mu}{\partial N}\right)_v dN + \int \left(\frac{\delta \mu}{\delta v(\vec{r})}\right)_N dv(\vec{r})\,d\vec{r} \;= 2\eta\, dN + \int f(\vec{r})dv(\vec{r})d\vec{r} \tag{19}$$

$$f(\vec{r}) = \left(\frac{\partial \mu}{\partial v(\vec{r})}\right)_N = \left(\frac{\partial \rho(\vec{r})}{\partial N}\right)_v \Rightarrow \text{Maxwell's relation} \tag{20}$$

where η is the hardness and $f(\vec{r})$ is the Fukui function [2].

The equation (20) is similar to the Maxwell's relation and can be interpreted either as the change of the electron density $\rho(\vec{r})$ at each point $\vec{r}$ with changed N (total number of electrons) or as the sensitivity of chemical potential of a system to an external perturbation at a particular point $\vec{r}$.

Any flow of a substance takes place from the phase of higher μ to the phase of lower μ.

Here electron flows from B to A if $\mu_B^0 \succ \mu_A^0$. Therefore, the associated energies are

$$E_A = E_A^0 + \mu_A^0(N_A - N_A^0) + \eta_A(N_A - N_A^0)^2 + ... \tag{21}$$

$$E_B = E_B^0 + \mu_B^0(N_B - N_B^0) + \eta_B(N_B - N_B^0)^2 + ... \tag{22}$$

$$E_A + E_B = E_A^0 + E_B^0 + (\mu_A^0 - \mu_B^0)\Delta N + (\eta_A + \eta_B)\Delta N^2 \tag{23}$$

where $\Delta N = N_B^0 - N_B = N_A - N_A^0$

Upto the 1st order, if $\mu_B^0 \succ \mu_A^0$, a positive $\Delta N(B \xrightarrow{e^-} A)$ will stabilize the system.

Minimization of (E_A+E_B) with respect to ΔN provides $\mu_A = \mu_B$, where

$$\mu_A = \left(\frac{\partial E_A}{\partial N_A}\right)_v = \mu_A^0 + 2\eta_A \Delta N + ... \tag{24}$$

$$\mu_B = \left(\frac{\partial E_B}{\partial N_B}\right)_v = \mu_B^0 - 2\eta_B \Delta N + ... \tag{25}$$

and we have the expressions for ΔN and ΔE as follows:

$$\Delta N = \frac{\mu_B^0 - \mu_A^0}{2(\eta_A + \eta_B)} \tag{26}$$

$$\Delta E = -\frac{(\mu_B^0 - \mu_A^0)^2}{4(\eta_A + \eta_B)} \tag{27}$$

Conceptual density functional theory (DFT) [2-4] has been quite successful in providing theoretical bases for popular qualitative chemical concepts like electronegativity [10], hardness [11, 12] and electrophilicity [13, 14]. In terms of several global and local chemical reactivity and selectivity descriptors a complete theory of chemical reactivity has been envisaged. These descriptors and the associated electronic principles are presented below:

Global Reactivity Descriptors

These descriptors describe the reactivity of the molecule as a whole.

Electronegativity (χ) and Chemical Potential (μ)

In order to understand the nature of a chemical bond, Pauling introduced the concept of electronegativity [15] as, 'the power of an atom in a molecule to attract electrons to itself'. Based on thermodynamical data the calculated electronegativity values of atoms by Pauling follow the general intuition in chemistry.

Later on, the concept of 'absolute' electronegativity which is independent of molecular environment has come into picture as proposed by Mulliken [16]. This 'absolute' electronegativity of any atom or molecule can be expressed in terms of two experimentally measurable quantities, ionization potential (I) and electron affinity (A) as follows:

$$\chi = \frac{I + A}{2} \tag{28}$$

According to density functional theory (DFT) [2-4], the Lagrange multiplier associated with the normalization constraint is identified as the chemical potential (μ), *viz.*,

$$\mu = \left(\frac{\delta E}{\delta \rho}\right)_{v(\vec{r})} = v(\vec{r}) + \frac{\delta F[\rho]}{\delta \rho(\vec{r})} \tag{29}$$

where E and $v(\vec{r})$ are the total energy and external potential respectively. Therefore

$$-\chi = \left(\frac{\partial E}{\partial N}\right)_{v(\vec{r})} = \int \left(\frac{\delta E}{\delta \rho}\right)_{v(\vec{r})} \left(\frac{\partial \rho}{\partial N}\right)_{v(\vec{r})} d\vec{r} = \left(\frac{\delta E}{\delta \rho}\right)_{v(\vec{r})} = \mu \tag{30}$$

Using the finite difference approximation of $\left(\frac{\partial E}{\partial N}\right)_{v(\vec{r})}$, μ may be expressed in terms of *I* and *A* as:

$$\mu = -\chi = -\frac{I + A}{2} \tag{31}$$

Using Koopmans' theorem μ and χ can be expressed as:

$$\mu = -\chi = \frac{1}{2}\left(\in_{HOMO} + \in_{LUMO}\right) \tag{32}$$

Chemical Hardness (η) and Softness (S)

It is found that, in many cases electronegativity alone cannot account for the stability of a molecule. To account for the stability of a molecule and the direction of acid-base reactions Pearson [17] introduced two parameters, 'hardness' and 'softness' in chemistry.

For an N-electron system, the second derivative of energy with respect to N, keeping external potential $v(\vec{r})$ fixed, is considered to be a measure of the chemical hardness [18]:

$$\eta = \frac{1}{2}\left(\frac{\partial^2 E}{\partial N^2}\right)_{v(\vec{r})} = \frac{1}{2}\left(\frac{\partial \mu}{\partial N}\right)_{v(\vec{r})} \tag{33}$$

which would be always positive due to the convex nature of E *vs*. N curve.

In the Koopmans' framework, η can be expressed as:

$$\eta = \frac{1}{2}\left(\in_{LUMO} - \in_{HOMO}\right) \tag{34}$$

The inverse of hardness [19] can be defined as softness:

$$S = \frac{1}{2\eta} = \left(\frac{\partial N}{\partial \mu}\right)_{v(\vec{r})} \tag{35}$$

The softness is closely associated with the polarizability of a system. A larger (more polarizable) chemical system is softer and *vice versa*.

As χ and η measure the response of the system when N varies at constant $v(\vec{r})$, the response function [20] does that job when $v(\vec{r})$ changes for a fixed N. For weak electric and magnetic fields it is provided by polarizability and magnetizability respectively.

Polarizability (α)

The linear response of the electron density in the presence of an infinitesimal electric field F is defined as electric dipole polarizability and it represents a second order variation in energy

$$\alpha_{a,b} = -\left(\frac{\partial^2 E}{\partial F_a \partial F_b}\right) \; a,b = x,y,z \tag{36}$$

A soft molecule is more polarizable compared to the corresponding harder counterpart. Similarly the magnetizability can be defined.

Electrophilicity Index (ω)

Maynard *et al.* [21] have shown that the reaction rate of fluorescence decay experiment on human immunodeficiency virus type-1 (HIV-1) nucleocapsid protein p7 (N C_p7) when interacting with some electrophilic agents, *e.g.*, azodicarbonamide (ADA), N-ethylmaleimide (NEM) *etc.* gives an almost linear response with the square of electronegativity (χ) to the chemical hardness (η) ratio. The quantity χ^2/η is related to the capacity of an electrophile to promote a soft (covalent) reaction.

Prompted by the work of Maynard and co-workers [21], Parr *et al.* defined electrophilicity index (*ω*) [13] as:

$$\omega = \frac{\mu^2}{2\eta} = \frac{\chi^2}{2\eta} \tag{37}$$

which measures the stabilization in energy when the system acquires an additional electronic charge ΔN from the environment.

This descriptor has been shown to provide valuable insights into various quantitative–structure-activity/ property/ toxicity – relationship (QSAR/QSPR/QSTR) models [22].

Local Reactivity Descriptors

They take care of the site selectivity of an atom in a molecule.

Electron Density ($\rho(\vec{r})$)

The most important local descriptor is the electron density $\rho(\vec{r})$ itself, in the DFT framework. Electron density $\rho(\vec{r})$ is given as [2, 3]:

$$\rho(\vec{r}) = \left(\frac{\delta E(\rho)}{\delta v(\vec{r})} \right)_N \tag{38}$$

Fukui Function ($f(\vec{r})$)

The Fukui function (FF) [23] is one of the widely used local reactivity descriptors in modeling chemical reactivity and site selectivity. Fukui function (FF) is defined as [23]:

$$f(\vec{r}) = \left(\frac{\partial \rho}{\partial N} \right)_{v(\vec{r})} = \left(\frac{\delta \mu}{\delta v(\vec{r})} \right)_N , \tag{39}$$

$$\text{such that } \int f(\vec{r})\,d\vec{r} = 1.$$

In equation (39), the discontinuity on the slope of $\rho(\vec{r})$ *vs*. *N* curve at integral *N*, provides three types of Fukui functions which account for nucleophilic, electrophilic and radical attacks respectively, at a particular reaction site. These three functions can be expressed in a different form by the use of finite difference and frozen core approximations as follows [23]:

$$f^{+}(\vec{r}) = \left(\frac{\partial \rho}{\partial N}\right)^{+}_{v(\vec{r})} \cong \rho_{N+1}(\vec{r}) - \rho_{N}(\vec{r}) \approx \rho_{LUMO}(\vec{r}) \text{ [for nucleophilic attack]} \qquad (40a)$$

$$f^{-}(\vec{r}) = \left(\frac{\partial \rho}{\partial N}\right)^{-}_{v(\vec{r})} \cong \rho_{N}(\vec{r}) - \rho_{N-1}(\vec{r}) \approx \rho_{HOMO}(\vec{r}) \text{ [for electrophilic attack]} \qquad (40b)$$

$$f^{0}(\vec{r}) = \left(\frac{\partial \rho}{\partial N}\right)^{0}_{v(\vec{r})} \cong \frac{1}{2}\left(\rho_{N+1}(\vec{r}) - \rho_{N-1}(\vec{r})\right) \approx \frac{1}{2}\left(\rho_{HOMO}(\vec{r}) + \rho_{LUMO}(\vec{r})\right) \text{ [for radical attack]} \quad (40c)$$

Equations (40a) to (40c) provide a correspondence between the local parameters and the frontier orbital theory of chemical reactivity [24]. A large value of f^{+}, f^{-}, or f^{0} at any site indicates the probability of respective attacks at that site which would correspond to a large change in chemical potential.

The condensed Fukui functions are proposed by Yang *et al.* [25], considering a finite difference method and the Mulliken population analysis (MPA) scheme as:

$$f_k^{+} = q_k(N+1) - q_k(N) \text{ [for nucleophilic attack]} \qquad (41a)$$

$$f_k^{-} = q_k(N) - q_k(N-1) \text{ [for electrophilic attack]} \qquad (41b)$$

$$f_k^{o} = [q_k(N+1) - q_k(N-1)]/2 \text{ [for radical attack]} \qquad (41c)$$

where q_k is the electronic population of atom k in a molecule.

Local Softness ($s(\vec{r})$)

In "frontier-controlled" reactions, where frontier orbital densities play an important role, the tendency of a particular site to be involved is given by a local softness parameter. Local softness $s(\vec{r})$ is defined as [19]:

$$s(\vec{r}) = \left(\frac{\partial \rho}{\partial \mu}\right)_{v(\vec{r})} \tag{42}$$

which is related to the global softness S as:

$$S = \int s(\vec{r})d\vec{r} \tag{43}$$

Local softness is related to FF as follows:

$$s(\vec{r}) = \left(\frac{\partial \rho(\vec{r})}{\partial \mu}\right)_{v(\vec{r})} = \left(\frac{\partial \rho}{\partial N}\right)_{v(\vec{r})} \left(\frac{\partial N}{\partial \mu}\right)_{v(\vec{r})} = f(\vec{r})S \tag{44}$$

Both global and local softnesses may also be expressed as appropriate number fluctuations [19].

Local Hardness ($\eta(\vec{r})$)

The local hardness $\eta(\vec{r})$ is defined as [26]:

$$\eta(\vec{r}) = \frac{1}{2}\left(\frac{\delta \mu}{\delta \rho}\right)_{v(\vec{r})} \tag{45}$$

which is related to the global hardness as:

$$\eta = \int \eta(\vec{r}) f(\vec{r})\, d\vec{r} \tag{46}$$

which is not a simple integral over $\eta(\vec{r})$ as in the case of the local softness [eq. 43].

The definition (45) of local hardness is ambiguous [27] because of the inter-dependence between $\rho(\vec{r})$ and $v(\vec{r})$ according to DFT [1]. The situation may improve in an appropriate ensemble like an isomorphic ensemble [28]. While local softness is an electronic reactivity index, the local hardness may be considered to be a nuclear reactivity index and hence together they will take care of variations in N and $v(\vec{r})$ which will encompass all possible situations [29].

Philicity ($\omega(\vec{r})$)

Chattaraj *et al.* [30] proposed the generalized concept of philicity which contains almost all information regarding the global as well as local reactivity and selectivity, specially the electrophilic/nucleophilic power of a given atomic site in a molecule. This quantity may be considered as the local variant of the global electrophilicity index, called philicity ($\omega(\vec{r})$) and is defined as [30]:

$$\omega = \int \omega(\vec{r})\, d\vec{r} \tag{47}$$

Philicity is obtained through the resolution of the identity associated with the normalization of Fukui function [23,24], $f(\vec{r})$, as:

$$\omega^{\alpha}(\vec{r}) = \omega . f^{\alpha}(\vec{r}) \tag{48}$$

where α = +, -, and 0 refer to nucleophilic, electrophilic and radical reactions respectively. Corresponding condensed-to-atom variants may be written for the kth atomic site in a molecule as

$$\omega_k^{\alpha} = \omega . f_k^{\alpha} \tag{49}$$

In eq. (48) any normalized-to-one quantity (*e.g.* the shape function, $\sigma(\vec{r}) = \rho(\vec{r}) / N$) may be used. But FF is preferred owing to the explicit information of electron addition/removal in it.

Electronic Structure Principles

The global and local reactivity descriptors are better appreciated through various related electronic structure principles, such as Sanderson's electronegativity equalization principle [31], hard and soft acids and bases (HSAB) principle [17,32], maximum hardness principle (MHP) [33,34], minimum polarizability principle [35], minimum magnetizability principle (MMP) [36], *etc.*

Electronegativity Equalization Principle (EEP)

The difference in electronegativity plays a major role in chemical reactions. Electrons are transferred from a species of lower electronegativity to a species of higher electronegativity until both possess equal electronegativity values.

Sanderson postulated that [31], during molecule formation the electronegativities of the constituent atoms become equal, yielding a molecular electronegativity (χ_M) which is roughly the geometric mean of the electronegativities of the isolated atoms,

$$\chi_M = (\chi_A^0 \chi_B^0 \chi_C^0 \ldots)^{1/(a+b+c+\ldots)} \tag{50}$$

where a, b, c are the numbers of atom of a given element (A, B, C, *etc.*).

As an application of the electronegativity equalization principle (EEP), Parr and Pearson [18] derived Eqs. (26) and (27) to measure the amount of charge transfer ΔN and the energy change ΔE associated with the formation of A:B complex from acid A and base:B.

These expressions are very useful in understanding the acid-base reaction mechanism. It is important to note that the electronegativity difference drives the electron transfer whereas the hardness sum provides a resistance to it. Therefore both χ and η are to be considered in analyzing these processes.

Hard-Soft-Acid-Base (HSAB) Principle

Pearson introduced the hard-soft-acid-base (HSAB) principle [17, 32] which in general can describe a variety of acid-base reactions. This principle is stated as, ‘hard acids prefer to coordinate with hard bases and soft acids prefer to coordinate with soft bases for both their thermodynamic and kinetic properties’.

In order to quantify the concept of hardness and softness Pearson proposed [37] a relation that correlates the stability of a molecule with the hardness and softness, as well as the inherent strengths of acids and bases. The stability constant of a reaction is given by,

$$\log K = S_A S_B + \sigma_A \sigma_B \tag{51}$$

where σ_A and σ_B are the inherent strengths of acids and bases whereas S_A and S_B are the softness factors. It is expected that σ_A, σ_B would be related to χ_A, χ_B.

Maximum Hardness Principle (MHP)

Pearson's HSAB principle has been analyzed and it has been argued [38] that the hard-hard reactions are governed by the charge-controlled interactions and the soft-soft interactions are of the covalent type. Various studies on the chemical reactivity suggest that soft molecules are more reactive compared to the corresponding harder counterparts. Hence, isomeric molecules with higher chemical hardness are found to be more abundant compared to the molecules with lower hardness values. This fact leads to the principle of maximum hardness. The maximum hardness principle (MHP) is stated [33, 34] as 'there seems to be a rule of nature that molecules arrange themselves so as to be as hard as possible'.

Minimum Polarizability Principle (MPP)

As a consequence of the maximum hardness principle (MHP) [33, 34] and an inverse relationship [39] between hardness and polarizability, a minimum polarizability principle (MPP) is stated as [35], 'the natural direction of evolution of any system is towards a state of minimum polarizability'.

Various physicochemical properties like molecular vibrations, internal rotations, chemical reactions, aromaticity, atomic shell structure, excited states, dynamical problems *etc.* are analyzed with the use of both MHP and MPP [4, 14].

Minimum Electrophilicity Principle (M-El-P)

The minimum electrophilicity principle [40] is stated as 'Electrophilicity will be a minimum (maximum) when both chemical potential and hardness are maxima (minima)'. In order to analyze the possible connection between the extremal values of electrophilicity (ω) and stability it has been shown [40] that the extremum on the electrophilicity occurs during chemical reactions, molecular vibrations and internal rotations at the points for which the following condition is satisfied

$$\frac{\partial \mu}{\partial \lambda} = \frac{\mu}{2\eta}\left[\frac{\partial \eta}{\partial \lambda}\right] \tag{52}$$

where λ can be a reaction coordinate (reaction), bond length (stretching), bond angle (bending) or dihedral angle (internal rotation).

Since μ is negative and η is positive, the extremum of electrophilicity occurs when the slopes of the variations of μ and η are of opposite signs. Therefore the extrema in chemical potential and hardness will ensure extremum in ω. In general μ and η are maxima for the equilibrium geometry implying the minimum value for ω. In those cases the stability would be related to the minimum value of ω.

Minimum Magnetizability Principle (MMP)

Magnetizability of a system can be expressed in terms of its diamagnetic (ξ_{dm}) and paramagnetic (ξ_{pm}) components as follows:

$$\xi_{Total} = \xi_{dm} + \xi_{pm} \tag{53}$$

Very recently, a new electronic structure principle, *viz*. the minimum magnetizability principle (MMP) [36] has been proposed to extend the domain of applicability of the conceptual density functional theory (DFT) in explaining the magnetic interactions and magnetochemistry. This principle is stated as, "a stable configuration/conformation of a molecule or a favorable chemical process is associated with a minimum value of the magnetizability". It is also established that a soft molecule can be easily polarizable and magnetizable than a hard one.

Electrophilicity Equalization Principle (El-E-P)

It is known that during the interaction of an electrophile with a nucleophile, the electrophilicity of the former is reduced (*via* electronic charge transfer and/or other related processes, from the nucleophile to the electrophile), and that of the latter is increased until they are equalized to a final value in between the two. This principle is stated as "The electrophilicity gets equalized during molecule formation, and the final equalized electrophilicity may be expressed as the geometric mean of the isolated atom values [41]." An important outcome of this result is that the local electrophilicity [30, 40] may alternatively be considered to be constant everywhere and is equal to its global variant.

Dynamical Situations

In order to tackle the time-dependent problems within a density based quantum mechanical framework we start with the Quantum Fluid Dynamics (QFD) approach using the time dependent Schrödinger equation (TDSE) for a single particle.

Substituting the following polar form

$$\Psi(\vec{r},t) = R(\vec{r},t)\, e^{is(\vec{r},t)/\hbar}$$

in the TDSE:

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V\right]\Psi = i\hbar\frac{\partial\Psi}{\partial t} \tag{54}$$

and separating out the real and the imaginary parts we obtain an equation of continuity:

$$\frac{\partial\rho}{\partial t} + \nabla.(\rho\vec{v}) = 0 \tag{55}$$

and an Euler-type equation of motion,

$$m\rho\left[\frac{\partial\vec{v}}{\partial t} + (\vec{v}.\nabla)\vec{v}\right] = -\rho\nabla(V + V_{qu}) \tag{56}$$

where V_{qu} is the quantum potential.

In order to follow the behaviour of various reactivity descriptors in a time dependent situation and also to analyze the dynamical variants of the above mentioned electronic structure principles a quantum fluid density functional theory (QFDFT) [42] has been made use of which is obtained by combining QFD and time dependent (TD) DFT [43] to have the following equations

$$\frac{\partial\rho}{\partial t} + \nabla.(\rho\nabla\xi) = 0 \tag{57a}$$

and

$$\frac{\partial \xi}{\partial t}+\frac{1}{2}(\nabla \xi)^2+\frac{\delta G[\rho]}{\delta \rho}+\int \frac{\rho(\vec{r}\,',t)}{|\vec{r}-\vec{r}\,'|}d\vec{r}\,'+v_{ext}(\vec{r},t)=0 \tag{57b}$$

where ξ is the velocity potential, $G[\rho]$ contains kinetic and exchange-correlation energy functionals and $v_{ext}(\vec{r},t)$ is the total external potential including $v(\vec{r})$.

The above equations may be alternatively written in the form of the following generalized nonlinear Schrödinger equation (GNLSE) [42],

$$\left[-\frac{1}{2}\nabla^2+v_{eff}(\vec{r},t)\right]\Phi(\vec{r},t)=i\frac{\partial \Phi(\vec{r},t)}{\partial t} \tag{58a}$$

where the effective potential takes the form as

$$v_{eff}(\vec{r},t)=\frac{\delta T_{NW}}{\delta \rho}+\frac{\delta E_{xc}}{\delta \rho}+\int \frac{\rho(\vec{r}\,',t)}{|\vec{r}-\vec{r}\,'|}d\vec{r}\,'+v_{ext}(\vec{r},t) \tag{58b}$$

where T_{NW} and E_{XC} are the non-Weizsäcker part of the kinetic energy and the exchange-correlation energy functionals respectively. The 3-D complex valued hydrodynamical function $\Phi(\vec{r},t)$ may be written in the following polar form:

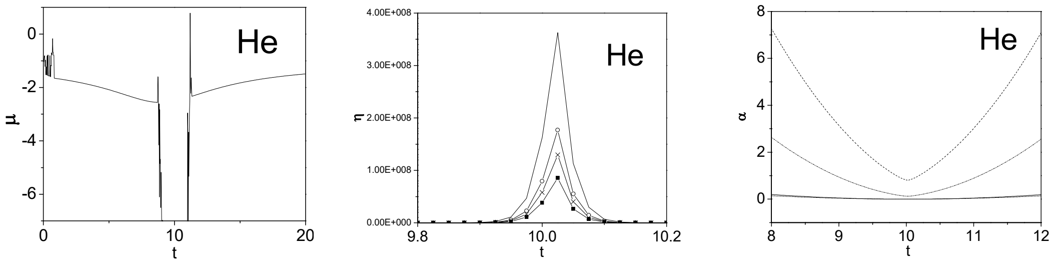
$$\Phi(\vec{r},t)=\rho(\vec{r},t)^{1/2}\exp(i\xi(\vec{r},t));i=\sqrt{-1} \tag{59a}$$

$$\rho(\vec{r},t)=|\Phi(\vec{r},t)|^2 \tag{59b}$$

$$\vec{j}(\vec{r},t)=\left[\Phi_{re}\nabla\Phi_{im}-\Phi_{im}\nabla\Phi_{re}\right]=\rho\nabla\xi \tag{59c}$$

The TD processes chosen for this purpose are: (I) Ion-atom collision and (II) Atom-field interaction. The GNLSE is solved for problems I and II with pertinent $v_{ext}(\vec{r},t)$.

Fig. (**1**) represents the dynamical profiles of various global reactivity parameters, *viz*., chemical potential (μ), hardness (η) and polarizability (α) during a collision process between a proton and a *He*-atom in its ground and different excited states. As shown in the Fig. **1a**, the TD chemical potential divides the collision process into three regimes [35, 44], *viz*., approach, encounter and departure. In the encounter regime the actual chemical process takes place where the hardness maximizes and polarizability minimizes which may be thought of as the dynamical variants of the MHP and the MPP respectively (Figs. **1b** and **1c**).



(a) **(b)** **(c)**

Figure 1: Dynamical profiles of a) chemical potential (μ), b) hardness (η) and c) polarizability (α) during a collision process between a proton and a He-atom in its ground state and also in different excited states. Reproduced with permission from ref. [32(c)]. Copryright 2003 American Chemical Society.

Since proton is a hard acid and the hardness decreases and polarizability increases with electronic excitation [45], a dynamical variant of the HSAB principle is also observed [32] which in turn helps analyzing the regioselectivity in a chemical reaction [46].

For the problem II the external field may be an electric field or may be a generic field simulating the presence of another reactant or a reagent or a solvent. While chemical potential (first order effect) oscillates in phase with that of electric field for a moderate field strength, a relatively higher field intensity is needed to obtain an in-phase oscillation in hardness.

The nature of the oscillations in ω depends on those of μ and η. The interplay between the central nuclear Coulomb field and the axial external electric field governs the overall reactivity dynamics for both the electronic states [47].

Important insights into chaotic ionization of Rydberg atoms in presence of external field may also be obtained from these studies [48].

CONCLUSIONS

Conceptual density functional theory is developed here to understand the chemical reactivity in static and dynamic situations. Various reactivity descriptors and the associated electronic structure principles are now better understood.

ACKNOWLEDGEMENTS

We are thankful to BRNS, Mumbai and SERB (DST), New Delhi for financial assistance.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

ABBREVIATIONS AND SYMBOLS OF SOME IMPORTANT QUANTITIES

A	= Electron affinity
α	= Polarizability
B3LYP	= Becke three-parameter Lee-Yang-Parr functional
$[\partial\eta/\partial N]$	= Variation of hardness with electron number
DFT	= Density functional theory
$\in_{\mathrm{HOMO}}$	= Highest occupied molecular orbital energy
$\in_{\mathrm{LUMO}}$	= Lowest unoccupied molecular orbital energy
E_{xc}	= Exchange-correlation energy functionals
FF	= Fukui function
f_k^{α} (α=+,-,0)	= Condensed Fukui function

$F[\rho]$	=	Hohenberg-Kohn-Sham universal functional
$f(\vec{r})$	=	Fukui function
η	=	Hardness
GNLSE	=	Generalized nonlinear Schrödinger equation
HOMO	=	Highest occupied molecular orbital
HSAB	=	Hard and soft acids and bases
I	=	Ionization potential
LUMO	=	Lowest unoccupied molecular orbital
μ	=	Chemical potential
MHP	=	Maximum hardness principle
MPP	=	Minimum polarizability principle
N	=	Number of electrons
PW91	=	Perdew and Wang 91
PBE	=	Perdew, Burke and Ernzerhof
PBE0	=	Perdew, Burke and Ernzerhof 0
QFD	=	Quantum fluid dynamics
QSAR	=	Quantitative structure-activity relationship
QSPR	=	Quantitative structure-property relationship
QSTR	=	Quantitative structure-toxicity relationship
$\rho(\vec{r})$	=	Electron density

S	= Softness
$s(\vec{r})$	= Local softness
t_F	= Thomas-Fermi kinetic energy density
T_{NW}	= Non-Weizsäcker part of the kinetic energy
VWN	= Vosko-Wilk-Nusair
V_{qu}	= Quantum potential
$v_{ext}(\vec{r},t)$	= Time dependent external potential
$v(\vec{r})$	= External potential
$\Phi(\vec{r},t)$	= 3-D hydrodynamical wave function
χ	= Electronegativity
ω	= Electrophilicity index
$\omega_k{}^{\alpha}$	= Condensed philicity
$\omega(\vec{r})$	= Philicity

REFERENCES

[1] Hohenberg, P.; Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **1964**, *136*, B864-B871.

[2] (a) Kohn, W.; Sham, L. Self-consistent equations including exchange and correlation effects J. *Phys. Rev.* **1965**, *140*, A1133-A1138. (b) Parr, R. G.; Yang, W. *Density Functional Theory of Atoms and Molecules*, Oxford University Press, New York, **1989**.

[3] (a) Kohn, W.; Becke, A. D.; Parr, R. G. Density functional theory of electronic structure *J. Phys. Chem.* **1996**, *100*, 12974-12980. (b) Chattaraj, P. K.; Nath, S.; Maiti, B. Reactivity Descriptors. In: *Computational Medicinal Chemistry for Drug Discovery*, Tollenaere, J.; Bultinck, P.; Winter, H. D.; Langenaeker, W. Eds.; Marcel Dekker: New York, **2003**; Chapter 11, pp. 295- 322.

[4] Geerlings, P.; De Proft, F.; Langenaeker, W. Conceptual density functional theory *Chem. Rev.* **2003**, *103*, 1793-1874.

[5] Vosko, S. H.; Wilk, L.; Nusair, M. Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis *Can. J. Phys.* **1980**, *58* 1200-1211.

[6] Perdew, J. P. *Electronic Structure of Solids*, Ziesche, P.; Eschrig, H. (Eds.), Akademie: Berlin, Germany, **1991**.

[7] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865-3868.

[8] Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648-5652.

[9] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1997**, *78*, 1396-1396.

[10] *Electronegativity: Structure and Bonding*, Sen, K. D.; Jorgenson, C. K. Eds.; Springer-Verlag: Berlin, **1987**; Vol. 66.

[11] *Chemical Hardness: Structure and Bonding*, Sen, K. D.; Mingos, D. M. P. Eds.; Springer-Verlag: Berlin, **1993**, Vol. 80.

[12] Pearson, R. G. *Chemical Hardness: Applications from Molecules to Solids*, Wiley-VCH Verlag GMBH: Weinheim, **1997**.

[13] Parr, R. G.; Szentpaly, L. v.; Liu, S. Electrophilicity index *J. Am. Chem. Soc.* **1999**, *121*, 1922-1924.

[14] Chattaraj, P. K.; Sarkar, U.; Roy, D. R. Electrophilicity index *Chem. Rev.* **2006**, *106*, 2065-2091.

[15] Pauling, L. *The Nature of the Chemical Bond*, 3rd ed., Cornell University Press, Ithaca, New York, **1960**.

[16] (a) Mulliken, R. S. A new electron affinity scale; Together with data on valence states and on valence ionization potentials and electron affinities *J. Chem. Phys.* **1934**, *2*, 782-793. (b) Mulliken, R. S. Electronic structures of molecules XI. Electroaffinity, molecular orbitals and dipole moments *J. Chem. Phys.* **1935**, *3*, 573-585.

[17] (a) Pearson, R. G. Hard and soft acids and bases—the evolution of a chemical concept *Coord. Chem. Rev.* **1990**, *100*, 403-425. (b) Pearson, R. G. *Hard and Soft Acids and Bases*, Dowden, Hutchinson and Ross: Stroudsberg, PA, **1973**. (c) Hancock, R. D.; Martell, A. E. Hard and soft acid-base behavior in aqueous solution: Steric effects make some metal ions hard: A quantitative scale of hardness-softness for acids and bases *J. Chem. Educ.* **1996**, *73*, 654-661.

[18] Parr, R. G.; Pearson, R. G. Absolute hardness: companion parameter to absolute electronegativity *J. Am. Chem. Soc.* **1983**, *105*, 7512-7516.

[19] Yang, W.; Parr, R. G. Hardness, softness, and the Fukui function in the electronic theory of metals and catalysis *Proc. Natl. Acad. Sci. USA* **1985**, *82*, 6723-6726.

[20] Berkowitz, M.; Parr, R. G. Molecular hardness and softness, local hardness and softness, hardness and softness kernels, and relations among these quantities *J. Chem. Phys.* **1998**, *88*, 2554-2557.

[21] Maynard, A. T.; Huang, M.; Rice, W. G.; Covell, D. G. Reactivity of the HIV-1 nucleocapsid protein p7 zinc finger domains from the perspective of density-functional theory *Proc. Natl. Acad. Sci. USA.* **1998**, *95*, 11578-11583.

[22] (a) Parthasarthi, R.; Subramanian, V.; Roy, D. R.; Chattaraj, P. K. Electrophilicity index as a possible descriptor of biological activity *Bioorg. Med. Chem.* **2004**, *12*, 5533-5543. (b) Roy, D. R.; Parthasarathi, R.; Maiti, B.; Subramanian, V.; Chattaraj, P. K. Electrophilicity as a possible descriptor for toxicity prediction *Bioorg. Med. Chem.* **2005**, *13*, 3405-3412. (c) Padmanabhan, J.; Parthasarathi, R.; Subramanian, V.; Chattaraj, P. K. Group philicity and electrophilicity as possible descriptors for modeling ecotoxicity applied to chlorophenols *Chem. Res. Tox.* **2006**, *19*, 356-364.

[23] (a) Parr, R. G.; Yang, W. Density functional approach to the frontier-electron theory of chemical reactivity *J. Am. Chem. Soc.* **1984**, *106*, 4049-4050. (b) Ayers, P. W.; Levy, M.

Perspective on "Density functional approach to the frontier-electron theory of chemical reactivity" *Theor. Chem. Acc.* **2000**, *103*, 353-360.

[24] (a) Fukui, K. *Theory of Orientation and Stereoselection*, Springer – Verlag, Berlin, **1975**. (b) Fukui, K. Role of frontier orbitals in chemical reactions *Science* **1987**, *218*, 747-754.

[25] Yang, W.; Mortier, W. J. The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines *J. Am. Chem. Soc.* **1986**, *108*, 5708-5711.

[26] (a) Berkowitz, M.; Ghosh, S. K.; Parr, R. G. On the concept of local hardness in chemistry *J. Am. Chem. Soc.* **1985**, *107*, 6811-6814. (b) Ghosh, S. K.; Berkowitz, M. A classical fluid-like approach to the density-functional formalism of many-electron systems *J. Chem. Phys.* **1985**, *83*, 2976-2983.

[27] (a) Harbola, M. K.; Chattaraj, P. K.; Parr, R. G. Aspects of the softness and hardness concepts of density functional theory *Israel. J. Chem.* **1991**, *321*, 395-402. (b) Ghosh, S. K. Energy derivatives in density-functional theory *Chem. Phys. Lett.* **1990**, *172*, 77-82.

[28] De Proft, F.; Liu, S.; Parr, R. G. Chemical potential, hardness, hardness and softness kernel and local hardness in the isomorphic ensemble of density functional theory *J. Chem. Phys.* **1997**, *107*, 3000-3006.

[29] De Proft, F.; Liu, S.; Geerlings, P. Calculation of the nuclear Fukui function and new relations for nuclear softness and hardness kernels *J. Chem. Phys.* **1998**, *108*, 7549-7554.

[30] (a) Chattaraj, P. K.; Maiti, B.; Sarkar, U. Philicity: A unified treatment of chemical reactivity and selectivity *J. Phys. Chem. A* **2003**, *107*, 4973-4975. (b) Roy, D. R.; Parthasarathi, R.; Padmanabhan, J.; Sarkar, U.; Subramanian, V.; Chattaraj, P. K. Careful scrutiny of the philicity concept *J. Phys. Chem. A* **2006**, *110*, 1084-1093.

[31] (a) Sanderson, R. T. An interpretation of bond lengths and a classification of bonds *Science* **1951**, *114*, 670-672. (b) Sanderson, R. T. Carbon-carbon bond lengths **1952**, *116*, 41-42. (c) Sanderson, R. T. Partial charges on atoms in organic compounds **1955**, *121*, 207-208. (d) Sanderson, R. T. Electronegativities in inorganic chemistry *J. Chem. Educ.* **1952**, *29*, 539-544. (e) Sanderson, R. T. Electronegativities in inorganic chemistry. III *J. Chem. Educ.* **1954**, *31*, 238-245.

[32] (a) Chattaraj, P. K.; Lee, H.; Parr, R. G. HSAB principle *J. Am. Chem. Soc.* **1991**, *113*, 1855-1856. (b) Chattaraj, P. K.; Schleyer, P. v. R. An ab initio study resulting in a greater understanding of the HSAB principle *J. Am. Chem. Soc.* **1994**, *116*, 1067-1071. (c) Chattaraj, P. K.; Maiti, B. HSAB principle applied to the time evolution of chemical reactions *J. Am. Chem. Soc.* **2003**, *125*, 2705-2710.

[33] (a) Pearson, R. G. Recent advances in the concept of hard and soft acids and bases *J. Chem. Educ.* **1987**, *64*, 561-567. (b) Pearson, R. G. The principle of maximum hardness *Acc. Chem. Res.* **1993**, *26*, 250-255. (c) Pearson, R. G. Maximum chemical and physical hardness *J. Chem. Educ.* 1999, *76*, 267-275.

[34] (a) Parr, R. G.; Chattaraj, P. K. Principle of maximum hardness *J. Am. Chem. Soc.* **1991**, *113*, 1854-1855. (b) Ayers, P. W.; Parr, R. G. Variational principles for describing chemical reactions: the Fukui function and chemical hardness revisited *J. Am. Chem. Soc.* **2000**, *122*, 2010-2018.

[35] (a) Chattaraj, P. K.; Sengupta, S. Popular electronic structure principles in a dynamical Context *J. Phys. Chem.* **1996**, *100*, 16126-16130. (b) Ghanty, T. K.; Ghosh, S. K. A density functional approach to hardness, polarizability, and valency of molecules in chemical reactions *J. Phys. Chem.* **1996**, *100*, 12295-12298.

[36] Tanwar, A.; Roy, D. R.; Pal, S.; Chattaraj, P. K. Minimum magnetizability principle *J. Chem. Phys.* **2006**, *125*, 056101-056102.

[37] Pearson, R. G. [Quantitative evaluation of the HSAB (hard-soft acid-base) concept]. Reply to the paper by Drago and Kabler *Inorg. Chem.* **1972**, *11*, 3146-3146.

[38] (a) Klopman, G. Chemical reactivity and the concept of charge- and frontier-controlled reactions *J. Am. Chem. Soc.* **1968**, *90*, 223-224. (b) Klopman, G. *Chemical Reactivity and Reaction Paths*, Klopman, G. (Ed.), Wiley, New York, **1974**, Chap. 4. (c) Chattaraj, P. K. Chemical reactivity and selectivity: Local HSAB principle *versus* frontier orbital theory *J. Phys. Chem. A.* **2001**, *105*, 511-513.

[39] (a) Pearson, R. G. in Ref. 5. Politzer, P. A relationship between the charge capacity and the hardness of neutral atoms and groups *J. Chem. Phys.* **1987**, *86*, 1072-1074. (b) Ghanty, T. K.; Ghosh, S. K. Correlation between hardness, polarizability, and size of atoms, molecules, and clusters *J. Phys. Chem.* **1993**, *97*, 4951-4953.

[40] (a) Chamorro, E.; Chattaraj, P. K.; Fuentealba, P. Electrophilicity index along the reaction path *J. Phys. Chem. A* **2003**, *107*, 7068-7072. (b) Chattaraj, P. K.; Gutierrez- Oliva, S.; Jaque, P.; Toro-Labbe, A. Towards understanding the molecular internal rotations and vibrations and chemical reactions through the profiles of reactivity and selectivity indices: an ab initio SCF and DFT study *Mol. Phys.* **2003**, *101*, 2841-2853. (c) Parthasarathi, R.; Elango, M.; Subramanian, V.; Chattaraj, P. K. Variation of electrophilicity during molecular vibrations and internal rotations *Theor. Chem. Acc.* **2005**, *113*, 257-266.

[41] Chattaraj, P. K.; Giri, S.; Duley, S. Electrophilicity equalization principle *J. Phys. Chem. Lett.* **2010**, *1*, 1064-1067.

[42] Deb, B. M.; Chattaraj, P. K. Density-functional and hydrodynamical approach to ion-atom collisions through a new generalized nonlinear Schrödinger equation *Phys. Rev. A* **1989**, *39*, 1696-1713.

[43] Runge, E.; Gross, E. K. U. Density-functional theory for time-dependent systems *Phys. Rev. Lett.* **1984**, *52*, 997-1000.

[44] Chattaraj, P. K.; Sengupta, S. Dynamics of chemical reactivity indices for a many-electron system in its ground and excited states *J. Phys. Chem. A* **1997**, *101*, 7893-7900.

[45] (a) Chattaraj, P. K.; Poddar, A. Chemical reactivity and excited-state density functional theory *J. Phys. Chem. A* **1999**, *103*, 1274-1275. (b) Chattaraj, P. K.; Poddar, A. A density functional treatment of chemical reactivity and the associated electronic structure principles in the excited electronic states *J. Phys. Chem. A* **1998**, *102*, 9944-9948. (c) Chattaraj, P. K.; Poddar, A. Molecular reactivity in the ground and excited electronic states through density-dependent local and global reactivity parameters *J. Phys. Chem. A* **1999**, *103*, 8691-8699. (d) Fuentealba, P.; Simon-Manso, Y.; Chattaraj, P. K. Molecular electronic excitations and the minimum polarizability principle *J. Phys. Chem. A* **2000**, *104*, 3185-3187.

[46] Chattaraj, P. K.; Maiti, B. Regioselectivity in the chemical reactions between molecules and protons: A quantum fluid density functional study *J. Phys. Chem. A* **2004**, *108*, 658-664.

[47] (a) Chattaraj, P. K. Quantum fluid density functional theory of helium atom in an intense laser field *Int. J. Quantum Chem.* **1992**, *41*, 845-859. (b) Chattaraj, P. K.; Maiti, B. Reactivity dynamics in atom−field interactions: A quantum fluid density functional study *J. Phys. Chem. A* **2001**, *105*, 169-183.

[48] (a) Chattaraj, P. K.; Sengupta, S. Chemical hardness as a possible diagnostic of the chaotic dynamics of rydberg atoms in an external field *J. Phys. Chem. A* **1999**, *103*, 6122-6126. (b) Chattaraj, P. K.; Sarkar, U. Ground- and excited-states reactivity dynamics of hydrogen and helium atoms *Int. J. Quantum Chem.* **2003**, *91*, 633-650.

CHAPTER 10

Mathematical (Structural) Descriptors in QSAR: Applications in Drug Design and Environmental Toxicology

Marjan Vračko*

Kemijski Inštitut/National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

Abstract: In the chapter we present a short overview of QSAR (Quantitative Structure-Activity Relationship) modeling. The QSAR paradigm grounds on an assumption that properties of a compound depend on its chemical structure. In its final form a QSAR model is expressed as a mathematical relationship between molecular structure and property. A model is built on existing knowledge, *i.e.*, on a set of compounds with known structures and known properties. The QSAR models are widely used in rational drug design and in the environmental toxicology. As examples we present a case study of QSAR modeling in searching for new anti-tuberculosis drugs and the predictions of five toxicological endpoints with the internet available program CAESAR.

Keywords: QSAR modeling, topological, electro-topological, quantum chemical descriptors, anti-tuberculosis drugs, fluoroquinolones, environmental sciences, oecd principles for validation of qsar models, caesar programs for bio-concentration factor, mutagenicity, carcinogenicity, skin sensitization, developmental toxicity.

INTRODUCTION

QSAR is an acronym for Quantitative Structure-Activity Relationship. The paradigm grounds on an assumption that properties of a compound depend on its chemical structure. In its final form a QSAR model is expressed as a mathematical relationship between molecular structure and property. A model is built on existing knowledge, *i.e.* on a set of compounds with known structures and known properties. The developing of a QSAR model has four basic steps. The first step is building up the data set, the second one is the determination of

***Corresponding author Marjan Vračko:** Kemijski inštitut/National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia; Tel: +386 1 4760315; Fax: +386 1 4760300; E-mail: marjan.vracko@ki.si

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50010-2

molecular structures and molecular descriptors, the third one is the construction of models using the mathematical (chemometrical) methods, and the fourth one is the testing and validation of them. The QSAR model is mostly used in two areas: in rational drug design and in chemical regulation for evaluation of toxicological and eco-toxicological parameters of compounds.

The collection and checking of data (chemical structures and properties) are crucial parts of QSAR modeling. It is to emphasize that the quality of the data determines the quality of the final model. The data used for modeling should be obtained under the same laboratory conditions and using the same experimental protocols. With the special care the molecular structures should be checked. Databases usually consist of miscellaneous information including substances like metals, salts, organic compounds, or mixtures. Solely the organic compounds can have different isomeric, tautomeric or enantiomeric forms. When compiling the data set we must be aware of these pitfalls. In short, the compilation of data is an important and time consuming work, which usually takes more than a half of the total time used for model development.

The second step is structure determination. One has to decide how molecules will be considered in the model. It is obvious that there is a certain hierarchy in description of a molecule. A molecule can be considered as collection of fragments, or represented by two-dimensional, or three-dimensional structures. Three-dimensional structure is determined by positions of all atoms, which constitute a molecule. Molecular structures form a basis for calculation of descriptors. An insight on structural descriptors is given in the section below. Nowadays a variety of computer software packages are available to calculate descriptors and hundreds of them can be easily calculated. A selection of the most relevant descriptors represents a basic problem in the developing QSAR models.

In the third step the modeling method must be selected. The most applied method is multi-dimensional linear regression. Recently, the more advanced methods as the principal component analysis, partial least square, Ridge regression, artificial neural networks of different architectures and learning algorithms have become part of QSAR modeling. These methods are often applied in combination with algorithms for descriptor selection. One of such algorithms is the genetic algorithm. Basically, it is an analysis of an ensemble of models with an algorithm, which mimics the natural evolution. Some other methods consider all descriptors giving different importance (weights) to the descriptors. In most of the QSAR models, the property to model is expressed as a continuous variable, as for example dose of activity. Alternatively, the property can be given as affiliation to a particular class of activity. For classification problems, a variety of methods are available, *e.g.* linear discriminate analysis, support vector machine, artificial neural networks of specific architectures, *etc.*

The last step includes testing and validating the models. The questions are: how to test a model and how to express the quality of a model? Today, a basic concept is accepted that a model should be tested with an independent test set. An independent test set means a set that was never used in the model developing procedure. Before the start of the modeling development a test set is excluded from the compiled data set. Again, different strategies are possible. Usually, a random selection is performed, or, alternatively, the objects for the test set are selected equivocally from the entire model's domain [1]. When the model is presented in its final form it is tested with this test set. The quality of a model is usually expressed as the correlation coefficient between predicted and measured values. The correlation coefficients r^2 and q^2, which are often used in leave-out testing strategies, are defined as follows [2].

$$r^2 = \frac{\sum_i (y_i^e - \overline{y}^e)(y_i^p - \overline{y}^p)}{\sqrt{\sum_i (y_i^e - \overline{y}^e)^2 \sum_i (y_i^p - \overline{y}^p)^2}} \tag{1}$$

$$q^2 = 1 - \frac{PRESS}{\sum_i (y_i^e - \overline{y}^e)^2} \qquad PRESS = \sum_i (y_i^e - y_i^p)^2$$

where y_i^e and y_i^p are respectively experimental and predicted values for the object *i*, the bared symbols represent the mean values. When the model is used for classification its performance is usually expressed as a ratio between correct and false classified objects. In binary classification, *i.e.* the response can be positive or negative, the model responses can be understood as true positive (TP), true negative (TN), false positive (FP), or false negative (FN). The models are evaluated with parameters precision (P), specificity (Sp) and sensitivity (Se), which are defined as:

$$P=(TP+TN)/(TP+TN+FP+FN),\ Se=TP/(TP+FN),\ Sp=TN/(TN+FP) \quad (2)$$

STRUCTURAL DESCRIPTORS

Descriptors are numerical parameters, which describe a chemical structure in the model. Today several hundreds of descriptors are in use. Considering the nature of descriptors they can be classified as physico-chemical or structural ones. From physico-chemical descriptors the most used parameters are: octanol/water partition coefficients (log P, log D), refractivity index, molecular weight, or different spectroscopic data. These descriptors can be determined experimentally, what further means that they are known only for existing material. This is a disadvantage because in the research we often treat hypothetical structures. The advantage of calculated descriptors is obvious, structural descriptors can be determined solely on the basis of molecular structure. The question of how to represent or encode a chemical structure in a numerical fashion is central for chemical informatics. An ideal representation should be: unique, uniform, reversible, and invariant on translation and rotation of the structure. Unique means that different structures have different representations, uniform means that the representation has the same dimension for all structures; reversibility means that the structure can be reconstructed from the representation; and invariance means that the representation should not be influenced if the structure is rotated or translated in space. No representation fulfills all of the four requirements simultaneously. For example, the basic geometrical representation when a molecule is represented with coordinates of all its atoms is unique and reversible, but not uniform and invariant. The representation of a structure with a set of descriptors is unique, which means that these descriptors are different for different structures (this is true only after limitation. Some descriptors may have the same values for different structures, *i.e.* they may be degenerate). In most cases the representation is uniform and invariant but not reversible, *i.e.* the structure cannot be determined from descriptors [3].

To understand the strategy for developing descriptors we must clarify the concept of ‘molecular structure’. If we talk about ‘chemical structure’, we must consider a hierarchy for the description. At the two-dimensional (2-D) level, we describe the structure with atoms and bonds between them (‘structural formulas’). At the three-dimensional (3-D) level, we describe the structure with positions of all atoms. The step from 2-D to 3-D poses a problem. Rigid molecules are rare, 3-D structures may be different for molecules in crystalline form, in solution, in gas phase, or in an environment of proteins. Often 3-D structures are determined theoretically and they are different when the different theoretical approximations are applied. In the determination of 3-D structures it is often assumed that molecules are isolated (*in vacuo*). In reality, molecules are embedded in an environment, which in biological systems often consists of proteins or other bio-molecules. At the next level, the structures are optimized in their environment, sometimes such optimization is referred as 4-dimensionional representation.

Fragments as Descriptors

Structural fragments can be used as descriptors. In this description a structure is encoded as a multi-dimensional vector where the binary vector components indicate the presence or absence of a particular fragment. In their pioneer work, Free and Wilson represent molecular structures as a sum of constitutional fragments and correlate the representation with activity [4]. The basic idea is that a property (activity, toxicity, *etc.*) is due to particular molecular fragments. In drug design, one searches for fragments (pharmacophores) which are most relevant for particular activity or property. The examples are presented in references [5, 6] where a property for general drug-likeness is considered. In an advanced approach the fragments can be combined with fragment (substructure or pharmacophore) descriptors, which include the information on physico-chemical properties of fragments [7]. An efficient encoding of chemical structures in a unique and reversible way still represents a problem in chemical informatics. In reference [8] authors propose the algorithm for reading of fragments directly from SMILES codes. In environmental toxicology the fragments (structural alerts) are applied for study of toxicological properties, as for example mutagenicity and carcinogenicity [9]. It is an intention to create a database of structural alerts, which are responsible for activity. Originally, the database was compiled on the basis of experts' knowledge [10]; in later approaches statistical tools have been implemented [11-13].

One-Dimensional Descriptors

They provide the basic information about molecules, like number of atoms, molecular weight, or lipophilic properties. Lipophilic parameters (log P and log D) counts under the most used descriptors in QSAR. They were introduced to QSAR modeling in the first pioneer works by Hansch [14]. Log P is expressed as a ratio of the solubility in octanol and water of the substance under study and it describes the readiness of a molecule to prefer the polar (water) or non-polar (octanol) environment. Similarly, log D is expressed as a ratio of both solubilities given as function of the solute's acidity pH. They were introduced as descriptors under the assumption that lipophilicity determines the transport of a drug from the site of application to the cell inside and thus basically determines the bio-activity of a compound. Log P (log D) can be measured or calculated theoretically using one of the many commercial and free programs. When using the calculated log P (log D) as descriptor one must be aware that in some cases the predicted values depend on the applied software [15]. With the consideration of pioneer works in QSAR modeling one has to mention two other empirical constants: Hammet and Taft substituent constants, which describe the electronic and steric properties of molecules. As further physico-chemical constants used as descriptors, one should mention water solubility, refractivity index and other partitioning coefficients, such as partitioning coefficients between blood and tissue, *etc.*

Two-Dimensional Descriptors – Topological Indices

Topological indices are deduced from two-dimensional representations of molecular structure ('structural formulas'), which show topological properties of molecules, *i.e.* how atoms are inter-connected, but they do not show metric properties, *i.e.* lengths and angles between bonds. In the language of mathematicians, molecules are graphs where atoms are vertices and bonds are edges. The topology is completely described with the adjacency matrix, which describes how atoms are connected. A further important quantity is the topological distance between two atoms. It is the shortest distance going from one atom to another following the molecular structure. Usually, the topological distances are collected in the distance matrix D.

The first topological index was proposed by Wiener and Plat about 60 years ago. Wiener index is defined as a half-sum of all elements of matrix D. After Wiener's contribution, the idea has been extended and has resulted in many different indices such as hyper-Wiener index, Zagreb index, Szeged index. All these indices are expressed as counts of distances and are therefore integer numbers. The next class of topological indices is made by connectivity indices. In 1975 the connectivity index was proposed by Randić as a sum over weighted edges – bonds [16].

$$\chi = \sum_{ij} \frac{1}{\sqrt{v_i v_j}} \tag{3}$$

v_i is the degree of the vertex *i*, *i.e.* the number of edges connected to the vertex. Afterwards, other connectivity indices were introduced, *e.g.* Kier and Hall (Kappa index), and Balaban index (J):

$$J = \frac{b}{\mu + 1} \sum \frac{1}{\sqrt{v_i v_j}} \tag{4}$$

where *b* is the number of bonds (edges) and μ is the cyclomatic number of the graph, which is defined as the number of edges in a cyclic graph that must be deleted from the graph to transform it into an acyclic one.

Another class of topological indices is that of information content indices, which has its origin in informational theory. Following this theory, a graph is considered as an ensemble *A* of subset A_i. Each subset A_i is defined by a particular equivalence relation, where p_i is the probability that a randomly selected element of *A* occurs in subset A_i. Similarly to the definition of entropy, the mean information content (IC) is defined as:

$$IC = -\sum p_i \log_2 p_i \tag{5}$$

For the practical application of information theory to molecular graphs, the crucial question is the definition of equivalent relationship. An example of an equivalent relationship is that of two vertices that are equivalent if they possess topologically equivalent first neighborhoods [17]. Considering different equivalence relationships, numerous information content indices are defined. Some examples for calculation of informational content indices are shown in reference [18].

An important topic in structure activity relationship is stereoisomerism. In attempts to include chirality in the description of structures, one usually introduces an extra factor, which multiplies the original topological index. Chirality is addressed in reference [19]. For further reading about topological indices the reference [20] is recommended.

Two-Dimensional Descriptors – Electrotopological Descriptors

The electrotopological indices encode, besides the topology, the basic electronic properties of atoms. Kier and Hall introduced E-state indices, which include information about the valence [21] of an atom and also information about neighboring atoms. The electrons are assigned as core electrons, or valence electrons, which are further assigned as σ, π or lone pairs electrons.

$$Z^v = number\,of\,(\sigma + \pi + lonepairs), \delta = number\,of\,(\sigma) \tag{6}$$

In this description hydrogen electrons are not considered. The intrinsic state of an atom is defined as:

$$I = \frac{(2/L)^2 Z^v + 1}{Z} \tag{7}$$

L being the principal quantum number (2 for second raw elements, 3 for third raw elements, *etc.*), *Z* the atomic number and Z^v the valence electron number.

E-state of an atom is defined as an intrinsic state (*I*), modified by intrinsic states of neighboring (all) atoms in the molecule (*ΔI*).

$$S_i = I_i + \sum_j \Delta I_{ij}, \quad \Delta_{ij} I_{ij} = \frac{I_i - I_j}{(d_{ij}+1)^k} \tag{8}$$

Where d_{ij} is the topological distance and the exponent k measures the importance of distance; usually k is equal 2.

In reference [22] authors introduced ETA (Extended Topochemical Atom) indices. The ETA scheme includes several parameters, which are defined in following expressions:

$$\alpha = \frac{Z - Z^v}{Z^v} . \frac{1}{PN - 1} \text{ Core count (zero for hydrogen atoms)} \tag{9}$$

$$\varepsilon = -\alpha + 0.3\delta^v \text{ electronegativity} \tag{10}$$

$$\beta = \sum x\nu + \sum y\pi + \delta \text{ VEM (Valence Electron Mobile) count} \tag{11}$$

PN is the period number, *x* and *y* are weights for σ and π bonds and are defined for a particular classes of chemicals. With these quantities the different ETA indices are defined as for example:

$$\gamma_i = \frac{\alpha_i}{\beta_i} \text{ VEM vertex count} \tag{12}$$

$$\eta = \sum_{i<j} \left[\frac{\gamma_i \gamma_j}{d_{ij}} \right]^{0.5} \text{ The composite index} \tag{13}$$

Eigenvalues of Topological Matrices

Eigenvalues represent one of the matrix invariants. In applications of matrix algebra to physics and theoretical chemistry, eigenvalues find different important interpretations. In the up-rising time of quantum mechanics, Hückel presented one of the first quantum mechanical models of molecules where molecules were described with matrices. The basis for such matrices was the adjacency matrix, in which the nonzero elements were replaced by empirical parameters. The eigenvalues of matrices were considered as molecular orbital energies, and eigenvectors of the matrix as molecular orbitals, which describe the quantum mechanical states of electrons [23]. Later, eigenvalues of adjacency matrices have been introduced as descriptors. Lovasz and Pelikan proposed a leading eigenvalue of adjacency matrix as branching index [24]. In further applications the elements of matrices are weighted with different physico-chemical constants, like resonance integrals, or dipole moment. An overview of different approaches on how to use eigenvalues as descriptors is given in reference [25].

Three-Dimensional Descriptors - Quantum Chemical Descriptors

Quantum mechanics is widely applied in chemistry. It provides methods for optimization of geometrical parameters and thus for determination of 3D molecular structures. The further task is the calculation of quantum chemical descriptors, which in some cases can put insight into particular chemical mechanisms [26]. Exact quantum chemical treatment, *i.e.* solving Schrödinger equation for electrons in Coulomb field of nuclei is not possible for real word molecules, indeed, the exact solution of Schrödinger equation exists only for hydrogen atom. The reason for this fact lies in the interaction between electrons. In chemistry, two approximations are mostly applied: Hartree-Fock (HF) approximation and electron density functional approximation. Hartree-Fock approximation is a very well elaborated method used for the treatment of many-fermion systems. Here, the electron-electron interaction is approximated with an average electronic potential, which is calculated in an iterative way. The basic natural law, which requires the change of the wave function's sign when two electrons are interchanged, appears in the HF equation as exchange potential. Results of the calculation are orbital energies and molecular orbitals. The configuration of the electronic states is determined with the occupation of orbitals. This serves as a basis for calculation of charge distributions in molecules. Hartree-Fock calculations are computer intensive and, in spite of the fast development of computers, the technique still remains limited to medium size molecules. To work out the electronic structure of large numbers of molecules, which are compiled in large scale data bases, semi-empirical methods are often used, like CNDO, MNDO, MINDO, AM1 or PM approximations [27]. The idea of this approximation is to replace the complicated electronic potential with empirical parameters. The density functional approximation follows a different philosophy.

Here, electrons are represented by a cloud of electron density, which is calculated directly using the Kohn-Sham equation [27]. Some results of quantum chemical calculations, which are mostly used as descriptors, are orbital energies, and from them the HOMO and LUMO play a special role. HOMO is the highest occupied molecular orbital and it is based on Koopmas' theorem related to ionization potential (IP), while LUMO is the lowest unoccupied molecular orbital and is related to electron affinity (EA).

$$IP = - EHOMO \text{ and } EA = - ELUMO \tag{14}$$

Further descriptors deduced from orbital energies are Mulliken electronegativity (ME) and electronic hardness (N) defined as:

$$ME = (IP+EA)/2 \text{ and } N = (IP-EA)/2 \tag{15}$$

To extend the set of descriptors, lower and higher molecular orbital energies can be taken into consideration [28]. Molecular orbitals can be used to calculate the charge on particular atoms and the charge distribution of molecules as a whole. Charges on particular atoms and multipole moments (dipole moment, quadrupole moment, *etc.*) are often taken as descriptors. Often one takes the maximal or minimal charge on particular atoms, which are regarded as important for the mechanism of activity. Alternatively, the electron-electron repulsion energy calculated at a quantum chemical level, or two-electron integrals, which describe particular interactions, can be taken as descriptors [29]. At the end, it is to emphasize that quantum chemical results depend on the method of calculation [30]. Therefore, it is recommended that all molecules, which are used in QSAR models, are treated within the same quantum chemical approximation.

Three-Dimensional Descriptors - Geometrical Descriptors

They are deduced from three-dimensional molecular structure, which is defined through positions of all atoms in the molecule. In the crystalline form of a given material, its three-dimensional structure can be measured by X-ray diffraction measurements. When the material is in gaseous phase or in solution, its 3D structure may be different. In QSAR studies, 3D structures are mostly determined theoretically by applying molecular mechanics or quantum chemical methods to minimize the total molecular energy. To the class of 3-D descriptors belong mass distribution descriptors such as moments of inertia and gravitation index, shape indices, surface area indices and van der Waals indices.

Usually, a large number of descriptors are calculated. For a selection of relevant descriptors different methods such as genetic algorithms, heuristic selection, clustering, *etc.*, can be applied [31]. Alternatively, all descriptors can be considered, but in such a case the role of each descriptor in the model is evaluated by a weight. Descriptors of a model may indicate the mechanism of activity of the target studied. For example, strong dependence on topological indices indicates more steric interactions; or strong dependence on orbital energies indicates importance of charge transfer mechanisms. Alternatively, a molecular structure can be encoded in a spectrum like object, which is a descriptor vector. An example is the 'spectrum-like representation of molecular structures' where positions of atoms are projected on three perpendicular planes and the projections are converted to spectral forms. Alternatively, the positions of atoms can be transformed to a reciprocal space where the components in such a space can be taken as descriptors (3D-MoRSE code) [33]. In reference [32] authors compared models built with topological indices, geometric+electrostatic indices, and spectrum like representations and found that the latter method outperformed the others. In references [34, 35] authors analyzed the role of different types of descriptors in modeling mutagenicity. They compared the models built with topostructural (TS), topochemical (TC), three-dimensional (3D), quantum chemical (QC) descriptors and combinations of descriptors [35]. According to their quality, which was evaluated with the statistical parameters, the models can be ordered into a series: TS+TC > TS+TC+3D > TS+TC+3D+QC > TS+TC+3D+QC+logP > 3D > TC > TS > QC. The results reported in references [34, 35] are similar; the lowest correlation coefficients are obtained under consideration of QC descriptors.

Regarding molecular geometry, it can be optimized in the environment of receptor, if a protein interaction is under study. The environment, which causes an additional electrostatic field and includes hydrogen donors or acceptors, influences the optimization procedure. This approach is often applied in drug discovery, particularly in the case when the targeting bio-molecule is known (sometimes these approaches are addressed as 4- or 5-dimensional representations).

QSAR IN DRUG DISCOVERY

It is well known that the most revolutionary drugs were discovered by chance. It is a challenging idea to attack this problem systematically using computational tools; QSAR modeling is part of this strategy. We have witnessed a fast development of computational software and hardware, which supports the development in computational chemistry and pharmacy. The research strategy is focused on the following tasks: collecting of data and maintaining of databases, computational treatment of chemical structures and calculation of descriptors, QSAR modeling, collecting the information on biological targets and maintaining of target (protein) databanks, modeling the drug-target interaction [36]. Nowadays, large databanks, gathering data on structures, physico-chemical properties and biological activities of molecules, are publicly available. Some examples are ZINK, PubChem, NIH/CADD, *etc.* On the other hand, the variety of currently available chemometric tools can assist the searching for new drug candidates.

As case studies we present the QSAR modeling of anti-tubercular activity. The compounds of interest are fluoroquinolones. Some of the fluoroquinolones are potent antibiotics; however, the search for new more active analogues with less side effect is still on-going. Reference [37] presents QSAR models built on three data sets of flouroquinolones, which were compiled from NIAID database, available from the Internet. The modeled activity was expressed as the MIC (minimal inhibitory concentration expressed as mg ml^{-1}) value. For the structures, a set of descriptors were calculated with DRAGON and CODESSA software [38]. To select the most relevant descriptors, the descriptor set was filtered considering the correlation between descriptors and activity and inter-correlation between descriptors. Ten descriptors, appearing in most of the models, belong to the topological, electrotopological and constitutional classes. The study reported in reference [39] shows that electrotopological descriptors are more important in comparison to quantum chemical ones. Models were also tested with leave-one-out procedure and with the test set. In the further study [40] a combinatorial library of 5,590 compounds was generated by applying the virtual combinatorial synthetic pathway. The library was filtered in the first step with Lipinski-Veber rules to select the drug-like candidates. The remaining candidates were evaluated with QSAR models keeping the 15 candidates with the highest activity. A similar study is reported in reference [41], where authors developed QSAR models for activity against *Mycobacterium smegmatis* and *Mycobacterium fortuitum* with a set of 117 and 110 compounds, respectively. A pool of 1,056 descriptors was calculated with the Life Sciences Molecular Design Suite. It turned out that the topological and fragment descriptors play the most important role. Furthermore, authors created a virtual library of 5,280 compounds, which was screened by Lipinski rules and QSAR models. At the end, after the docking study in environment of the gyraze, they proposed seven candidates and one ‘winner’. In [42] authors studied the data set of 71 compounds using different modeling techniques like Ridge regression, principle component analysis and partial least squares. They calculated a comprehensive set of descriptors using POLLY, Triplet and MolconnZ programs, which are classified as topological, topochemical, and geometrical ones, respectively. The best models were found with topochemical descriptors followed by models with topostructural ones; both types of descriptors outperformed models built from 3D descriptors. In [43] authors report the QSAR study on quinoxaline compounds using a Partial Least Squares method. A pool of constitutional, physicochemical, electrostatic and topological descriptors was employed. The most robust model was found with constitutional descriptors. In [44] authors analyzed a large diverse data set of 4,100 compounds, which were classified as active or non-active. The structure was represented with a large number of descriptors, which were calculated with DRAGON and ADRIANA software. Descriptors were classified into 22 classes, however, the number of used descriptors exceeds 1,000, therefore the roles of individual descriptors cannot be evaluated. For classification, authors applied random forest and associative neural networks. The sensitivity, specificity, and precision of models for test set are Sn=0,75, Sp = 0.75, P=0.76, respectively.

QSAR IN ENVIRONMENTAL TOXICOLOGY

QSAR plays an important role in chemical regulation. For the risk assessment several toxicological, eco-toxicological and physico-chemical data must be known and in some cases QSAR can be used as an alternative method to expensive tests. Currently, there are more than 85,000 compounds in commerce in USA and EU and this number is growing rapidly by nearly 3,000 substances per year. For about 15% of these compounds, the reliable data necessary for risk assessment are known. In EU market the situation is similar. The Institute for Health and Consumer Protection reported that toxicity data are available for 15% - 70% of High Production Chemicals for different toxicological endpoints. The new European chemical legislation REACH (Registration, Evaluation, and Authorization of Chemicals) made a roadmap for registration of chemicals, which must be done for all chemicals on European market [45]. REACH recommends the using of QSAR methods as alternative method for gathering missing data for risk assessment, and for classification and labeling of compounds. QSAR models, which are used in regulatory processes, must meet some criteria. The prediction must be reliable, or at least, estimation on reliability must be provided. The models used for regulatory purposes must be transparent in such a way that the experts could eventually rebuild them. To establish a platform for discussions about the QSAR models, the OECD adopted the document: Principles for validation of (Q)SAR models used for regulatory purposes [46], which basically consists of five principles.

Principle 1: Defined endpoint. Endpoint refers to any physicochemical, biological or environmental effect. The intent of this principle is to ensure transparency in the endpoint being predicted, since a given endpoint could be determined by different experimental protocols and under different experimental conditions.

Principle 2: Unambiguous algorithm. The aim of this principle is to ensure the transparency of modeling algorithms.

Principle 3: Definite applicability domain. QSAR models are inevitably associated with limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of actions. This principle seeks for determining the kind of structures, properties and mechanisms of action where a given QSAR model yields accurate results.

Principle 4: Measure of goodness-of-fit, robustness and predictivity. This principle expresses the need to provide two types of information: the internal performance of a model (as expressed as goodness-of-fit and robustness) and the predictivity of model using an appropriate test set.

Principle 5: Mechanistic interpretation, when possible. The intent of this principle is to ensure that there is an assessment of the possibility of a mechanistic association between the descriptors used in a model and the endpoint being predicted, and that any association is documented.

CAESAR MODELS

CAESAR [47] is a project funded by the European Commission dedicated to developing *in silico* models for prediction of five endpoints relevant for REACH legislation. The endpoints are: bioconcentration in fish, mutagenicity, carcinogenicity, developmental toxicity and skin sensitization. The models were developed according to the OECD principles for (Q)SAR models on high quality data sets compiled following the OECD or US EPA standards. For developing models, advanced computational techniques were used including programs for calculation of molecular descriptors and techniques for descriptor selection. The models are publicly available in the Internet and they require as input only the SMILES code of molecular structures. For the bioconcentration factor the prediction is expressed as a real number, for the other four endpoints the prediction is in form of binary classifications. Besides classifications, CAESAR provides two additional pieces of information: a comment if the descriptors are out of the range and six compounds belonging to the training set for each prediction, which are the most similar to the evaluated one.

Bioconcentration Factor

The bioconcentration factor (BCF) describes the readiness of chemicals to concentrate in organisms when the compounds are present in the environment. It is a required eco-toxicological parameter for chemical regulation, for example REACH requires the BCF for all compounds produced or imported over 100 tons per year.

The CAESAR model for BCF is based on a database of 378 compounds [48]. The data set consists of the following chemical classes: alkanes, alkenes, mono and diaromatic hydrocarbons, polycyclic aromatic hydrocarbons, polychlorinated dibenzofurans, polychlorinated dibenzodioxins, polychlorinated biphenyls, chloroalkanes, chloroalkenes, and halogenated aromatic compounds. The CAESAR model was built with a radial basis function for neural networks using eight descriptors including log P, topological, electrotopological and geometrical ones. In most BCF models the log P plays the fundamental role. Indeed, the BCF is strongly correlated with log P, which means that lipophilic compounds strongly tend to accumulate in organisms. The model was tested showing a correlation coefficient r = 0.8 for the test set. One can find more details in reference [49].

Mutagenicity

Mutagenicity is the ability of substances to cause cell mutation. It is directly connected to carcinogenicity and developmental toxicity. One of the *in vitro* tests for determination of mutagenicity is the Ames test performed on Salmonella [50]. The CAESAR model is built on a large data set of 4,204 compounds with their Ames test results, which were extracted from an original set reported by Kazius *et al.* [51]. Compounds belong to diverse chemical classes. In reference [51] authors identified eight toxicophores, which may be indicators on mutagenicity (aromatic nitro, aromatic amine, three-membered heterocycle, nitroso, unsubstituted heteroatom-bonded heteroatom, azo-type, aliphatic halide, polycyclic aromatic system). The modeling technique is a hybrid system combining a support vector algorithm for classification and a rule based system checking for structural alerts. The mutagenicity is binary expressed as non-positive or positive. For all structures the descriptor pool was calculated with the MDL software; the BestFirst algorithm from WEKA was applied to select the 25 most relevant descriptors. Among them are log P, two topological indices (the cyclomatic number and the Bonchev-Trinajstić mean information content), one electro-topological index (the minimum E-state value for all atoms) and 21 counts of particular E-state fragments. The achieved accuracy was 82%, which was close to 85% of reliability of the experimental Ames test; false negative rate was 10%. Further details are described in reference [52].

Carcinogenicity

Carcinogenicity is one of the toxicological endpoints required in chemical regulation. Carcinogenesis is a very complex biological process, which indeed includes many different mechanisms. Therefore, it is not expected that a QSAR model can explain the mechanism of carcinogenesis. QSAR models are just tools giving insight about carcinogenicity based on the similarity of particular compounds to non-carcinogenic or carcinogenic compounds.

The details of CAESAR carcinogenicity models are reported in [53]. The models were built on a set of 805 non-congeneric compounds extracted from the Carcinogenic Potency Data base (CPDBAS). In terms of chemical classification the compounds belong to diverse chemical classes. The Hybride Selection Algorithm developed from BioChemics Consulting SAS (BCX) was applied to select eight descriptors from a set of 254 MDL descriptors. Furthermore, a cross correlation matrix, multicolinearity and fisher ratio technique was applied to select 12 descriptors from a set of 835 DRAGON descriptors. As modeling technique for classification the counter propagation artificial neural networks were applied. A compound was classified as non-carcinogenic (or non-positive) when the results obtained in mice and rat tests were negative, and in contrary, it was classified as carcinogenic when at least one test showed positive response.

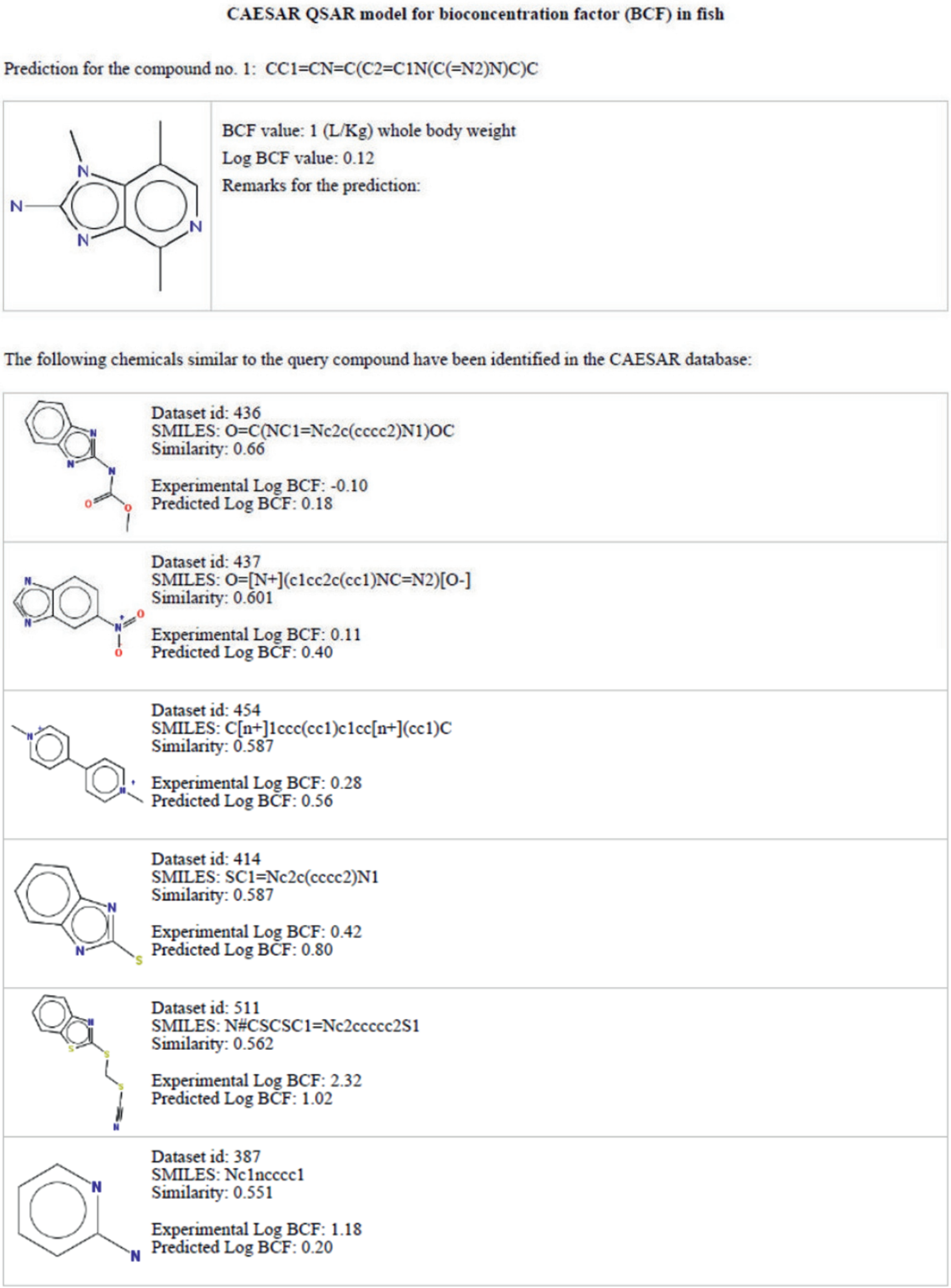
Developmental Toxicity

Developmental toxicity (DT) is an endpoint required for risk assessment. A detailed description of its tests is given in OECD documentation [54]. These tests are some of the most expensive ones and for a single substance several thousand animals must be sacrificed. Alternative methods for assessment of DT include (Q)SAR methods, intelligent strategies and priority setting [55, 56].

The details of CAESAR program for DT are presented in reference [57]. In the CAESAR model, a compound is binary classified as non-developmental toxicant if it belongs to the FDA category A or B, or, as developmental toxicant, if it belongs to the FDA category C or D. Two models are implemented, the first one is the application of a random forest algorithm (13 descriptors), and the second one is the adaptive fuzzy partition algorithm (six descriptors). The model is built on Arena data set, which includes 292 compounds, from which 41% show some evidence on developmental toxicity and 59% does not [58]. The data set was developed by combining data from Teratogen Information System (TERIS) and the Food and Drug Administration data. Regarding chemical classification, the compounds belong to diverse chemical classes. Descriptors were calculated with DRAGON, T.E.S.T. and MDL programs. At the end six or 13 molecular descriptors were selected using the WEKA (Waikato Environment for Knowledge Analysis) software. The selected descriptors belong to topological and electro-topological classes. Most of them are eigenvalues of topological matrices weighted with physico-chemical parameters, or E-state indices related to particular fragments. The accuracy for leave-one-out cross validation test for random forest and adaptive fuzzy partition algorithm was 77% and 72%, respectively.

Skin Sensitization

Skin sensitization is the ability of a substance to induce allergic reaction after skin contact. For the experimental determination of skin sensitization the OECD recommends the Bühler test on guinea pigs [59] or the mice local lymph node assay [60]. The CAESAR models were constructed on a data set of 209 compounds selected from the Gerberick data set of 211 compounds [61]. The compounds are classified regarding the local lymph node assay into extreme sensitizers, strong sensitizers, moderate sensitizers, weak sensitizers, or no sensitizers. In terms of chemical classification they belong to aldehydes, ketones, aromatic amines, quinones, and acrylates, which may be active due to different mechanisms. From a pool of 502 descriptors, which were calculated with DRAGON, the seven descriptors were selected using a combination of genetic algorithm and stepwise regression. Selected descriptors belong to constitutional and topological classes. The binary classification model was built with adaptive fuzzy partition algorithm and for test set the accuracy of 90% was achieved. Additionally, a model based on adaptive fuzzy partition has been presented. The details of CAESAR model for skin sensitization are given in reference [62].



CAESAR QSAR model for bioconcentration factor (BCF) in fish

Prediction for the compound no. 1: CC1=CN=C(C2=C1N(C(=N2)N)C)C

BCF value: 1 (L/Kg) whole body weight
Log BCF value: 0.12
Remarks for the prediction:

The following chemicals similar to the query compound have been identified in the CAESAR database:

Dataset id: 436
SMILES: O=C(NC1=Nc2c(cccc2)N1)OC
Similarity: 0.66

Experimental Log BCF: -0.10
Predicted Log BCF: 0.18

Dataset id: 437
SMILES: O=[N+](c1cc2c(cc1)NC=N2)[O-]
Similarity: 0.601

Experimental Log BCF: 0.11
Predicted Log BCF: 0.40

Dataset id: 454
SMILES: C[n+]1ccc(cc1)c1cc[n+](cc1)C
Similarity: 0.587

Experimental Log BCF: 0.28
Predicted Log BCF: 0.56

Dataset id: 414
SMILES: SC1=Nc2c(cccc2)N1
Similarity: 0.587

Experimental Log BCF: 0.42
Predicted Log BCF: 0.80

Dataset id: 511
SMILES: N#CSCSC1=Nc2ccccc2S1
Similarity: 0.562

Experimental Log BCF: 2.32
Predicted Log BCF: 1.02

Dataset id: 387
SMILES: Nc1ncccc1
Similarity: 0.551

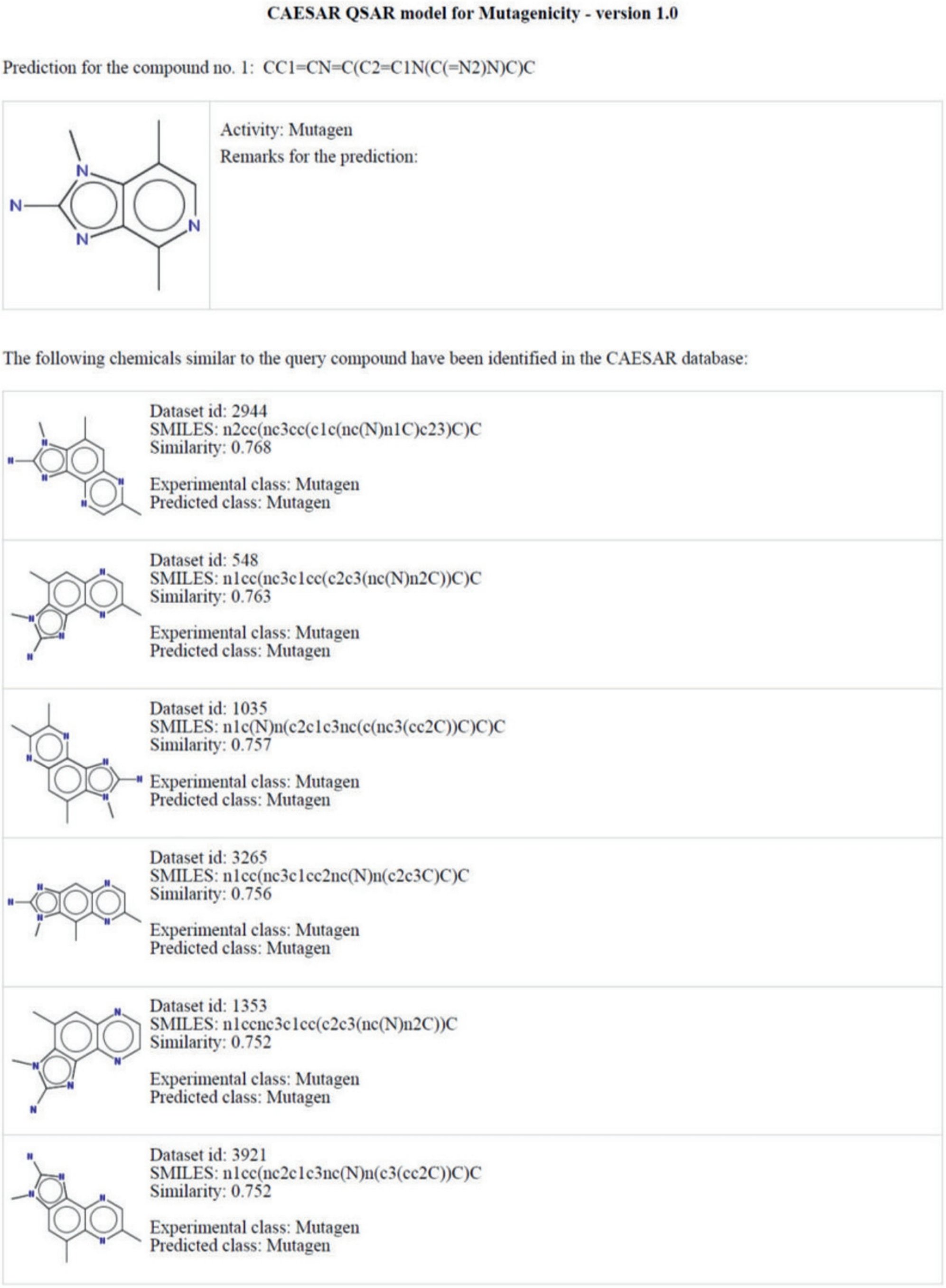
Experimental Log BCF: 1.18
Predicted Log BCF: 0.20

Figure 1: CAESAR prediction for BCF for 1,4,7-trimethylimidazo[4,5-c]pyridin-2 amine.

As an example we present the predictions of five endpoints for 1,4,7-trimethylimidazo[4,5-c]pyridin-2-amine. Fig. **(1)** shows the prediction of its BCF and the six most similar compounds from the training set; for each of them the predicted and experimental values are reported. The correlation coefficient between experimental and predicted values for the six compounds is r = 0.600. Fig. **(2)** shows the prediction for mutagenicity of 1,4,7-trimethylimidazo[4,5-c]pyridin-2-amine. The CAESAR prediction gives the six most similar compounds regarding the target substance together with their similarity indices and their predicted and experimental classification. Carcinogenicity prediction is shown in Fig. **(3)**; the compound is predicted as 15.1% positive (carcinogenic) and 84.9% as non-positive (non-carcinogenic). The predictions for the six most similar compounds are: TP = 4, FN = 1, TN = 1. Fig. **(4)** shows the prediction for skin sensitization. The compound is predicted as 96.5% active (skin sensitizer) and as 3.5% inactive (non-sensitizer). The confusion matrix is TP=3, FP=1, FN=1, TN=1. Fig. **(5)** shows the developmental toxicity. The compound is predicted as toxic. The confusion matrix for the most similar six compounds to the target substance is: TP=3, FP=1, FN=0, TN=2.

CONCLUDING REMARKS

In the chapter we present different aspects of structural descriptors. It is to emphasize that in a selected QSAR model the descriptors carry the entire information about molecular structures. This further means that the descriptors together with the training set, define the domain of the model and the similarity relationships among structures. In some cases they may indicate the mechanism of the predicted property (activity), as shown in the presented case study on drug research. The second example shows five predictions performed with the CAESAR programs for the same structure. Figs. (**1** to **5**) show for each prediction the six most similar compounds from the actual training set. 'The six most similar compounds' can give an insight into the mechanism of property (activity) and also show how good the predicted compound fits into domain of the model.



CAESAR QSAR model for Mutagenicity - version 1.0

Prediction for the compound no. 1: CC1=CN=C(C2=C1N(C(=N2)N)C)C

Activity: Mutagen
Remarks for the prediction:

The following chemicals similar to the query compound have been identified in the CAESAR database:

Dataset id: 2944
SMILES: n2cc(nc3cc(c1c(nc(N)n1C)c23)C)C
Similarity: 0.768

Experimental class: Mutagen
Predicted class: Mutagen

Dataset id: 548
SMILES: n1cc(nc3c1cc(c2c3(nc(N)n2C))C)C
Similarity: 0.763

Experimental class: Mutagen
Predicted class: Mutagen

Dataset id: 1035
SMILES: n1c(N)n(c2c1c3nc(c(nc3(cc2C))C)C)C
Similarity: 0.757

Experimental class: Mutagen
Predicted class: Mutagen

Dataset id: 3265
SMILES: n1cc(nc3c1cc2nc(N)n(c2c3C)C)C
Similarity: 0.756

Experimental class: Mutagen
Predicted class: Mutagen

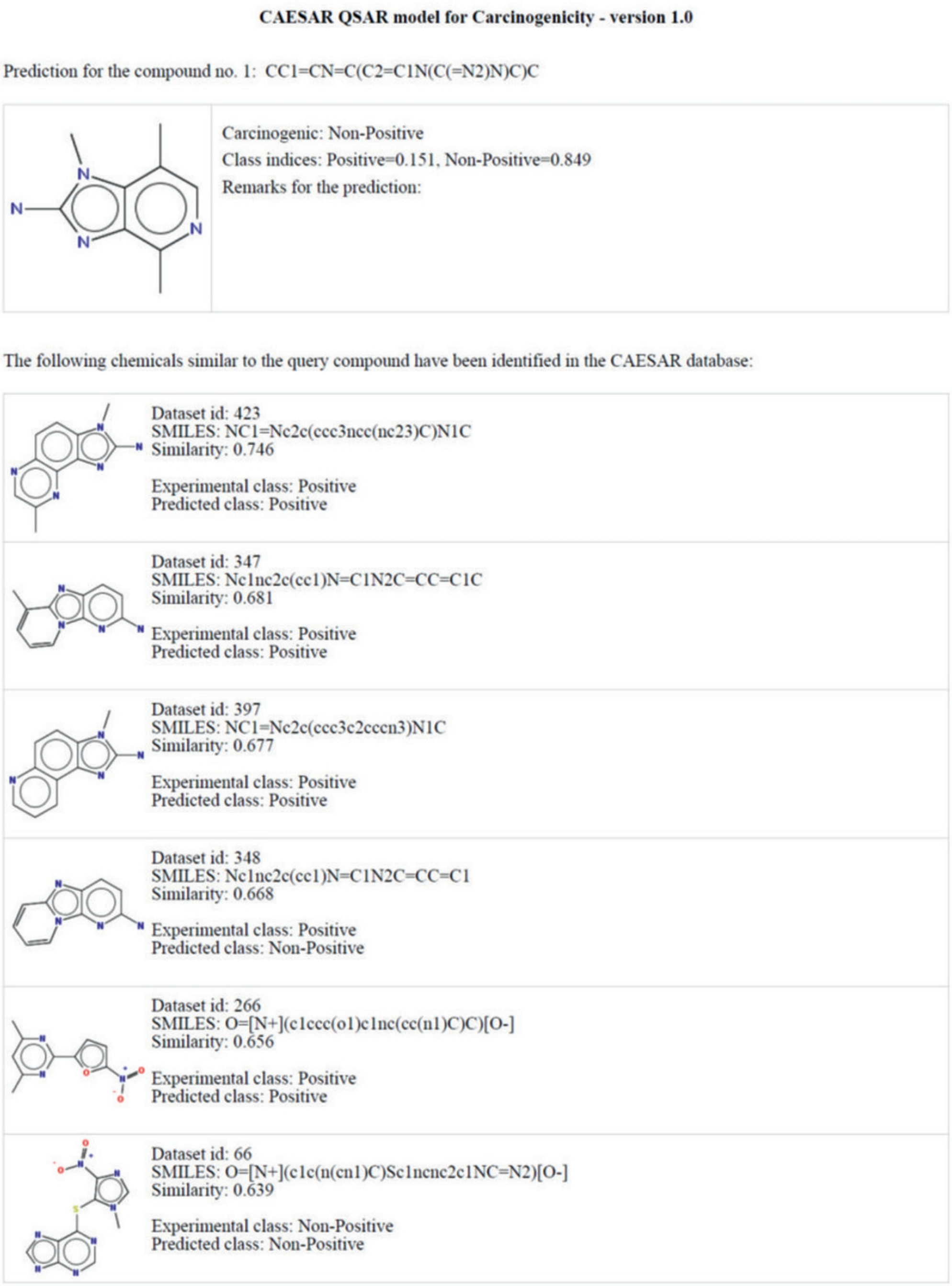
Dataset id: 1353
SMILES: n1ccnc3c1cc(c2c3(nc(N)n2C))C
Similarity: 0.752

Experimental class: Mutagen
Predicted class: Mutagen

Dataset id: 3921
SMILES: n1cc(nc2c1c3nc(N)n(c3(cc2C))C)C
Similarity: 0.752

Experimental class: Mutagen
Predicted class: Mutagen

Figure 2: CAESAR prediction of mutagenicity for 1,4,7-rimethylimidazo[4,5-c]pyridin-2 amine.



CAESAR QSAR model for Carcinogenicity - version 1.0

Prediction for the compound no. 1: CC1=CN=C(C2=C1N(C(=N2)N)C)C

Carcinogenic: Non-Positive
Class indices: Positive=0.151, Non-Positive=0.849
Remarks for the prediction:

The following chemicals similar to the query compound have been identified in the CAESAR database:

Dataset id: 423
SMILES: NC1=Nc2c(ccc3ncc(nc23)C)N1C
Similarity: 0.746

Experimental class: Positive
Predicted class: Positive

Dataset id: 347
SMILES: Nc1nc2c(cc1)N=C1N2C=CC=C1C
Similarity: 0.681

Experimental class: Positive
Predicted class: Positive

Dataset id: 397
SMILES: NC1=Nc2c(ccc3c2cccn3)N1C
Similarity: 0.677

Experimental class: Positive
Predicted class: Positive

Dataset id: 348
SMILES: Nc1nc2c(cc1)N=C1N2C=CC=C1
Similarity: 0.668

Experimental class: Positive
Predicted class: Non-Positive

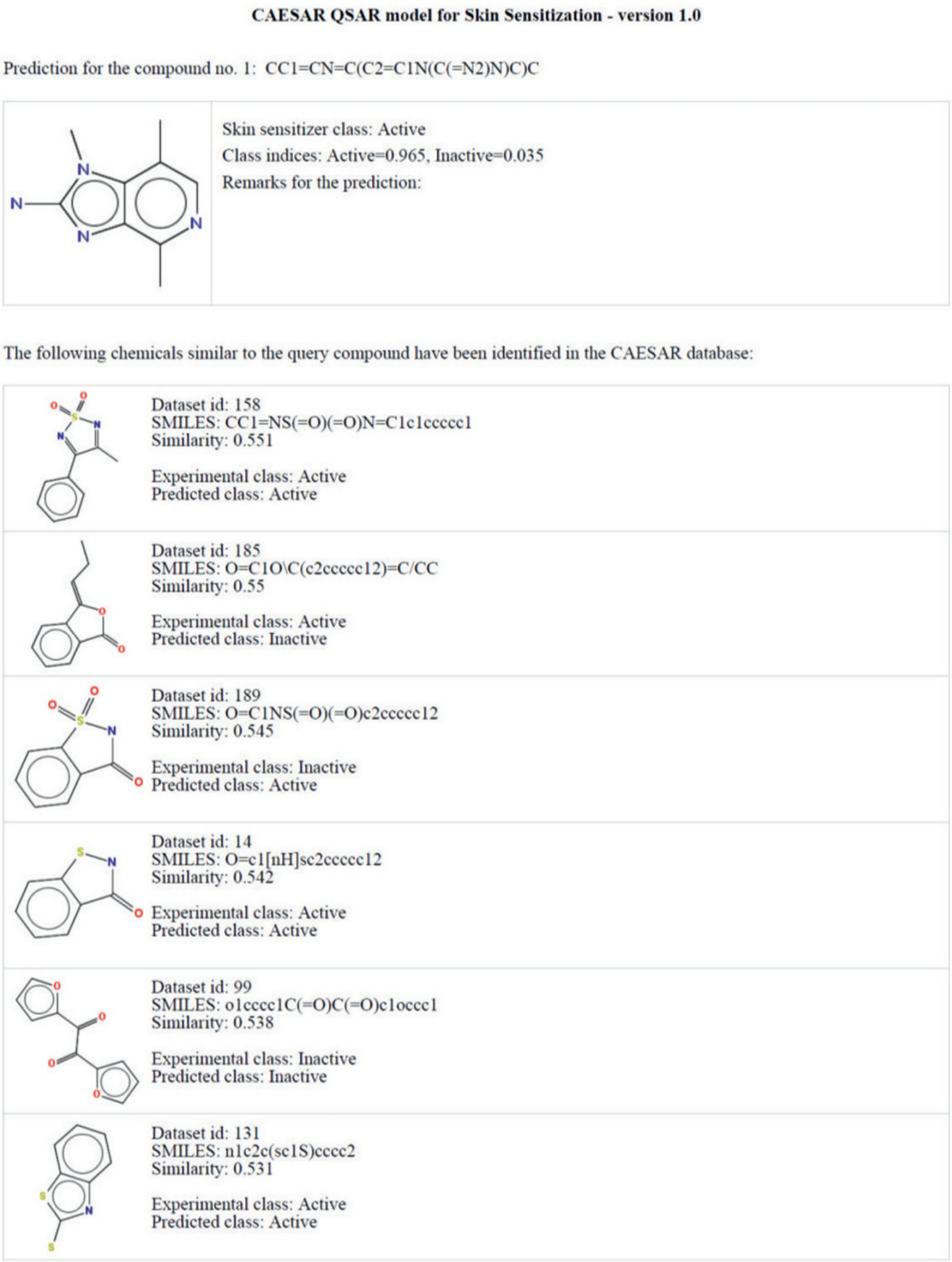
Dataset id: 266
SMILES: O=[N+](c1ccc(o1)c1nc(cc(n1)C)C)[O-]
Similarity: 0.656

Experimental class: Positive
Predicted class: Positive

Dataset id: 66
SMILES: O=[N+](c1c(n(cn1)C)Sc1ncnc2c1NC=N2)[O-]
Similarity: 0.639

Experimental class: Non-Positive
Predicted class: Non-Positive

Figure 3: CAESAR prediction for carcinogenicity for 1,4,7-trimethylimidazo[4,5-c]pyridin-2 amine.



CAESAR QSAR model for Skin Sensitization - version 1.0

Prediction for the compound no. 1: CC1=CN=C(C2=C1N(C(=N2)N)C)C

Skin sensitizer class: Active
Class indices: Active=0.965, Inactive=0.035
Remarks for the prediction:

The following chemicals similar to the query compound have been identified in the CAESAR database:

Dataset id: 158
SMILES: CC1=NS(=O)(=O)N=C1c1ccccc1
Similarity: 0.551

Experimental class: Active
Predicted class: Active

Dataset id: 185
SMILES: O=C1O\C(c2ccccc12)=C/CC
Similarity: 0.55

Experimental class: Active
Predicted class: Inactive

Dataset id: 189
SMILES: O=C1NS(=O)(=O)c2ccccc12
Similarity: 0.545

Experimental class: Inactive
Predicted class: Active

Dataset id: 14
SMILES: O=c1[nH]sc2ccccc12
Similarity: 0.542

Experimental class: Active
Predicted class: Active

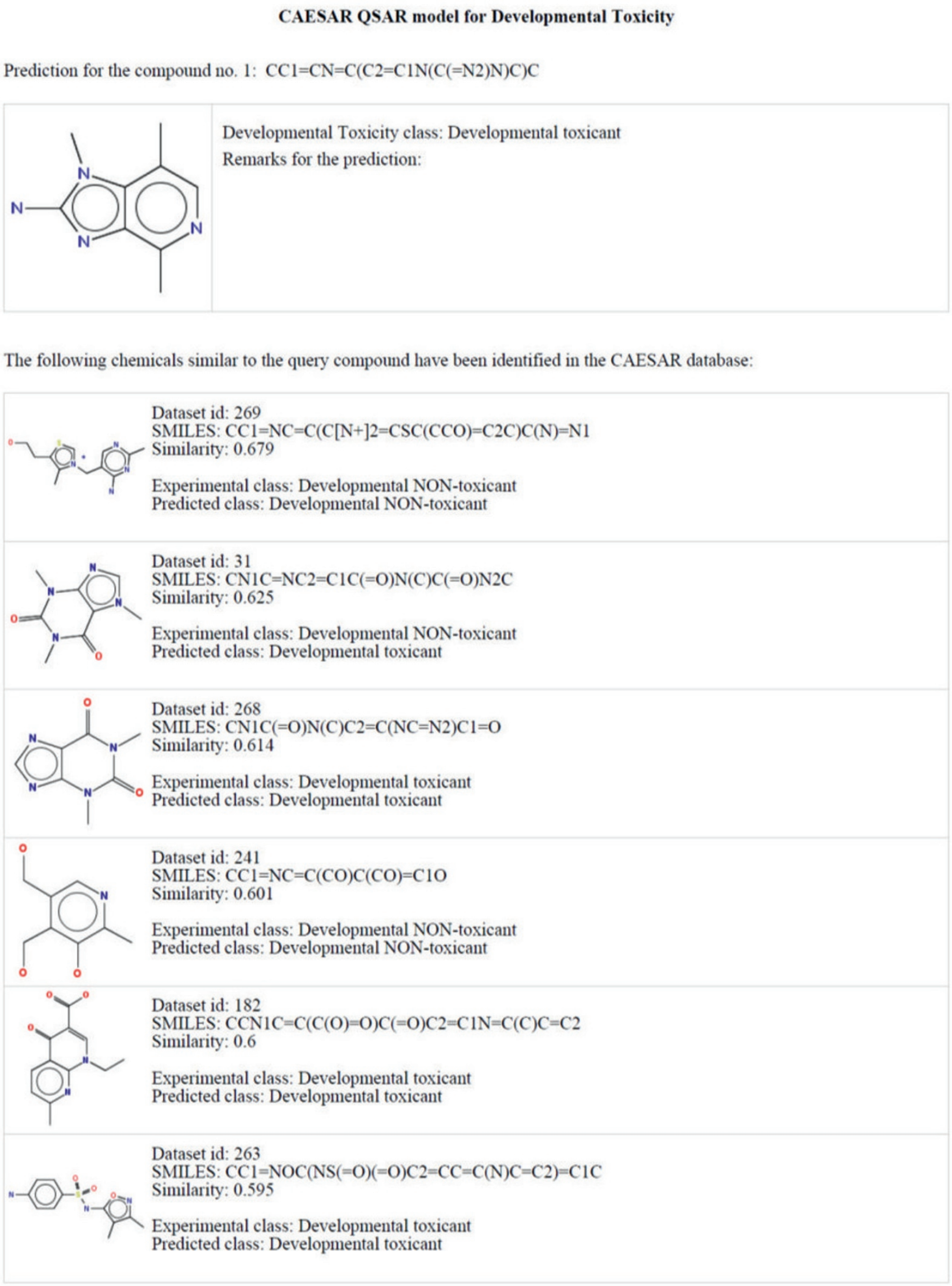
Dataset id: 99
SMILES: o1cccc1C(=O)C(=O)c1occc1
Similarity: 0.538

Experimental class: Inactive
Predicted class: Inactive

Dataset id: 131
SMILES: n1c2c(sc1S)cccc2
Similarity: 0.531

Experimental class: Active
Predicted class: Active

Figure 4: CAESAR prediction for skin sensitization for 1,4,7-trimethylimidazo[4,5-c]pyridin-2 amine.



CAESAR QSAR model for Developmental Toxicity

Prediction for the compound no. 1: CC1=CN=C(C2=C1N(C(=N2)N)C)C

Developmental Toxicity class: Developmental toxicant
Remarks for the prediction:

The following chemicals similar to the query compound have been identified in the CAESAR database:

Dataset id: 269
SMILES: CC1=NC=C(C[N+]2=CSC(CCO)=C2C)C(N)=N1
Similarity: 0.679

Experimental class: Developmental NON-toxicant
Predicted class: Developmental NON-toxicant

Dataset id: 31
SMILES: CN1C=NC2=C1C(=O)N(C)C(=O)N2C
Similarity: 0.625

Experimental class: Developmental NON-toxicant
Predicted class: Developmental toxicant

Dataset id: 268
SMILES: CN1C(=O)N(C)C2=C(NC=N2)C1=O
Similarity: 0.614

Experimental class: Developmental toxicant
Predicted class: Developmental toxicant

Dataset id: 241
SMILES: CC1=NC=C(CO)C(CO)=C1O
Similarity: 0.601

Experimental class: Developmental NON-toxicant
Predicted class: Developmental NON-toxicant

Dataset id: 182
SMILES: CCN1C=C(C(O)=O)C(=O)C2=C1N=C(C)C=C2
Similarity: 0.6

Experimental class: Developmental toxicant
Predicted class: Developmental toxicant

Dataset id: 263
SMILES: CC1=NOC(NS(=O)(=O)C2=CC=C(N)C=C2)=C1C
Similarity: 0.595

Experimental class: Developmental toxicant
Predicted class: Developmental toxicant

Figure 5: CAESAR prediction for developmental toxicity for 1,4,7-trimethylimidazo[4,5-c]pyridin-2 amine.

ACKNOWLEDGEMENTS

Author thanks Agency of Research of R. Slovenia (ARRS) for the financial support through the Grant P1-0017.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

REFERENCES

[1] Valkova, I.; Vračko, M.; Basak, S.C. Modeling of structure-mutagenicity relationship: counter propagation neural network approach using calculated structural descriptors. *Anal. chim. acta*., **2004**, *509,* (2), 179-186.

[2] Hawkins, D.M.; Basak, S.C.; Mills D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.*, **2003**, 43(2), 579-586.

[3] Novič, M.; Zupan, J. A new general and uniform structure representation. In: *Software Entwicklung in der Chemie 10*; Gasteiger, J. Ed.; GDCh: Frankfurt am Main, **1996**; pp. 47-58.

[4] Free, S.M.; Wilson, J.W. A mathematical contribution to structure-activity study. *J. Med. Chem.,* **1964**, *7,* 395-399.

[5] Ursu, O.; Oprea, T.I. Model-free drug-likeness from fragments. *J. Chem. Inf. Model.*, **2010**, *50,* 1387-1394.

[6] Du, Q.; Huang, R.; Wei, Y.; Pang, Z.; Du, L.; Chou, K. Fragment-based Quantitative Structure-activity relationship (FB-QSAR) for fragment based drug design. J. Comput. Chem., **2009**, 30(2), 295-304.

[7] Catana, C. Simple idea to generate fragment and pharmacophore descriptors and their implications in chemical informatics. *J. Chem. Inf. Model.*, **2009**, *49,* 543-548.

[8] Toporov, A.A.; Toporova, A.P.; Benfenati, E.; Manganaro, A. QSAR modelling of the toxicity to Tetrahymena pyriformis by balance of correlations. *Mol. Divers.,* **2010**, *14,* 821-827.

[9] Benigni, R.; Bossa, C. Structure alerts for carcinogenicity, and the Salmonella assay system: A novel insight through the chemical related databases technology. *Mutat. Res.*, **2008**, *659,* 248-261.

[10] Ashby, J. Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ. Mutagen.*, **1985**, *7,* 919-921.

[11] Kazius, J.; McGuire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, **2005**, *48,* 312-320.

[12] Helma, C.; Cramer, T.; Kramer, S.; De Readt, L. Data mining and machine learning techniques for the identification of mutagenicity including substructures and structure activity relationship of noncongeneric compounds. *J. Chem. Inf. Comp. Sci.*, **2004**, *44,* 1402-1411.

[13] Benigni, R.; Bossa, C. Structure alerts for carcinogenicity, and the Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutat. Res. Revs.*, **2008**, *659,* 248-261.

[14] Hansch, C.; Maloney, P.P.; Fujita, T.; Muir, R.M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nat.,* **1962**, *194,* 178-80.

[15] Eros, D.; Kovesdi, I.; Orfi, L.; Takacs-Novak, K.; Acsady, G.; Keri, G. Reliability of logP predictions based on calculated molecular descriptors: A critical review. *Current Med. Chem.*, **2002**, *9,* 1819-1829.

[16] Randič, M. On characterization of molecular branching. *J. Am. Chem. Soc.,* **1975**, *97,* 6609-6615.

[17] Balasubramanian, K.; Basak, S. C. Characterisation of isospectral graphs using graph invariants and derived ortogonal parameters. *J. Chem. Inf. Comput. Sci.*, **1998**, *38(3),* 367-373.

[18] Basak, S.C. Use of molecular complexity indices in predictive pharmacology and toxicology: a QSAR approach. *Med. Sci. Res.*, **1987**, *15*(11), 605-609.

[19] Natarajan, R.; Basak, S.C. Numerical descriptors for the characterization of chiral compounds and their applications in modeling biological and toxicological activities. *Curr. Topics Med. Chem.*, **2011**, *11*(7), 771-787.

[20] Balaban, A.T. A personal view about topological indices for QSAR/QSPR. In: QSAPR/QSAR studied by molecular descriptors; Diudea, M. V. Ed.; Nova Science Publisher, Inc.: Huntington, New York, **2001**; pp. 1-31

[21] Kier, L.B.; Hall, L.H. Molecular connectivity. Part 7. Specific treatment of heteroatoms. *J. Phar. Sci.*, **1976**, *65,* 1806-1809.

[22] Roy, K.; Gosh, G. QSTR with extended topochemical atom indices. Part 5: Modeling of the acute toxicity of phenylsulfonyl carboxylates to Vibrio fischeri using genetic function approximation *Bioorg. Med. Chem.,* **2005**, *13,* 1185-1194.

[23] Hückel, E. Quantentheoretische Beitrage zum Benzolproblem, I. Die Elektronenkonfiguration des Benzols und vervandter Verbindungen. *Zeit. für Pysik*, **1931**, *70,* 204.

[24] Lovasz, L.; Pelikan, J. On the eigenvalues of trees. *Period. Math. Hun.*, **1973**, *3,* 175-182.

[25] Randić, M.; Plavšić, D.; Razinger, M. Double invariants. *MATCH*, **1997**, *35,* 243-259. Randić, M.; Vračko, M.; Novič, M. Eigenvalues as molecular descriptors. In: QSAPR/QSAR studies by molecular descriptors. Ed. Diudea M. V., Nova Science Publishers, Inc. Huntington, New York, **2001**, pp. 147-211.

[26] Katritzky, A.R.; Lobanov, V. S.; Karelson.; M. Quantum chemical descriptors in QSAR/QSPR studies. *Chem. Rev.*, **1996**, *96,* 1027-1043.

[27] Lewars, E.G. Computational Chemistry, Introduction to the Theory and Applications of Molecular and Quantum Mechanics, Second Edition, Springer, Dordrecht, Heidelberg, London, New York, **2011**.

[28] Vračko, M.; Szymoszek, A.; Barbieri. P. Structure-mutagenicity study of 12 Trimethylimidazopyridine isomers using orbital energies and spectrum-like representation as descriptors. *J. Chem. Inf. Comput. Sci.,* **2004**, *44,* 352-358.

[29] Girones, X.; Amat, L.; Robert, D.; Carbo-Dorca, R. Use of electron-electron repulsion energy as a molecular descriptor in QSAR and QSPR studies. *J. Comp.-Aided Mol. Design*, **2000**, *14,* 477-485.

[30] Netzeva, T.I.; Aptula, A. O.; Benfenati, E.; Cronin, M. T. D., Gini, G.; Lessigiarska, I.; Maran, U.; Vračko, M.; Schüürmann, G. Description of the electronic structure of organic

chemicals using semiempirical and ab initio methods for development of toxicological QSARs. *J. Chem. Inf. Mod.*, **2005**, *45,* 106-114.

[31] Leardi, R., Ed. *Nature-inspired methods in Chemometrics: Genetic Algorithms and Artificial Neural Networks*; Elsevier: Amsterdam-Boston-Heidelberg-London-New York-Oxford-Paris-San Diego-San Francisco-Singapore-Sydney-Tokyo, **2003**.

[32] Novič, M.; Vračko, M. Comparison of spectrum-like representation of 3D chemical structure with other representations when used for modelling biological activity. *Chemom. Intell. Lab. Syst.*, **2001**, *59,* 33-44.

[33] Schuur, J. H.; Selzer, P., Gasteiger, J. The coding of three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.,* **1996**, 36(2), 334-344.

[34] Vracko, M.; Mills, D.; Basak, S. C. Structure-mutagenicity modelling using counter propagation neural network. *Environ. Toxicol. Pharmacol.,* **2004**, *16,* 25-36.

[35] Basak, S.C.; Mills, D.R.; Balaban, A.T.; Gute, B.D. Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: a hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, **2001**, *41,* 671-678.

[36] Warr, W. A. Some Trends in Chem(o)informatics. In: *Chemoinformatics and Computational Chemical Biology*; Bajorath, J. Ed.; Humana Press (Springer): New York, **2010**, pp.1-37.

[37] Minovski, N.; Vračko, M.; Šolmajer, T. Quantitative structure-activity relationship study of antitubercular fluoroquinolones. *Mol. Divers.,* **2011**, 15(2), 417-426.

[38] Todeschini, R.; Consonni, V. Handbuch of molecular descriptors. Wiley-VCH, Weinheim, **2000**.

[39] Minovski, N.; Jezierska-Mazzarello, A.; Vračko, M.; Šolmajer, T. Investigation of 6-fluoroquinolones activity against Mycobacterium tuberculosis using theoretical molecular descriptors: a case study. *Cent. Eur. J. Chem.*, **2011**, 9(5), 855-866.

[40] Minovski, N.; Šolmajer, T. Chemometrical exploration of combinatorially generated drug-like space of 6-fluoroquinolone analogs: a QSAR study. *Acta Chim. Slov.*, **2010**, *57,* 529-591.

[41] Ghosh, P.; Bagchi, M.C. Anti-tubercular drug designing by structure based screening of combinatorial libraries. *J. Mol. Model.*, **2011**, 17(7), 1607-1620.

[42] Bagchi, M. C.; Mills, D.; Basak, S.C. Quantitative structure-activity relationship (QSAR) studies of quinolone antibacterials against M-fortuitum and M-smegmatis using theoretical molecular descriptors. *J. Mol. Model.,***2007**, *13,* 111-120.

[43] Ghosh, P.; Vracko,M.; Chattopadhyay, A. K.; Bagchi, M.C. On application of constitutional descriptors for merging of quinoxaline data sets using linear statistical methods. *Chem. Biol. Drug Des.*, **2008**, *72,* 155-162.

[44] Kovalishyn, V.; Aires-de-Sousa, J.; Ventura, C.; Leitao, R.E.; Martins, F. QSAR modeling of antitubercular activity of diverse organic compounds. *Chemom. Intelligent Lab. Syst.*, **2011**, 107(1), 69-74.

[45] http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm

[46] OECD: Guidance document on the validation of (Q)SAR models. Paris, France. Organization foe Economic Cooperation and Safety Publications. *Series on testing and assessment.* **2007**, *69,* 154.

[47] Benfenati, E. The CAESAR project for *in silico* models for the REACH legislation. *Chem. Cent. J.*, **2010**, 4(Suppl 1)I1.

[48] Dimitrov, S.; Dimitrova, N.; Parkerton, T.; Comber, M.; Bonnell, M.; Mekenyan, O. Baseline model for identifying the bioaccumulation potential of chemicals. *SAR QSAR Environ. Res.,* **2005**, *16,* 531-554.

[49] Lombardo, A.; Roncaglioni, A.; Boriani, E.; Milan, C.; Benfenati, E. Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. *Chem. Cent. J.,* **2010**, 4(Suppl 1)S1.

[50] Ames, B. N. The detection of environmental mutagens and potential. *Cancer*, **1984**, *53,* 2030-2040.

[51] Kazius, J.; Mcguire, R.; Bursi, R. Derivation and validation of toxicophores for mutagenicity prediction. *J. Med. Chem.*, **2005**, *48,* 312-320.

[52] Ferrari, T.; Gini, G. An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. *Chem. Cent. J.,* **2010**, 4(Suppl 1)S2.

[53] Fjodorova, N.; Vračko, M.; Novič, M.; Roncaglioni, A.; Benfenati, E. New public QSAR model for carcinogenicity. *Chem. Cent. J.*, **2010**, 4(Suppl 1)S3.

[54] OECD, Draft Guidance Document on Mammalian Reproductive Toxicity Testing and Assessment. *Series on Testing and Assessment*, ***43,*** OECD Publication Office, Paris, France, **2007**.

[55] Cronin, M.T.D.; Worth, A.P. (Q)SARs for predicting effects relating to reproductive toxicity. *QSAR Comb. Sci.,* **2008**, *27,* 91-100.

[56] Bolčič-Tavčar, M.; Vračko, M. Assessing the reproductive toxicity of some (con)azole compounds using a structure-activity relationship approach. *SAR & QSAR Environ. Res.*, **2009**, *20,* 711-723.

[57] Cassano, A.; Manganaro, A.; Martin, T.; Young, D.; Piclin, N.; Pintore, M. CAESAR models for developmental toxicity. *Chem. Cent. J.*, **2010**, 4(Suppl 1)S4.

[58] Arena, V.C.; Sussman, N.B.; Mazumdar, S.; Yu, S.; Macina, O.T. The utility of structure-activity relationship (SAR) models for prediction and covariate selection in developmental toxicity: comparative analysis of logistic regression and decision tree models. *SAR & QSAR Environ. Res.*, **2004**, *15,* 1-18.

[59] OECD, OECD guideline for testing of chemicals, OECD **1992**, 406.

[60] OECD, OECD guideline for testing of chemicals, OECD **2002**, 429.

[61] Geberick, G.F.; Ryan, C.A.; Kern, P.S.; Schaltter, H.; Dearman, R.J.; Kimber, I.; Patlewicz, G.Y.; Basketter, D.A. Compilation of historical local node data for evaluation of skin sensitization alternative methods. *Dermatitis*, **2005**, *16,* 157-202.

[62] Chaudhry, Q.; Piclin, N.; Cotterill, J.; Pintore, M.; Price, N.R.; Chretien, J.R.; Roncaglioni, A. Global QSAR models of skin sensitisers for regulatory purposes. *Chem. Cent. J.*, **2010**, 4(Suppl 19)S5.

CHAPTER 11

Current Landscape of Hierarchical QSAR Modeling and its Applications: Some Comments on the Importance of Mathematical Descriptors as well as Rigorous Statistical Methods of Model Building and Validation

Subhash C. Basak[1,*] and Subhabrata Majumdar[2]

[1]International Society of Mathematical Chemistry, 1802 Stanford Avenue, Duluth, MN 55811 and UMD-NRRI, 5013 Miller Trunk Highway, Duluth MN 55811, USA and [2]School of Statistics, University of Minnesota Twin Cities, 224 Church Street SE, Minneapolis, MN 55455, USA

Abstract: Mathematical chemistry or more accurately discrete mathematical chemistry had a tremendous growth spurt in the second half of the twentieth century and the same trend is continuing in the twenty first century. This continual growth was fueled primarily by two major factors: 1) Novel applications of discrete mathematical concepts to chemical and biological systems, and 2) Availability of high speed computers and relevant software whereby hypothesis driven as well as discovery oriented research on large data sets could be carried out. This led to the development of not only a plethora of new concepts, but also to various useful applications. This chapter will discuss the major milestones in the development of hierarchical QSARs for the prediction of physical as well as biological properties of various classes of chemicals by the Basak group of researchers using mathematical descriptors and different statistical methods.

Keywords: Property-activity relationship (PAR), graph theory, molecular graphs, weighted pseudograph, graph theoretic matrices, adjacency matrix, distance matrix, topological indices, topostructural indices, topochemical indices, information theoretic indices, connectivity indices, valence connectivity indices, E-state indices, quantum chemical descriptors, hierarchical quantitative structure-activity relationship (HiQSAR), partial least square (PLS), principal components regression (PCR), principal components analysis (PCA), ridge regression (RR), naïve q^2, true q^2, proper cross validation, leave one out (LOO) method, Envelope models, interrelated two-way clustering, linear discriminant analysis, mutagenicity, congenericity principle, diversity begets diversity principle, big data.

[*] **Corresponding author Subhash C. Basak:** International Society of Mathematical Chemistry, 1802 Stanford Avenue, Duluth, MN 55811, USA; Tel: 1-218-727-1335; Fax: 1-218-720-4238; E-mail: sbasak@nrri.umn.edu

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50011-4

1. INTRODUCTION

> *"Knowledge is of no value unless you put it into practice".*
>
> Anton Chekhov

Quantitative structure-activity/property relationships (QSARs/QSPRs) are mathematical models which attempt to predict biomedicinal activity/toxicity/ physicochemical properties of chemicals from their structures or assorted physical properties or substituent constants derived from experimental data as independent variables [1-6]. A contemporary trend in QSAR/QSPR is the use of properties which can be calculated from structure without the input of any other data [2-4]. The major reason behind this is that for the majority of candidate chemicals, both in new drug discovery protocols and hazard assessment of environmental pollutants, experimental properties needed for QSAR formulation are not available [4-7]. The various pathways for the development of structure-activity relationship (SAR) and property-activity relationship (PAR) models either from calculated molecular descriptors or from experimentally derived properties as independent variables may be expressed by Fig. **1** below:

$$\begin{tikzcd}
\mathbf{C} \arrow[rr, "\alpha"] \arrow[dr, "\beta_1"] \arrow[ddr, "\theta_1"'] & & \mathbb{R} \\
 & \mathbf{D} \arrow[ur, "\beta_2"] \arrow[d, "\gamma_1"'] & \\
 & \mathbf{M} \arrow[uur, "\theta_2"'] &
\end{tikzcd}$$

Figure 1: Composition functions for structure-activity relationship (SAR) and property-activity relationship (PAR).

Use of calculated descriptors and experimental properties in PAR/SAR/QSAR may be illuminated through a formal exposition of the structure-property similarity principle—the central paradigm of the field of structure activity relationship [8]. Fig. **1** depicts the determination of an experimental property, *e.g.*, measurement of octanol-water partition coefficient of a chemical, as a function $\alpha: C \rightarrow \mathbb{R}$ which maps the set C of compounds into the real line$\mathbb{R}$. A non-empirical QSAR may be looked upon as a composition of a description function $\beta_1: C \rightarrow D$ mapping each chemical structure of C into a space of non-empirical structural descriptors (D) and a prediction function $\beta_2: D \rightarrow \mathbb{R}$ which maps the descriptors into the real line. One example can be the use of electrotopological state indices for the development of QSARs [9]. When $[\alpha(C) - \beta_2 \circ \beta_1(C)]$ is within the range of experimental errors, we say that we have a good non-empirical QSAR model. On the other hand, the property-activity relationship (PAR) is the composition of $\theta_1: C \rightarrow M$ which maps the set M into the molecular property space M and $\theta_2: M \rightarrow \mathbb{R}$ mapping those molecular properties into the real line$\mathbb{R}$. Property-activity relationship seeks to predict one property (usually a complex property) or bioactivity of a molecule in terms of other (usually simpler or easily determined experimentally) properties. For example, in the estimation of bioconcentraton factor using connectivity indices by Sabljic and Protic [10], it is a theoretically based PAR approach. On the other hand, the structure toxicity relationship (SAR) of narcotic industrial chemicals by Veith *et al.,* [11] in fathead minnow (*Pimephales promelas*) is actually a mixed PAR model because the authors stated that log P (octanol-water) data consisted of a mixture of measured and fragment based calculated values. PAR models using only calculated property (*e.g.* all calculated partition coefficient, log P) are represented in the mapping: $\theta_2 \circ \gamma_1 \circ \beta_1: C \rightarrow \mathbb{R}$, which is a composition of β_1, $\gamma_1: D \rightarrow M$ mapping the descriptor space into the molecular property space (*e.g.* calculation of log P from fragments using the additivity rule), and θ_2, as described in Fig. **1**.

2. THE TORTUOUS HISTORY OF QSAR: FROM 1868 TO THE PRESENT TIME

> *Everything should be made as simple as possible, but not simpler.*
>
> — Albert Einstein

Crum-Brown and Fraser [12] reported their seminal observation in 1968 that the ***structure*** of quaternary compounds was related to their "physiological activity".

In the next phase, the emphasis shifted from structure to physicochemical properties of molecules as the factors underlying their biological activity. About two decades after Crum-Brown and Fraser's initial observation, Richet [13] observed in 1893 that the toxicological activity of different classes of organic compounds was inversely related to their water solubility. Also, in the 1890s and at the turn of the century, Meyer [14] and Overton [15] made the pioneering observation that the biological activity of narcotic chemicals and nonspecific compounds like general anesthetics were related to their partition coefficients between polar and nonpolar solvents. From the thermodynamic point of view Ferguson [16] and Mullins [17] proposed that the thermodynamic activities of various narcotic chemicals in blood are approximately equal when they had equal observable biological effects.

Regarding observed associations between physicochemical properties and biological activities of chemicals Molinengo and Orsetti [18] pointed out: "*These correlations between physical properties of the molecules and their pharmacological activity have been taken as evidence that in certain cases biological activity is unrelated to molecular structure*". They pointed out that the principle of nonspecificity of drug action is often contradicted by various pharmacological data, *e.g.*, oils and other substances with a very high oil-water partition coefficient and a low water solubility, are not hypnotic; the liposoluble camphor is a central nervous system (CNS) stimulant. Such data indicate that the 'nonspecificity' principle is useful only for chemicals having molecular structure capable of eliciting the particular biological effect being studied. For example, barbiturates cause a depression of spinal reflexes. But methylation of both of the nitrogen atoms, which may increase its lipophilicity, transforms these molecules into excitatory drugs [18].

As discussed by Basak [4, 6], in the 1960s, Corwin Hansch's group [19] formulated the linear free energy relationship (LFER) method of QSAR which was a multi-parameter approach involving lipophilicity as well as electronic [20] and steric [21] parameters originating from physical organic chemistry. The linear solvation energy relationship (LSER) approach [22], related to the LFER methodology, is also dependent on experimental data.

As pointed out by Basak [6], the commonality among the approaches from Richet's rule to the LFER as well as LSER techniques is that these are fundamentally property-property relationships (PPRs) where physical and biological properties of molecules are predicted from a set of their measured physicochemical properties. Such PPR or PAR methods worked well in estimating toxicity and biological activity of congeneric sets of chemicals, but are difficult to apply when the data set under investigation is structurally diverse [23].

It may be mentioned here that both for LFER and LSER approaches attempts have been made to calculate the parameters (independent variables of models) from molecular structure alone. Calculation of octanol-water partition coefficient from structure is a well-known case [1]. The use of calculated quantum chemical descriptors instead of those derived from physical organic chemistry is also common [1]. Hickey and Passino-Reader [24] put forward a "rule of thumb" approach for the calculation of LSER descriptors. Famini and Wilson [25] proposed the use of the semiempirical MNDO method of quantum chemistry for the estimation of solvatochromic descriptors.

As described by Basak *et al*., [4, 6, 26, 27], both in new drug discovery and hazard assessment of chemicals we face situations where very few or no experimental properties of chemicals under investigation are known. A reasonable alternative under such circumstances is the use of those descriptors for QSAR formulation which can be calculated directly from chemical structure. With this end in view Basak group of researchers developed the hierarchical QSAR (HiQSAR) approach [6, 27] where topological, geometrical, and quantum chemical descriptors are used for model building in a graduated manner.

3. CALCULATION OF MOLECULAR DESCRIPTORS FOR QSAR

3.1. Definitions of Graph Theoretic Terms and Basic Concepts

> *"The difference between the right word and the almost right word is the difference between lightning and a lightning bug".*

Mark Twain

Mathematicians may flatter themselves that they possess new ideas which mere human language is as yet unable to express.

James C. Maxwell

The field of discrete mathematical chemistry or mathematical chemistry has emerged as an important discipline during past half century or so [4, 6, 27]. A major part of this discipline consists of concepts derived from chemical graph theory and molecular topology where various authors use different terminologies for the same concept. Therefore, it is useful for the reader if the definitions of the terms which will be used in this article are given at the outset.

A graph G is defined as an ordered pair consisting of two sets V and E, $G = [V, E]$, where V represents a finite nonempty set of points, and E is a binary relation, sometime symbolized also by R, defined on the set V. The elements of V are called vertices and the elements of E are called edges. Such an abstract graph is commonly visualized by representing elements of V as points and by connecting each pair (u, v) of elements of V with a line if and only if $(u, v) \in E$. The vertex v and the edge e are incident with each other, as are u and e. Two vertices in G are called adjacent if $(u, v) \in E$. A walk of a graph is a sequence beginning and ending with vertices in which vertices and edges alternate and each edge is incident with vertices immediately preceding and following it. A walk of the form v_0, e_1, v_1, e_2, …, v_n joins vertices v_0 and v_n. The length of a walk is the number of edges in the walk. A walk is closed if $v_0 = v_n$; otherwise it is open. A closed walk with n points is a cycle if all its points are distinct and $n \geq 3$. A path is an open walk in which all vertices are distinct. A graph G is connected if every pair of its vertices is connected by a path. A graph G is a multigraph if it contains more than one edge between at least one pair of adjacent vertices, otherwise G is a simple graph. The distance d (u, v) between vertices u and v in G is the length of the shortest path connecting u and v. The degree of vertex v, denoted by δ^v, is equal to the number of edges incident with v. The eccentricity e (u) of a vertex u in G is defined as $e\ (u) = \max d\ (u, v)$, where $u, v \in V$. The radius, ρ, of a graph is given by $\rho = \min e\ (u)$, for all $u \in V$. For a vertex $v \in V$, the first order neighborhood, $r^1(v)$, is a subset of V such that $r^1(v) = \{ u \in V| \ d(u, v) =1)\}$. The first-order closed neighborhood of

v, $N^1(v)$, is defined as $N^1(v)$ = (v) U r^1 (v) = $r^0(v)$ U r^1 (v), where {v} is the one point set consisting of v only and may be taken as $r^0(v)$. If ρ is the radius of a graph, one can construct $N^i(u)$, i = *1, 2, ..., ρ*, for each vertex $u \in V$. Two graphs, G_1 and G_2, are said to be isomorphic if there exists a one-to-one mapping of the vertex set of G_1 onto that of G_2 such that adjacency is preserved. Automorphism is the isomorphism of a graph G with itself.

In a molecular graph, V represents the set of atoms and E or R represents the set of bonds present in the molecule. It should be noted, however, that the set E is not limited to covalent bonds only. In fact, elements of E may symbolize any type of bond, *viz*., covalent, ionic, or hydrogen bond, *etc*. It was pointed out by Basak *et al.,* [28] that weighted pseudographs, which contain both self-loops (an edge by which a vertex is connected to itself) and multiple bonds between at least one pair of vertices, constitute a versatile model for the representation of a wide range of chemical species. In depicting a molecule by a connected graph $G = [V, E]$, V may contain either all atoms present in the empirical formula or only non-hydrogen atoms. Hydrogen-filled graphs are preferable to hydrogen-depleted graphs when hydrogen atoms are involved in critical steric or electronic interactions or when hydrogen atoms have different physicochemical properties due to differences in their bonding neighborhoods.

3.2. Methods for the Calculation of Topological Indices

In many cases, the following is the sequence of steps in the representation and characterization of molecules using molecular topological methods:

a) Representation of the molecule by a chemical graph of choice

b) Development of various matrices, *viz*., adjacency matrix, distance matrix, *etc*. from the molecular graph

c) Extraction of invariants from the matrices as numerical molecular descriptors

For details of the above three steps see [3, 4, 6, 27-29]. The first chapter by Basak [6] in this volume of the eBook has discussed these steps in some details. So, these are not repeated here for brevity.

In the formulation of information theoretic topological indices, an appropriate set of elements are extracted from the molecular graph model and the set is then partitioned using a properly defined equivalence relation. An equivalence relation is reflexive, symmetric, and transitive relation defined on a set and partitions the set into mutually disjoint subsets [30]. Subsequently, Shannon's [31] relation is used for the calculation of information theoretic indices [6, 32, 33]. During the past four decades, Basak *et al.,* [34-36] used different types of equivalence relations in the calculation of various classes of information theoretic indices

3.3. Software for Calculation of Topological Indices, Atom Pairs and Quantum Chemical Descriptors

In their research, Basak *et al.,* have been using molecular descriptors calculated by MolconnZ [37], POLLY [38], an in-house software [39] developed for the calculation of Triplet indices [40] and APProbe [41], the last one being capable of calculating atom pairs (APs) [42] from molecular graphs. Quantum chemical descriptors were calculated using Sybyl v. 6.2 [43], MOPAC v 6.00 [44] and Gaussian [45].

Basak *et al.,* [6] divided the topological indices (TIs) into two major groups: Topostructural (TS) indices and topochemical (TC) indices. TS descriptors are topological indices which are calculated from skeletal graph models of molecules which do not distinguish among different types of atoms in a molecule or the various types of chemical bonds, *e.g.*; single bond, double bond, triplet bond, *etc*. Thus, TS descriptors quantify information regarding the connectivity, adjacency, and distances between vertices of molecular graphs, ignoring their distinct chemical nature. TC indices, on the other hand, are sensitive to both the pattern of connectedness of the vertices (atoms), as well as their chemical/bonding characteristics. Therefore, the TC indices are more complex than the TS descriptors.

Over the years Basak and coworkers have used different combinations of TS, TC, 3-D, and quantum chemical indices. Table **1** below gives a typical list of molecular descriptors used by Basak group of researchers in the formulation of QSARs. The set of bonding connectivity indices, *e.g.*, ${}^{h}\chi^{b}$ (bonding path connectivity index of order h = 0-6 and other indices of this bonding class) were defined for the first time by Basak *et al.,* [28].

Table 1: Symbols, definitions and classification of topological indices

Topostructural (TS)	
I_D^W	Information index for the magnitudes of distances between all possible pairs of vertices of a graph
$\overline{I}_D^W$	Mean information index for the magnitude of distance
W	Wiener index = half-sum of the off-diagonal elements of the distance matrix of a graph
I^D	Degree complexity
H^V	Graph vertex complexity
H^D	Graph distance complexity
$\overline{IC}$	Information content of the distance matrix partitioned by frequency of occurrences of distance *h*
M_1	A Zagreb group parameter = sum of square of degree over all vertices
M_2	A Zagreb group parameter = sum of cross-product of degrees over all neighboring (connected) vertices
${}^h\chi$	Path connectivity index of order *h* = 0-10
${}^h\chi_C$	Cluster connectivity index of order *h* = 3-6
${}^h\chi_{PC}$	Path-cluster connectivity index of order *h* = 4-6
${}^h\chi_{Ch}$	Chain connectivity index of order *h* = 3-10
P_h	Number of paths of length *h* = 0-10
J	Balaban's *J* index based on topological distance
nrings	Number of rings in a graph
ncirc	Number of circuits in a graph
DN^2S_y	Triplet index from distance matrix, square of graph order, and distance sum; operation *y* = 1-5
DN^21_y	Triplet index from distance matrix, square of graph order, and number 1; operation *y* = 1-5
$AS1_y$	Triplet index from adjacency matrix, distance sum, and number 1; operation *y* = 1-5
$DS1_y$	Triplet index from distance matrix, distance sum, and number 1; operation *y* = 1-5
ASN_y	Triplet index from adjacency matrix, distance sum, and graph order; operation *y* = 1-5
DSN_y	Triplet index from distance matrix, distance sum, and graph order; operation *y* = 1-5

DN^2N_y	Triplet index from distance matrix, square of graph order, and graph order; operation y = 1-5
ANS_y	Triplet index from adjacency matrix, graph order, and distance sum; operation y = 1-5
$AN1_y$	Triplet index from adjacency matrix, graph order, and number 1; operation y = 1-5
ANN_y	Triplet index from adjacency matrix, graph order, and graph order again; operation y = 1-5
ASV_y	Triplet index from adjacency matrix, distance sum, and vertex degree; operation y = 1-5
DSV_y	Triplet index from distance matrix, distance sum, and vertex degree; operation y = 1-5
ANV_y	Triplet index from adjacency matrix, graph order, and vertex degree; operation y = 1-5
kp_0	Kappa zero
kp_1-kp_3	Kappa simple indices
Topochemical (TC)	
O	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-filled graph
O_{orb}	Order of neighborhood when IC_r reaches its maximum value for the hydrogen-suppressed graph
I_{ORB}	Information content or complexity of the hydrogen-suppressed graph at its maximum neighborhood of vertices
IC_r	Mean information content or complexity of a graph based on the r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
SIC_r	Structural information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
CIC_r	Complementary information content for r^{th} (r = 0-6) order neighborhood of vertices in a hydrogen-filled graph
${}^{h}\chi^{b}$	Bond path connectivity index of order h = 0-6
${}^{h}\chi^{b}_{C}$	Bond cluster connectivity index of order h = 3-6
${}^{h}\chi^{b}_{Ch}$	Bond chain connectivity index of order h = 3- 6
${}^{h}\chi^{b}_{PC}$	Bond path-cluster connectivity index of order h = 4-6
${}^{h}\chi^{v}$	Valence path connectivity index of order h = 0-10
${}^{h}\chi^{v}_{C}$	Valence cluster connectivity index of order h = 3-6
${}^{h}\chi^{v}_{Ch}$	Valence chain connectivity index of order h = 3-10
${}^{h}\chi^{v}_{PC}$	Valence path-cluster connectivity index of order h = 4-6

J^B	Balaban’sJ index based on bond types
J^X	Balaban’sJ index based on relative electronegativities
J^Y	Balaban’sJ index based on relative covalent radii
AZV$_y$	Triplet index from adjacency matrix, atomic number, and vertex degree; operation y = 1-5
AZS$_y$	Triplet index from adjacency matrix, atomic number, and distance sum; operation y = 1-5
ASZ$_y$	Triplet index from adjacency matrix, distance sum, and atomic number; operation y = 1-5
AZN$_y$	Triplet index from adjacency matrix, atomic number, and graph order; operation y = 1-5
ANZ$_y$	Triplet index from adjacency matrix, graph order, and atomic number; operation y = 1-5
DSZ$_y$	Triplet index from distance matrix, distance sum, and atomic number; operation y = 1-5
DN^{2}Z$_y$	Triplet index from distance matrix, square of graph order, and atomic number; operation y = 1-5
nvx	Number of non-hydrogen atoms in a molecule
nelem	Number of elements in a molecule
fw	Molecular weight
si	Shannon information index
totop	Total Topological Index *t*
sumI	Sum of the intrinsic state values *I*
sumdelI	Sum of delta-*I* values
tets2	Total topological state index based on electrotopological state indices
phia	Flexibility index (kp_1* kp_2/nvx)
Idcbar	Bonchev-Trinajstić information index
IdC	Bonchev-Trinajstić information index
Wp	Wiener *p*
Pf	Platt *f*
Wt	Total Wiener number
knotp	Difference of chi-cluster-3 and path/cluster-4
knotpv	Valence difference of chi-cluster-3 and path/cluster-4
nclass	Number of classes of topologically (symmetry) equivalent graph vertices
NumHBd	Number of hydrogen bond donors
NumHBa	Number of hydrogen bond acceptors

SHCsats	E-State of C sp^3 bonded to other saturated C atoms
SHCsatu	E-State of C sp^3 bonded to unsaturated C atoms
SHvin	E-State of C atoms in the vinyl group, *=CH-*
SHtvin	E-State of C atoms in the terminal vinyl group, $=CH_2$
SHavin	E-State of C atoms in the vinyl group, *=CH-*, bonded to an aromatic C
SHarom	E-State of C sp^2 which are part of an aromatic system
SHHBd	Hydrogen bond donor index, sum of Hydrogen E-State values for *-OH*, *=NH*, $-NH_2$, *-NH-,-SH*, and *#CH*
SHwHBd	Weak hydrogen bond donor index, sum of *C-H* Hydrogen E-State values for hydrogen atoms on a C to which a F and/or Cl are also bonded
SHHBa	Hydrogen bond acceptor index, sum of the *E*-State values for *-OH*, *=NH*, $-NH_2$, -*NH-*, *>N*, *-O-*, *-S-*, along with -F and -Cl
Qv	General Polarity descriptor
$NHBint_y$	Count of potential internal hydrogen bonders (*y* = 2-10)
SHBinty	E-State descriptors of potential internal hydrogen bond strength (*y =2-10*)
ka_1-ka_3	Kappa alpha indices
	Electrotopological State index values for atom types: *SHsOH, SHdNH, SHsSH, SHsNH2, SHssNH, SHtCH, SHother, SHCHnX, HmaxGmax, Hmin, Gmin, Hmaxpos, Hminneg, SsLi, SssBe, Sssss, Bem, SssBH,SsssB, SssssBm, SsCH3, SdCH2, SssCH2, StCH, SdsCH, SaaCH, SsssCH, SddC, StsC, SdssC, SaasC, SaaaC, SssssC, SsNH3p, SsNH2, SssNH2p, SdNH, SssNH, SaaNH, StN, SsssNHp, SdsN, SaaN, SsssN, SddsN, SaasN, SssssNp, SsOH, SdO, SssO, SaaO, SsF, SsSiH3, SssSiH2, SsssSiH, SssssSi, SsPH2, SssPH, SsssP, SdsssP, SsssssP, SsSH, SdS, SssS, SaaS, SdssS, SddssS, SssssssS, SsCl, SsGeH3, SssGeH2, SsssGeH, SssssGe, SsAsH2, SssAsH, SsssAs, SdsssAs, SsssssAs, SsSeH, SdSe, SssSe, SaaSe, SdssSe, SddssSe, SsBr, SsSnH3, SssSnH2, SsssSnH, SssssSn, SsI, SsPbH3, SssPbH2, SsssPbH, SssssPb*
	Geometrical (3-D)
^{3D}W	3D Wiener number based on the hydrogen-suppressed geometric distance matrix
$^{3D}W_H$	3D Wiener number based on the hydrogen-filled geometric distance matrix
V_W	Van der Waal's volume
	Quantum Chemical (QC)
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{HOMO-1}	Energy of the second highest occupied molecular
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{LUMO+1}	Energy of the second lowest unoccupied molecular orbital
ΔHf	Heat of formation
μ	Dipole moment

4. HIERARCHICAL QSAR DEVELOPMENT AND VALIDATION

4.1. The Major Pillars of QSAR: 3Ds- Data Quality, Descriptor Relevance, and Data Fitting

“In God we trust; all others bring data”.

W. Edwards Deming

"For a successful technology, reality must take precedence over public relations, for nature cannot be fooled”.

Richard P. Feynman

The following are the prerequisites for the development of QSAR models:

a) Reasonably large and good quality bioassay or physicochemical property data (dependent variable) for a set of chemicals,

b) For the same set of chemicals, a collection of experimental data or relevant molecular descriptors (independent variables) which can adequately quantify aspects of molecular structure related to the physical property/biological activity, and

c) Proper methods of data fitting to models and their validation

A survey of modern QSAR literature would show that both experimentally determined physical properties [1] and substituent constants derived from test data as well as calculated molecular descriptors [2-6, 10-11, 22, 23, 25-27, 46-59] have been used for the formulation of QSARs.

4.2. Statistical Methods

It is not enough to do your best; you must know what to do, and then do your best.

W. Edwards Deming

शैले शैले न माणिक्यं मौक्तिकं न गजे गजे ।

साधवो नहि सर्वत्र चन्दनं न वने वने ॥

shaile shaile na maanikyam mauktikam na gaje gaje

Saadhavo naahi sarvatra chandanam na vane vane

(In Sanskrit)

Not all mountains contain gems in them, nor does every elephant has
pearl in it, noble people are not found everywhere, nor is sandalwood
found in every forest.

Chanakya

While building a scientifically interpretable and technically sound QSAR model, the researcher needs to keep in mind some specific issues. First and foremost of them is checking whether a specific method is applicable, or ideally, determining the best method to model a specific QSAR scenario. For example, in a regression setup where the number of descriptors (p) is much larger than number of samples (n) *i.e.* $p >> n$, the estimate of the coefficient vector is not unique. This is also the case when predictors in the study are heavily correlated with one another to the extent that the ‘design matrix’ becomes rank-deficient. Both of these situations are highly relevant to the QSAR paradigm. In many contemporary QSAR studies, the number of initial predictors typically is in hundreds or thousands, while more often than not, mostly to mitigate experimental cost, the experimenter can collect only tens or hundreds of samples. This effectively makes the problem high-dimensional ($p >> n$) in nature. Also, when a large number of descriptors on a set of chemicals are used to model their activity, it is only natural that some predictors within a single class or predictors in different classes are highly correlated to one another. Such situations can either be tackled by attempting to pick important variables through model selection or ‘sparsity’-type approaches (*e.g.* forward selection, LASSO [60], adaptive LASSO [61]), or finding a lower-dimensional transformation that preserves most of the descriptor information, *e.g.* Principal Component Analysis (PCA), envelope methods [62]. A third option here would be using machine learning methods, or even combining several models to improve predictions [63].

One can check the generalizability of a model, *i.e.* its ability to give competent predictions on 'similar' datasets (we shall discuss how this similarity is defined through applicability domains in a while) through validation on out-of-sample test datasets. For a small set of compounds, this is obtained by doing leave-one-out cross-validation, while for datasets with a larger number of compounds, a more computationally economical way is doing *k*-fold cross-validation: split the dataset randomly into *k* (previously fixed) equal subsets, take each subset in turn as test set and other compounds as training sets and obtain predictions. Comparing cross-validation with the somewhat prevalent approach of external validation, *i.e.* choosing a single train-test split of compounds, we observe that in external validation the splits are chosen with the help of the experimenters' knowledge or some ad-hoc criterion, while in cross-validation the splits are chosen randomly, thus intuitively providing a more unbiased estimate of the generalizability of the QSAR model. Furthermore, Hawkins *et al.,* [64] proved theoretically that compared to external validation, cross-validation is a better estimator of the actual predictive ability of a statistical model for small datasets, while for large sample size both perform similarly. Quoting the authors, "*The bottom line is that in the typical QSAR setting where available sample sizes are modest, holding back compounds for model testing is ill-advised. This fragmentation of the sample harms the calibration and does not give a trustworthy assessment of fit anyway. It is better to use all data for the calibration step and check the fit by cross-validation, making sure that the cross-validation is carried out correctly*".

Special care should be taken when combining conventional modelling with the additional step of variable selection dimension reduction. An *intuitive, but wrong, procedure* in this scenario would be to perform the first stage of pre-processing first, selecting important variables or determining the optimal transformation, and then use the transformed data/selected variables to build the predictive models and obtain predictions for each train-test split. The reason this is not appropriate is that the data is split only after the variable selection/dimension reduction step, thus essentially this method ends up using information from the holdout compound/split to predict activity of these very samples. This *Naïve cross-validation procedure* causes synthetic inflation of the cross-validated q^2, hence the predictive ability of the model [65, 66] (See Fig. **2**). A two-step procedure (referred in Fig. **2** as 'Two-deep CV') helps navigate this situation. Instead of doing the pre-model building step first and then taking multiple splits for out-of-sample prediction, for each split of the data the initial steps are performed only using the training set of compounds each time. Since calculations on two different splits are not dependent on each other, the increased computation load due to repeated variable selection can be tackled using parallel processing.

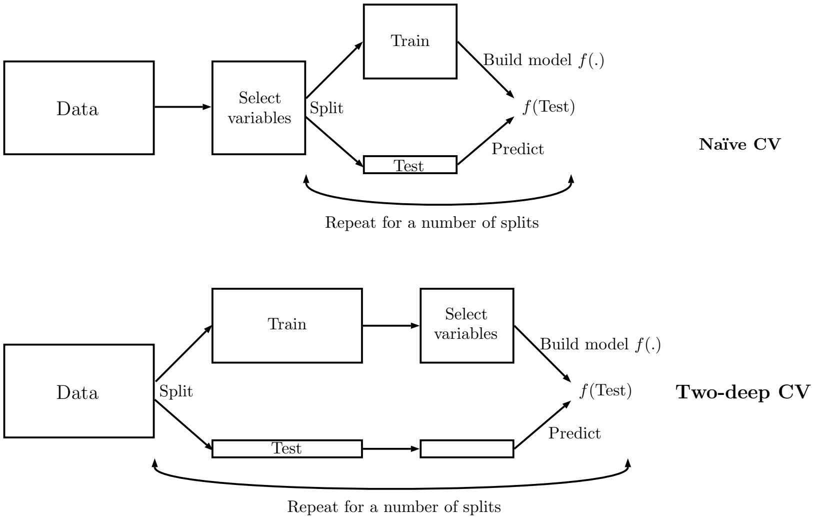


Figure 2: Difference between naïve and two-deep cross validation (C V).

4.3. Applicability Domain of QSAR Models

The final important issue one needs to handle while developing a QSAR model is that of defining applicability domain (AD) of the model. This is a required criterion of any valid implementable QSAR model according to OECD principles [67]. There are several methods of defining the AD of any statistical model, and these can be roughly put into two categories: explicitly attempting to define the active predictor space through some method like bounding box, PCA or convex hulls; and distance-based methods that calculate the similarity of a new compound to the set of compounds which have been used to build the training model. To obtain predictions for any incoming test sample using the model developed, the first set of methods are used to ensure that the compound belongs to the so-called ‘active subspace’: which essentially means we are doing interpolation, not extrapolation [68, 69]. For the distance-based approach, a pre-defined statistic is calculated to quantify the proximity of the new compound to the training set, and based on whether that statistic is above or below a certain cutoff, predictions for that compound are obtained [68, 70].

4.4. Hierarchical QSAR of Congeneric and Diverse Sets: An Example with Chemical Mutagens

> *"Computers are incredibly fast, accurate, and stupid. Human beings are incredibly slow, inaccurate, and brilliant. Together they are powerful beyond imagination".*
>
> Albert Einstein

In the 1990s, Basak group formulated the principle of hierarchical HiQSAR approach [46-53] and applied it in the prediction of physical property as well as bioactivity/toxicity of chemicals at the levels of enzymes, receptors, cells, and whole organisms. In this approach, one uses more complex and resource intensive descriptors only if they result in significant improvement in the quality of the predictive model as compared to the simpler indices. We begin by building QSAR models using only the TS descriptors, followed by the creation of additional models based on the successive inclusion of the hierarchically ranked descriptor classes (Fig. **3**). By comparing the resulting models, the contribution of each descriptor class is elucidated. In addition, the hierarchical approach enables us to determine whether or not the higher level descriptors are necessary in predicting the property or activity under consideration. In situations where the complex descriptors are not useful, we can avoid spending the resource required for their calculation. The full hierarchical QSAR scheme involving TS, TC, 3-D, and the different levels of quantum chemical indices as well as biodescriptors [54-59] derived from proteomics patterns are shown in Fig. **3** below. In this chapter, however, our discussion will be restricted to the chemodescriptors only.

The work by Majumdar *et al.,* [71] implemented some of the important points described above, *viz*., the importance of feature selection, relevance of integrating *recent methodological research taken from the omics area into the QSAR paradigm*, and two-fold cross-validation. In this chapter, we discuss the QSAR models for predicting mutagenicity of the homogeneous set of 95 aromatic and heteroaromatic amines using a combination of TS+TC+3-D + QC descriptors and predictive models of a diverse set of 508 chemical compounds built hierarchically using 5 types of descriptors: TS, TC, 3D, QC, and atom pairs: the total number of

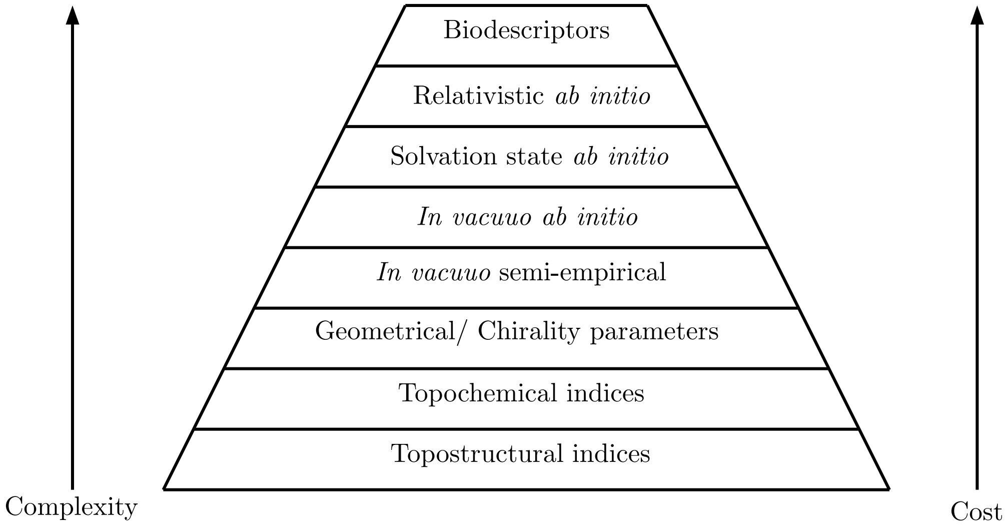


Figure 3: Hierarchical use of chemodescriptors and biodescriptors in QSAR.

predictors for the 508 set being 2,525. A machine learning method called Interrelated Two-way clustering (ITC), originally developed for application in gene microarray data [72], is used for variable selection, and resulting predictors are fed into a ridge regression model to get final predictions. The ITC algorithm involves the following steps:

i. Predictors are clustered into separate groups, say $G_1, G_2, \ldots, G_k$, which are substituted by several types of descriptors in QSAR;

ii. After that samples are clustered into two classes using each group, Say $S_{i,a}$ and $S_{i,b}$; $i = 1,2, \ldots, n$;

iii. All possible intersections of the 2^k clusters are taken. For example, for $k = 2$ the intersections are:

$$C_1 = S_{1,a} \cap S_{2,a};\ C_2 = S_{1,b} \cap S_{2,a};\ C_3 = S_{1,a} \cap S_{2,b};\ C_4 = S_{1,b} \cap S_{2,b}$$

iv. These are divided into heterogeneous groups: pairs of intersections with no common elements, *e.g.* $H_{14} = (C_1, C_4)$ and $H_{23} = (C_2, C_3)$ above;

v. For each $H_{st} = (C_s, C_t)$, cosine distances of subvectors with predictors from this heterogeneous group are calculated with the two model vectors: one with C_s zeros and C_t ones, and another with C_s ones and C_t zeros. Each distance vector is sorted in decreasing order, top one-third of predictors are taken from each of these vectors and are merged.

The algorithm is then repeated with selected predictors, and terminated when 90% of total number of samples is covered by the largest heterogeneous group, or maximum number of iterations reached. This is done because through the algorithm the groups become more and more similar, so sample classifications using them become more and more similar, thus heterogeneous groups cover an increasing proportion of total number of samples.

Two-deep cross-validation is used to obtain misclassification percentages. The findings obtained can be summarized into two points:

1. The prediction performances were almost same as those obtained in a previous study [49] that used the full set of predictors, demonstrating the utility of variable selection;

2. There is a significant improvement in model performance when TC predictors are added to a model built on only TS predictors, but beyond that, inclusion of 3D and QC predictors do not improve the model quality, although these are more computationally intensive to calculate.

The above findings were reinforced by results in a subsequent paper [73]. Apart from the ITC + ridge regression framework used in the previous paper [71], this study also introduces a new method of dimension reduction called envelope method into QSAR modelling, and combines it with linear discriminant analysis in a binary classification scenario. The methods are used on two datasets: the first one being the structurally diverse 508 mutagen data and the second one is a homogeneous dataset comprised of 95 aromatic and heteroaromatic amines mutagens [74]. Table **2** and **3** summarize the findings obtained from this analysis. The benefits of adding 3D and QC descriptors are visible only for envelope + LDA analysis on the 508 compound

Table 2: Comparison of performances of model 1 (RR and ITC+RR) for diverse and congeneric datasets

Dataset used	**Predictive model**	**Type of predictor used**	**No. of predictors**	**Correct classification %**	**Sensitivity**	**Specificity**
508 compound diverse dataset	Ridge regression [49]	TS	103	53.14	52.34	53.97
		TS+TC	298	76.97	83.98	69.84
		TS+TC+3D+QC	307	77.17	84.38	69.84
	ITC+ RR	TS	103	66.34	73.83	58.73
		TS+TC	298	73.23	77.34	69.05
		TS+TC+3D	301	74.80	77.34	72.22
		TS+TC+3D+QC	307	72.05	76.17	67.86
		TS+TC+AP [71]	2620	78.35	84.38	72.22
95 amines congeneric dataset	Ridge regression	TS	108	83.16	75.47	88.42
		TS+TC	266	84.21	77.36	92.86
		TS+TC+3D	269	84.21	77.36	92.86
		TS+TC+3D+QC	275	84.21	77.36	92.86
	ITC + RR	TS	108	88.42	92.45	83.33
		TS+TC	266	89.47	92.45	85.71
		TS+TC+3D	269	88.42	92.45	83.33
		TS+TC+3D+QC	275	85.26	88.68	80.95

Table 3: Comparison of performances of model 2 (Envelope LDA) diverse and congeneric datasets

Dataset used	**Type of predictor used**	**No. predictors**	**Correct classification %**	**Sensitivity**	**Specificity**
508 compound diverse dataset	TS	103	57.09	65.63	48.41
	TS+TC	298	60.24	69.92	46.43
	TS+TC+3D	301	61.02	71.09	50.79
	TS+TC+3D+QC	307	64.37	69.14	59.52
95 amines congeneric dataset	TS	108	81.05	92.86	71.70
	TS+TC	266	80.00	83.33	77.36
	TS+TC+3D	269	80.00	83.33	77.36
	TS+TC+3D+QC	275	71.58	78.57	66.04

dataset. We also observe that for the smaller, homogeneous dataset, adding other classes of descriptors on top of a model built using only TS descriptors does not

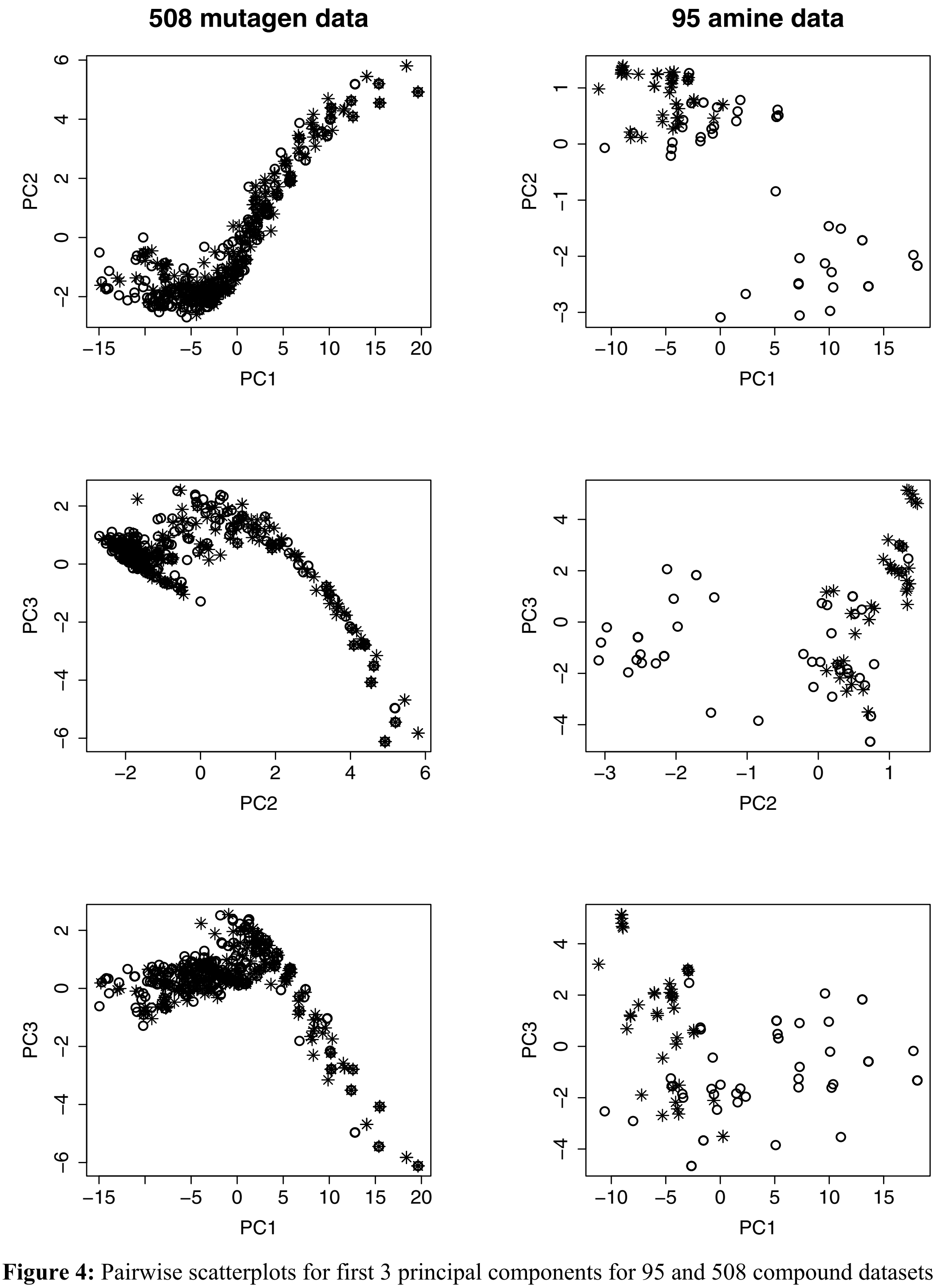


Figure 4: Pairwise scatterplots for first 3 principal components for 95 and 508 compound datasets.

improve out-of-sample predictions, while a considerable jump in correct classification percentage, sensitivity and specificity is observed in all the cases TC descriptors are added in a TS-only model on the 508 compound dataset. This lends support to the '*Diversity begets diversity principle*' [75], according to which there is a possible tradeoff between model complexity and composition of the dataset at hand: it is enough to use one single type of predictors to model 'well-behaved', homogeneous set of compounds, while a more diverse set of compounds requires a diverse collections of predictors to be used in modelling for the purpose of obtaining plausible prediction performance. Distinction between the two datasets is visible in pairwise plots of first 3 principal components in Fig. **4**. The distinction of two classes is quite clear for the 95 compound data (right column of Fig. **4**), but not for the diverse 508 compound data (left column).

DISCUSSION AND CONCLUSION

> *He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may cast.*
>
> Leonardo da Vinci

The objectives of this paper were three-fold: a) Investigate the use of HiQSAR approach developed by Basak *et al.,* [46-59] in the development of models for the prediction of mutagenicity of chemicals from their calculated descriptors, b) Study the relative niches of the congenericity *versus* diversity begets diversity principles [73], and c) Test the ability of different robust statistical methods in model development in rank deficient scenarios.

For, the congeneric 95 aromatic and heteroaromatic mutagens the addition of TC, 3-D, and quantum chemical indices after the use of TS descriptors did very little improvement in model quality. This is in line with our previous studies using HiQSAR approach [46-53] for various physicochemical, biomedical, and toxicological properties. The contrast in the results using TS and TC descriptors is also revealing. Whereas for the congeneric set of amines TS descriptors alone gave good quality models (Tables **2** and **3**), TC descriptors were very helpful

(Table **2**) in augmenting model quality for the 508 set. As discussed earlier, Basak *et al.,* [6, 27] coined the term "topostructural" for those topological indices which were defined on simple skeletal graphs that encode information regarding the size, shape, complexity, branching, *etc*. of molecular graphs ignoring, at the same time, chemically important features of vertices like atom types, bonding pattern, electron distribution. That such indices explain most of the variance in mutagenicity of the amines indicates that the variance in the dependent variable is determined by the general features of molecular constitution. On the other hand, for the 508 diverse set of mutagens (Table **2**) the TC indices make a significant improvement in model quality indicating an important role of molecular electronic character over and above the general structural features of the chemicals under investigation. Further investigation with other sets of congeneric and structurally diverse sets of chemicals are needed to understand the relative utility of the congenericity principle *versus* the diversity begets diversity principle in the general scheme of QSAR development.

Since the 1980s Basak and coworkers have been active in the exploration of various statistical methods in harnessing the collective power of calculated molecular descriptors, easily calculated topological indices in particular, in the selection of analogs and formulation of QSARs [76-81]. When regressions using individual descriptors failed for diverse sets of chemicals, Basak *et al.,* [28] began using robust statistical methods like PCA to extract useful information from the diverse collection of indices. Sometime methods like VARCLUS procedure of SAS [82, 83] was used to extract useful information. Subsequently, as more and more descriptors and software for their calculation became available and the situation became rank deficient, we started using appropriate methods [84] like principal components regression (PCR), partial least square (PLS), and ridge regression (RR) in QSAR model building. More recently, we tried interrelated two-way clustering or ITC [72] and in this paper used the envelope method [62] in QSAR formulation.

From the first formulation of graph theoretic indices or topological indices in 1947 by Wiener [85], many topological indices have been devised by mathematical chemists [3, 9, 27, 29, 32-36, 76, 86-88]. They have worked well in their specialized, local domains for which they were devised. But when they are compared with one another using a diverse data set and the PCA methodology [28], many of the indices, intuitively asserted to be mutually different, were loaded to the same PC. Here we give the results from our PCA study on the on the 90 indices calculated by POLLY [38] for the diverse set of 3,692 chemicals [28]: a) PC_1 was highly correlated ($0.96 > r > 0.69$) with the size and shape of the molecular graph; b) Higher order information theoretic indices (IC, CIC, and SIC), quantifying molecular complexity, were highly correlated with PC_2 with average correlation $r = 0.8$; c) PC_3 was correlated highly with cluster ($0.55 < r < 0.69$) and path/cluster ($0.27 < r < 0.59$) connectivity indices; because these indices have traditionally been associated with branching in a molecular graph, this PC was interpreted as reflecting molecular branching; d) PC_4 was clearly correlated with cyclic terms of the molecular connectivity indices. As more and more molecular descriptors are available, we need to do such studies with augmented sets of indices to find out which subsets of them are useful for QSAR and other purposes.

The era of “big data” has arrived [89]. The different aspects chemistry, computer aided drug design, and predictive toxicology/environmental ecotoxicology will be guided by big data and real time analytical/predictive tools associated with them [90, 91]. While discussion is going on about the 4Vs-- volume, velocity, variety and veracity-- of big data [92], proper data analytical tools are needed for the recognition of the interesting and latent relationships among the huge amount of data that is being generated in the virtual screening arena [93]. We hope some of the statistical methods that Basak and coworkers have been exploring since the 1980s will help in shedding some light in this area. The fifth V in the data area, *value or price of computation*, has been proposed by Basak *et al.,* [89]. The HiQSAR studies of Basak and coworkers are intimately connected with this V #5. In a HiQSAR study on a set of halocarbons, a group of chemicals important for both organic synthesis and environmental toxicology, Basak *et al.,* [94] used TS, TC, 3-D and different quantum chemical indices, *viz*., semiempirical AM_1, and *ab initio* STO-3G, 6-31G(d), 6-311G, 6-311G(d), and cc-pVTZ level descriptors. While a combination of TS and TC indices gave a reasonable model ($R^2_{cv} = 0.81$; s. e = 0.57), the addition of AM_1, and ab initio STO-3G, 6-31G(d), 6-311G, and 6-311G(d) indices did not make any improvement in model quality. But the addition of cc-pVTZ indices improved the correlation ($R^2_{cv} = 0.92$; SE = 0.38). Under such circumstances one will have to use their judgement whether such expensive calculations are called for or one will be satisfied with the TS and TC type indices which can be calculated fast and work well in many cases. Basak *et al.,* [95, 96] found that topological indices are capable of developing good quality QSARs for physicochemical, biomedical, and toxicological properties of both congeneric and diverse sets of chemicals. Some studies with anticancer 2-phenylindoles [97], active against breast cancer cells, and boron-containing dipeptide proteasome inhibitors [98] indicate that TS and TC based QSARs compare well with those developed using comparative molecular field analysis (CoMFA) methods.

In this age of high-performance computing, the landscape of predictive analytics is undergoing rapid changes and developments. To satisfy all OECD requirements of a valid QSAR model, namely a definite endpoint, a clear algorithm, specified applicability domain, measures of model performance and interpretation [66], we need to leverage these new methods as well as adapt to their theoretical framework. While statistical models often provide clear interpretability of predictor effects owing to their well-defined mathematical structures, and can be used to model complex scenarios like temporal dependency, off-the-shelf machine learning tools tend to perform better when prediction is the main goal. Finally, opening up to the applications of latest methods being developed and adhering to strict data- analytic procedures has the potential of developing more interdisciplinary collaborations, a wider audience, and, most importantly, a better understanding of the underlying scientific processes.

ACKNOWLEDGEMENTS

Subhash C. Basak deeply appreciates the help that he received from the co-editor of this eBook, Guillermo Restrepo. Basak would also like to thank Gregory Grunwald for sustained collaboration and technical support for almost three decades of Basak’s research at the University of Minnesota Duluth. Research of SM is supported by the United States National Science Foundation grant IIS-1029711.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

REFERENCES

[1] Hansch, C.; Leo, A. *Exploring QSARs: Fundamentals and Applications in Chemistry and Biology*, American Chemical Society: Washington, DC, **1995**, pp. 557.

[2] Kier, L.B.; Hall, L.H. *Molecular Structure Description: The Electrotopological State*, Academic Press: San Diego, CA, **1999**, pp. 245.

[3] Devillers, J.; Balaban, A.T., Eds. *Topological Indices and Related Descriptors in QSAR and QSPR*, Gordon and Breach: Amsterdam, **1999**, pp. 811.

[4] Basak, S. C. Role of mathematical chemodescriptors and proteomics-based biodescriptors in drug discovery, *Drug Develop. Res.*, **2010**, *72*, 1-9.

[5] Hawkins, D. M.; Basak, S. C.; Kraker, J. J.; Geiss, K. T.; Witzmann, F. A., Combining chemodescriptors and biodescriptors in quantitative structure-activity relationship modeling, *J. Chem. Inf. Model.*, **2006**, *46*, 9-16.

[6] Basak, S. C. Mathematical Structural Descriptors of Molecules and Biomolecules: Background and Applications, In: *Advances in Mathematical Chemistry and Applications*, vol. 1, pp. 3-23; Basak, S. C., Restrepo, G. and Villaveces, J. L., Eds.; Bentham eBooks, Bentham Science Publishers, **2015**.

[7] 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.*, **1990**, *87*, 183-197.

[8] Johnson; M, Basak, S. C.; Maggiora, G. A characterization of molecular similarity methods for property prediction. *Math. Comput. Model.* **1988**, *11*, 630-634.

[9] Kier, L.B.; Hall, L.H. *Molecular Structure Description: The Electrotopological State*, Academic Press: San Diego, CA, **1999**, pp. 245.

[10] Sabljic, A.; Protic, M. Molecular connectivity: a novel method for prediction of bioconcentration factor of hazardous chemicals. *Chemico-Biological Interactions*, **1982**, *42*, 301-310.

[11] Veith; G. D., Call, D. J.; Brooke, L. T. Structure-toxicity relationships for the fathead minnow, Pimephales promelas: Narcotic industrial chemicals. *Can. J. Fish. Aquat. Sci.*, **1983**, *40*, 743-748.

[12] Crum-Brown A.; Fraser, T. R. On the connection between chemical constitution and physiological action. Part1. On the physiological action of the ammonium bases, derived from Strychia, Brucia, Thebaia, Codeia, Morphia and Nicotia. *Trans. Roy. Soc. Edinb.*, **1868**, *25*, 151-203.

[13] Richet M, C. Note sur le rapport entre la toxicite´ et les proprie´ te´ s physiques des corps. *CR Soc. Biol. (Paris)*, **1893**, *45*, 775-776.

[14] Meyer, H. Zur theorie der alkoholnarkose (I): Welche eigenschaft der an¨asthetica bedingt ihre narkotische wirkung. *Arch. Exp. Pathol. Pharmakol.,* **1899**; *42*, 109-118; Meyer, H. Zur theorie der alkoholnarkose (III): der einfluss wechselnder temperature auf wirkungst¨arke und theilungscoefficient der narcotica. *Arch. Exp. Pathol. Pharmakol.*, **1901**, *46*, 338-346.

[15] Overton, E. *Studien ¨uber die narkose, zugleich ein beitrag zur allgemeiner Pharmakologie*. Jena: Gustav Fischer, **1901**.

[16] Ferguson, J. The use of chemical potential as indices of toxicity. *Proc. Roy. Soc. London Ser. B.*, **1939**, *127*, 387-404.

[17] Mullins, L. J. Some physical mechanisms in narcosis. *Chem. Rev.*, **1954**, *54*, 289-323.

[18] Molinengo, L.; Orsetti, M. The principle of nonspecificity and acute toxicity, *Trends Pharmacol. Sci.*, **1984,** *5*, 185-187.

[19] Hansch, C.; Fujita, T. ρ- σ-π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.*, **1964**, *86*, 1616-1626.

[20] Hammett, L. P. *Physical organic chemistry*. McGraw-Hill: New York, NY, pp. 404, **1940**.

[21] Taft, R. W. Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters, *J. Am. Chem. Soc.*, **1952**, *74*, 2729-2730.

[22] Kamlet, M. J.; Abboud, J. L. M.; Abraham, M. H.; Taft, R. W. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, π*, α, and β, and some methods for simplifying the generalized solvatochromic equation. *J. Org. Chem.*, **1983**, *48*, 2877- 2887.

[23] Franke, R.; Huebel, S.; Streich, W. J. Substructural QSAR approaches and topological pharmacophores, *Env. Health Perspect.*, **1985**, *61*, 239-255.

[24] Hickey, J. P.; Passino-Reader, D. R. Linear Solvation Energy Relationships: "Rules of Thumb" for Estimation of Variable Values, *Env. Sci. Technol.*, **1991**, *25*, 1753-1760.

[25] Famini, G. R.; Wilson, L. Y., Using theoretical descriptors in quantitative structure-property relationships: 3-carboxybenzisoxazole dexarboxylation kinetics. *J. Chem. Soc. Perkins Trans.*, **1994**, *2*, 1641-1650.

[26] Basak, S. C.; Gute, B. D.; Grunwald, G. D. Relative effectiveness of topological, geometrical, and quantum chemical parameters in estimating mutagenicity of chemicals, In: *Quantitative Structure-activity Relationships in Environmental Sciences VII*, F. Chen and G. Schuurmann, Eds., SETAC Press: Pensacola, FL. Chem. F., Schurrmam. G, Eds., **1998**; pp. 245-261.

[27] Basak, S. C. Mathematical descriptors for the prediction of property, bioactivity, and toxicity of chemicals from their structure: A chemical-cum-biochemical approach. *Curr. Comput. Aided Drug Des.*, **2013**, *9*, 449-462.

[28] Basak, S. C.; Magnuson, V. R.; Niemi, G. J.; Regal, R. R. Determining structural similarity of chemicals using graph-theoretic indices. *Discrete Appl. Math.*, **1988**, *19*, 17-44.

[29] Trinajstić, N. Chemical Graph Theory, 2nd ed., CRC Press: Boca Raton, FL, **1992**, pp. 352.

[30] Green, J. A. *Sets and Groups*, The English Language Book Society & Rutledge and Kegan Paul: Surrey, England, **1965**.

[31] Shannon, C. E., A Mathematical Theory of Communication, The Bell System Technical Journal, 1948, 27, 379-423.

[32] Bonchev, D.; Trinajstić, N. Information Theory, Distance Matrix, and Molecular Branching. *J. Chem. Phys.*, **1977**, *67*, 4517-4533.

[33] Bonchev, D. *Information Theoretic Indices for Characterization of Chemical Structures*, Research studies Press: Chichester, U.K., **1983**, pp. 249.

[34] Basak, S.C. Use of molecular complexity indices in predictive pharmacology and toxicology: A QSAR approach. *Med. Sci. Res.*, **1987**, *15*, 605-609.

[35] Raychaudhury, C.; Ray, S.K.; Ghosh, J.J.; Roy, A.B.; Basak, S.C. Discrimination of isomeric structures using information-theoretic topological indices. *J. Comput. Chem.*, **1984**, *5*, 581-588.

[36] Ray, S. K., Basak, S.C., Raychaudhury, C., Roy, A. B., Ghosh, J. J., A quantitative structure-activity relationship study of tumor inhibitory triazenes using bonding information content and lipophilicity, *IRCS Med. Sci.*, **1982**, *10*, 933-934.

[37] Hall Associates Consulting, *Molconn-Z Version 4.05*, Quincy, MA, **2003**.

[38] Basak, S. C.; Harriss, D. K.; Magnuson, V. R. 1988. *POLLY v. 2.3,* Copyright of the University of Minnesota, **1988**.

[39] Basak, S.C.; Grunwald, G.D.; Balaban, A.T. *TRIPLET*, Copyright of the Regents of the University of Minnesota, **1993**.

[40] Filip, P. A.; Balaban,T. S.; Balaban, A. T. A new approach for devising local graph invariants: derived topological indices with low degeneracy and good correlation ability. *J. Math. Chem.*, **1987**, *1*, 61-83.

[41] Basak, S. C.; Grunwald, G. D. *APProbe*, Copyright of the University of Minnesota, **1993**.

[42] Carhart, R.E.; Smith, D.H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.*, **1985**, *25*, 64-73.

[43] Tripos Associates, Inc. *Sybyl Version 6.2*, St. Louis, MO, **1995**.

[44] Stewart, J.J.P. *MOPAC Version 6.00, QCPE #455*, Frank J Seiler Research Laboratory: US Air Force Academy, CO, **1990**.

[45] M. J. Frisch *et al., Gaussian 98 (Revision A.11.2)*, Gaussian, Inc.: Pittsburgh, PA, **1998**.

[46] Gute, B. D.; Basak, S. C. Predicting acute toxicity of benzene derivatives using theoretical molecular descriptors: a hierarchical QSAR approach. *SAR QSAR Environ. Res.*, **1997**, *7*, 117-131.

[47] Gute, G. D.; Grunwald, G. D.; Basak, S. C. Prediction of the dermal penetration of polycyclic aromatic hydrocarbons (PAHs): A hierarchical QSAR approach, *SAR QSAR Environ. Res.*, **1999**, *10*, 1-15.

[48] Basak, S. C.; Mills, D. R.; Balaban, A. T.; Gute, B. D. Prediction of mutagenicity of aromatic and heteroaromatic amines from structure: A hierarchical QSAR approach, , *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 671-678.

[49] Hawkins, D. M.; Basak, S. C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environ. Toxicol. Pharmacol.*, **2004**, *16*, 37-44.

[50] Gute, B. D.; Basak, S. C.; Balasubramanian, K.; Geiss, K.; Hawkins, D. M. Prediction of halocarbon toxicity from structure: A hierarchical QSAR approach. *Environ. Toxicol. Pharmacol.*, **2004**, *16*, 121-129.

[51] Basak, S. C.; Natarajan, R.; Mills, D. Structure-activity relationships for mosquito repellent aminoamides using the hierarchical QSAR method based on calculated molecular descriptors. *ICCOMP'05 Proceedings of the 9th WSEAS International Conference on Computers*, **2005**, *7*, 958-963.

[52] Basak, S. C.; Mills, D.; Hawkins, D. M.; El-Masri, H. A. Prediction of tissue: air partition coefficients: A comparison of structure-based and property-based methods, *SAR QSAR Environ. Res.*, **2002**, *13*, 649-665.

[53] Basak, S. C.; Mills, D.; Mumtaz, M. M.; Balasubramanian, K. Use of topological indices in predicting aryl hydrocarbon (Ah) receptor binding potency of dibenzofurans: A hierarchical QSAR approach. *Indian. J. Chem.*, **2003**, *42A*, 1385-1391.

[54] Randic, M., Witzmann, F., M. Vracko, M., Basak, S. C. On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: Application to peroxisome proliferators. *Med. Chem. Res.*, **2001**, *10*, 456-479.

[55] Basak, S. C., Gute, B. D., Witzmann, F. Information-theoretic biodescriptors for proteomics maps: Development and applications in predictive toxicology. *ICCOMP'05 Proceedings of the 9th WSEAS International Conference on Computers*, **2005**, *7*, 996-1001.

[56] Vracko, M.; Basak, S. C.; Geiss, K.; Witzmann, F. Proteomics maps-toxicity relationship of halocarbons studied with similarity index and genetic algorithm. *J. Chem. Inf. Model.*, **2006**, *46*, 130-136.

[57] Basak, S.C.; Gute, B.D.; Geiss, K.T.; Witzmann, F. A. Information-theoretic biodescriptors for proteomics maps: Application to rodent hepatotoxicity. In: *Computation in Modern Science and Engineering, Proceedings of the International Conference on Computational Methods in Science and Engineering 2007 (ICCMSE 2007)*, Simos, T. E., Maroulis, G., Eds.; American Institute of Physics: Melville, New York, **2007**, 10-13.

[58] Hawkins, D. M.; Basak, S. C.; Kraker, J. J.; Geiss, K. T.; Witzmann, F. A. Combining chemodescriptors and biodescriptors in quantitative structure-activity relationship modeling. *J. Chem. Inf. Model.*, **2006**, *46*, 9-16.

[59] Basak, S. C., Gute, B. D. Mathematical descriptors of proteomics maps: Background and applications. *Curr. Opin. Drug Discov. Devel.*, **2008**, *11*, 320-326.

[60] Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. Royal Stat. Soc. Ser. B*, **1996**, *58*, 267-288.

[61] Zou, H. The Adaptive Lasso and Its Oracle Properties. *J. Amer. Stat. Assoc.*, **2006**, *101*, 1418-1429.

[62] Cook, R.D.; Li, B.; Chiaromonte, F. Envelope models for parsimonious and efficient multivariate linear regression. *Stat. Sinica*, **2010**, *20*, 927-1010.

[63] Sun, X.Q.; Chen, L.; Li, Y.Z.; Li, W.H.; Liu G.X.; Tu, Y.Q.; Tang, Y. Structure-based ensemble-QSAR model: a novel approach to the study of the EGFR tyrosine kinase and its inhibitors. *Acta Pharmacol. Sin.*, **2014**, *35*, 301-310.

[64] Hawkins, D.M.; Basak, S.C.; Mills, D. Assessing model fit by cross-validation. *J. Chem. Inf. Comput. Sci.*, **2003**, *3*, 579-586.

[65] Hawkins, D.M.; Basak, S.C.; Mills, D. QSARs for chemical mutagens from structure: ridge regression fitting and diagnostics. *Environ. Toxicol. Pharmacol.*, **2004**, *16*, 37-44.

[66] Basak, S.C.; Mills, D.; Hawkins, D.M.; Kraker, J.J. Proper statistical modeling and validation in QSAR: A case study in the prediction of rat fat-air partitioning, In: *Computation in Modern Science and Engineering, Proceedings of the International Conference on Computational Methods in Science and Engineering 2007 (ICCMSE 2007)*, Simos, T. E., Maroulis, G., Eds.; American Institute of Physics: Melville, New York, **2007**, 548-551.

[67] Sahigara, F.; Mansouri, K.; Ballabio, D; Mauri, A.; Consonni, V.; Todeschini, R. Comparison of different approaches to define the applicability domain of QSAR models. *Molecules*, **2012**, *17*, 4791-4810.

[68] Jaworska, J.; Nikolova-Jeliazkova, N.; Aldenberg, T. QSAR applicabilty domain estimation by projection of the training set descriptor space: A review. *Altern. Lab. Anim.*, **2005**, *33*, 445-459.

[69] Preparata, F.P.; Shamos, M.I. Convex Hulls: Basic Algorithms. In: *Computational Geometry: An Introduction*, Preparata, F.P., Shamos, M.I., Eds.; Springer-Verlag: New York, NY, **1991**, pp. 95-148.

[70] Worth, A.P.; Bassan, A.; Gallegos, A.; Netzeva, T.I.; Patlewicz, G.; Pavan, M.; Tsakovska, I.; Vracko, M. *The Characterisation of (Quantitative) Structure-Activity Relationships: Preliminary Guidance*. ECB Report EUR 21866 EN, European Commission, Joint Research Centre: Ispra, Italy, **2005**, p. 95.

[71] Majumdar, S.; Basak, S.C.; Grunwald, G.D. Adapting interrelated two-way clustering method for quantitative structure-activity relationship (QSAR) modeling of mutagenicity/non-mutagenicity of a diverse set of chemicals. *Curr. Comput. Aided Drug Des.*, **2013**, *9*, 463-471.

[72] Tang, C.; Zhang, L.; Zhang, A.; Ramanathan, M. Interrelated Two-way Clustering: An Unsupervised Approach for Gene Expression Data Analysis, In: *Proceedings of BIBE 2001: 2nd IEEE International Symposium on Bioinformatics and Bioengineering, Bethesda, Maryland, November 4-5, 2001*, Bilof, R.; Palagi, L., Eds.; IEEE Computer Society: Los Alamitos, CA, **2001**, pp. 41-48.

[73] Basak, S.C.; Majumdar, S. Prediction of mutagenicity of chemicals from their calculated molecular descriptors: a case study with structurally homogeneous versus diverse datasets. *Curr. Comput. Aided Drug Des.*, **2015**, *11*, 117-123.

[74] Debnath, A.K.; Debnath, G.; Shusterman, A.J.; Hansch, C. A QSAR Investigation of the Role of Hydrophobicity in Regulating Mutagenicity in the Ames Test: 1. Mutagenicity of Aromatic and Heteroaromatic Amines in *Salmonella typhimurium* TA98 and TA100. *Environ. Mol. Mutagen.*, **1992**, *19*, 37-52.

[75] Basak, S.C. Molecular Similarity and Hazard Assessment of Chemicals: A Comparative Study of Arbitrary and Tailored Similarity Spaces. *J. Eng. Sci. Manage. Educ.*, **2014**, *7(III)*, 178-184.

[76] Basak, S. C.; Roy, A. B.; Ghosh, J. J. Study of the structure-function relationship of pharmacological and toxicological agents using information theory, In: *Proceedings of the Second International Conference on Mathematical Modelling*, Avula, X.J.R., Bellman R., Luke, Y.L. and Rigler, A.K., Eds., pp 851-856, University of Missouri-Rolla: Rolla, Missouri, **1980**.

[77] Basak, S. C., Magnuson, V. R., Niemi, G. J., Regal, R. R. and Veith, G. D. Topological indices: their nature, mutual relatedness, and applications. *Mathl. Modelling*, **1987**, *8*, 300-305.

[78] Basak, S. C., Gieschen, D. P., and Magnuson, V. R. A quantitative correlations of the LC_{50} values of esters in Pimephales promelas using physicochemical and topological parameters. *Environ. Toxicol. Chem.*, **1984**, *3*, 191-199.

[79] Basak, S. C., Monsrud, L. J., Rosen, M. E., Frane, C. M., Magnuson, V. R. A comparative study of lipophilicity and topological indices in biological correlation. *Acta Pharm. Yugosl.*, **1986**, *36*, 81-95.

[80] Basak, S. C., Gieschen, D. P., Magnuson, V. R., Harriss, D. K. Structure-activity relationships and pharmacokinetics : A comparative study of hydrophobicity, van der Waals' volume and topological parameters. *IRCS Med. Sci.*, **1982**, *10*, 619-620.

[81] Basak, S. C., Harriss, D. K., Magnuson, V. R. Comparative study of lipophilicity versus topological molecular descriptors in biological correlations. *J. Pharm. Sci.*, **1984**, *73*, 429-437.

[82] SAS Institute Inc. *SASISTAT User's Guide, Release 6.03 Edition*. SAS Institute Inc.: Cary, NC, **1988**, Chapter 34, pp. 949-965.

[83] Basak, S. C. Grunwald, G. D. Tolerance space and molecular similarity. *SAR QSAR Environ. Res.*, **1995**, *3*, 265-277.

[84] Hawkins, D., Basak, S. C., Shi, X. QSAR with few compounds and many features. *J. Chem. Inf. Comput. Sci.*, **2001**, *41*, 663-670.

[85] Wiener, H. Structural determination of paraffin boiling point. *J. Am. Chem. Soc.*, **1947**, *69*, 17-20.

[86] Hosoya, H. Topological Index. A newly proposed quantity characterizing the topological nature of structural isomers of saturated hydrocarbons. *Bull. Chem. Soc. Jpn.*, **1971**, *44*, 2332-2339.

[87] Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.*, **1975**, *97*, 6609-6615.

[88] Balaban, A. T. Distance Connectivity Index. *Chem. Phys. Lett.*, **1982**, *89*, 399-404.

[89] Basak, S. C.; Bhattacharjee, A. K.; Vracko, M. Big data and new drug discovery: Tackling "big data" for virtual screening of large compound databases. *Curr. Comput. Aided Drug Des.*, **2015**, *in press*.

[90] Big Data for Development: Challenges & Opportunities, http://www.unglobalpulse.org/projects/BigDataforDevelopment

[91] Big data in global health: improving health in low- and middle-income countries, http://www.who.int/bulletin/volumes/93/3/14-139022/en/

[92] Ward, J. S.; Barker, A. Undefined by data: a survey of big data definitions. Ithaca: Cornell University Library, **2013**. Available from: http://arxiv.org/pdf/1309.5821v1.pdf [Accessed 2015/Aug 9].

[93] Richards, W. G. Virtual screening using grid computing: the screensaver project. *Nat. Rev. Drug Discov.*, **2002**, *1*, 551-555

[94] Basak, S.C.; Balasubramanian, K.; Gute, B.D.; Mills, D; Gorczynska, A; Roszak, S. Prediction of cellular toxicity of halocarbons from computed chemodescriptors: A hierarchical QSAR approach. *J. Chem. Inf. Comput. Sci.*, **2003**, *43*, 1103-1109.

[95] Basak, S. C.; Mills, D.; Gute, B. D.; Grunwald, G. D.; Balaban, A. T. Applications of topological indices in the property/bioactivity/toxicity prediction of chemicals, In: *Topology in Chemistry: Discrete Mathematics of Molecules*, Rouvray, D. H. and King, R. B., Eds.; Horwood Publishing Limited: Chichester, England, **2002**, pp. 113-184.

[96] Basak, S. C., Mills, D., Natarajan, R., Gute, B. D. Predicting chemical reactivity and bioactivity from structure: A mathematical-cum-computational approach. In: *Chemical Reactivity Theory: A Density Functional View*, Chattaraj, P. K., Ed.; CRC Press, **2009**, pp. 479-502.

[97] Basak, S. C., Zhu, Q., Mills, D. Prediction of Anticancer Activity of 2-phenylindoles: Comparative Molecular Field Analysis Versus Ridge Regression using Mathematical Molecular Descriptors. *Acta Chim. Slov.*, **2010**, *57*, 541-550.

[98] Basak, S. C., Mills, D. Quantitative Structure-Activity Relationship Studies of Boron-Containing Dipeptide Proteasome Inhibitors Using Calculated Mathematical Descriptors. *J. Math Chem.*, **2011**, *49*,185-200.

The authors are thankful to Bentham Science Publishers for their kind permission to reproduce in this book chapter by Basak and Majumdar Tables 1, 2, and 3 from the paper "Prediction of Mutagenicity of Chemicals from Their Calculated Molecular Descriptors: A Case Study with Structurally Homogeneous versus Diverse Datasets," by Subhash C. Basak and Subhabrata Majumdar, Current Computer-Aided Drug Design, 2015, 11, 117-123.

CHAPTER 12

Recent Advances in the Assessment of Druglikeness Using 2D-Structural Descriptors

Hariharan Rajesh[1,2], Lakshminarasimhan Rajagopalan[1] and Vellarkad N. Viswanadhan[1,*]

[1]Department of Computational Chemistry, Jubilant Biosys Limited, Bangalore 560 022, India and [2]Shanmugha Arts, Science, Technology, and Research Academy, Thanjavur 613 402, TN, India

Abstract: A review of methods employed for the assessment of druglikeness using 2D structural and atom type descriptors is presented. These methods are classified as Drug-like Filters (DLFs) and Druglike Indices (DLIs), depending on the characterization of druglikeness, using known drug and non-drug databases. The DLF methods specify a set of rules based on calculated property distributions, whereas the DLI methods aim to assess druglikeness through a single number derived from multiple descriptors. A review of ranges calculated from property profiles of known drugs is given, along with a careful re-assessment for twenty five descriptors based on an analysis of a recent drug database. A discussion of future direction for the development and utility of these approaches is presented.

Keywords: Lead-likeness, drug likeness, structural descriptors, drug like index, atom type diversity, relative drug-likeness potential, ALOGP, UALOGP, druglike descriptors, chembridge database, drugs database, drug properties, atom classification, structural diversity, lead optimization.

INTRODUCTION

Tapping the knowledge in available drug databases is clearly a vital aspect of new drug discovery, aiding in various phases of pre-clinical drug discovery, starting from generating hits and lead identification to lead optimization and pre-clinical candidate selection. Though the chemical constitution of a drug will always be unique in some respect, analysis of structural descriptors at atomic, moiety and

***Corresponding author Vellarkad N. Viswanadhan:** Department of Computational Chemistry, Jubilant Biosys Limited, #96, Industrial Suburb, 2nd Stage, Yeshwantpur, Bangalore 560 022, India; Tel: +91-80-6662 8908; Fax: +91-80-6662 8333; E-mail: vellarkad_viswanadhan@jubilantbiosys.com.

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50012-6

molecular levels enabled several useful characterizations of drugs, leading to metrics of druglikeness. The present chapter reviews various developments in the assessment of druglikeness since the seminal publications of Lipinski [1], and presents some new developments in the analysis of drug properties, especially at the atomic level. Such assessments employed a multiplicity of approaches [1-30], which can be broadly classified into two categories: Drug-like filters (DLF) and Drug-like indices (DLI). The former approach specifies a set of property ranges or preferences based on distributions of physico-chemical and structural properties of drugs through an analysis of drug databases, which form a DLF. The latter approach employs a fitting procedure using drug and non-drug databases, and structural descriptors to derive a DLI, which is a single index for the assessment of relative druglikeness.

DEVELOPMENT OF DRUG-LIKE FILTERS (DLFS) FROM STRUCTURAL DESCRIPTORS

The first DLF, the Ro5 (Rule of 5) developed by Lipinksi and co-workers [1] is specific with regard to the upper bounds of four important properties considered and how they are computationally assessed, based on an analysis of orally absorbed drugs. Ro5 is satisfied for a given molecule when, (i) calculated Log P less than 5, (ii) number of H-bond donors is less than 5 (iii) number of H-bond acceptors is less than 10 and (iv) molecular weight is less than 500. Of these, we note that, log P assessment is dependent on the method used and a lower value of 4.15 is indicated when Moriguchi Log P method is used [1], though better methods exist for accurate calculations of Log P such as ALOGP and CLOGP [see *e.g.,* 31-33]. Ro5 filter, however, does not apply when substrates for transporters, natural products and biological drugs are analyzed. Subsequent property profiling of drug databases led to important extensions and inclusion of additional properties in the creation of DLFs, such as molar refractivity, atom counts [3] and PSA [28-30] for such rule-based assessments.

Ghose *et al.* [3, 4] showed that property ranges also depended on the class of drugs, and compiled ranges for a number of properties, including topological, physicochemical and other structural descriptors. Ghose *et al.* [3] worked out the ranges of properties occupied by 80% and 50% of known drugs, and showed that molecular weight and calculated Log P of many drugs exceed the Ro5 limits by their estimates. Thus, the Log P range (based on ALOGP98 calculations [31, 32]) is shown to be -0.4 to 5.6, exceeding Ro5 limit of 5.0. Based on these ranges, over 20% of anti-hypertensive drugs exceed the Ro5 limit on molecular weight. DLF developed by Ghose *et al.* [3] has the following properties (i) Log P between -0.4 and 5.6; (ii) Molar Refractivity between 40 and 130; (iii) Molecular weight between 160 and 480; (iv) total number of atoms between 20 and 70; (v) structurally a combination of several of the following groups: phenyl, heterocyclic ring, aliphatic amine, alcoholic hydroxyl, a carboxy ester, a keto group; and (vi) absence of reactive group(s) that causing instability in physiological buffer.

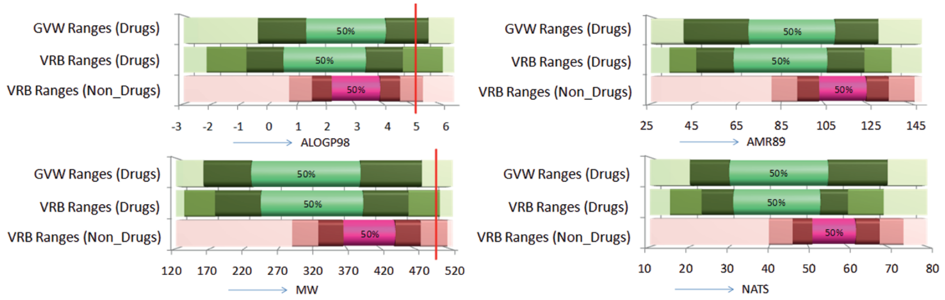


Figure 1: (a) ALOGP98, (b) AMR89, (c) MW, and (d) NATS ranges covering different fractions of Drugs and non drugs, as defined in Viswanadhan *et al.* [6]. The middle bright colored part covers 50% of each database. Dark colored extensions on either side constitute another 30%, covering 80% range, and another 15% is added by further light colored extensions, covering 95% range. From top, ranges based on GVW criteria [3] are shown, followed by the ranges obtained by Viswanadhan *et al.* [6] (VRB ranges). Lipinski range cutoff is shown as a bright red line for calculated Log P and MW.

Recently, Viswanadhan *et al.* [6] analyzed property preferences at atomic and molecular levels for drugs, leads, and nondrugs, to be considered for library design and lead optimization in drug discovery, using several drug and non-drug databases [34-36]. Fig. (**1**) shows a comparison of DLFs by Ghose *et al.* [3] and Viswanadhan *et al.* [6], with respect to four physicochemical properties. Work by Viswanadhan *et al.* [6] shows that the 95% LogP (by the ALOGP98 method [31, 32]) range is 2.2 to 6.1, while the Ro5 excludes at least 10% of orally absorbed known drugs, over a decade ago. The absence of lower limit for log P in Ro5 may additionally cause the inclusion of highly hydrophilic compounds as druglike, as the rules permit up to fifteen polar atoms in a molecule. Interestingly, non-drugs also strongly overlap the ranges shown for drugs and this requires some explanation. Of the several commercially available compound databases available in ZINC [34], Viswanadhan *et al.* [6] chose the Chembridge database [36], which is quite large (>230000 compounds) with a diverse collection of synthesizable organic compounds. Two independent random sets of 10000 molecules each, which pass the filters as described below, were utilized for assessment. These sets were filtered to exclude highly lipophilic (calculated log P > 8.0) and highly hydrophilic compounds (calculated log P < -5.0), similar to earlier analyses [3]. Also excluded were compounds which are unusually small (< 100 MW or < 14 atoms or < 10 heavy atoms) or large (> 800 MW or > 100 atoms), polymers, peptides, quaternary ammonium, multiple acids and phosphates. This additional filtering excluded entries which are not of particular interest as small molecule drug candidates, which would be of interest to synthetic and medicinal chemists needing guidance for compound acquisition and screening (virtual or real high throughput screening). Among other developments, a few of the significant analyses may be mentioned. Bemis and Murcko [17] identified and analyzed molecular frameworks and side chains found in drugs. Kutchikian *et al.* [21] developed a method for *de novo* generation of druglike molecules. Hann *et al.* [23] analyzed molecular complexity and its impact on the probability of finding leads for drug discovery. Recent studies identified *distinct* physicochemical profiles of different drug classes, *e.g.,* respiratory drugs, marketed *vs*. development drugs *etc.* [24, 25].

Unlike heteroatom counts, Polar Surface Area (PSA) is a direct, single number measure of overall polar character and is also shown to be useful in the assessment of oral absorption [28]. Egan *et al.* [28] proposed an elliptical filter, in a two dimensional plot of Log P and PSA. This construct is consistent with earlier work [29-30] that identified upper limits of PSA for oral absorption as 140 Å^2 or 120 Å^2.

Through several observations and careful analyses, Leeson [2] makes a strong case for lowering lipophilicity of small molecule as a means for attaining a small (~5%) improvement in attrition, which could double the output of new medicines and reduce compound-related toxicological attrition.

Table 1: Drug like descriptors, their optimal values and preferred range

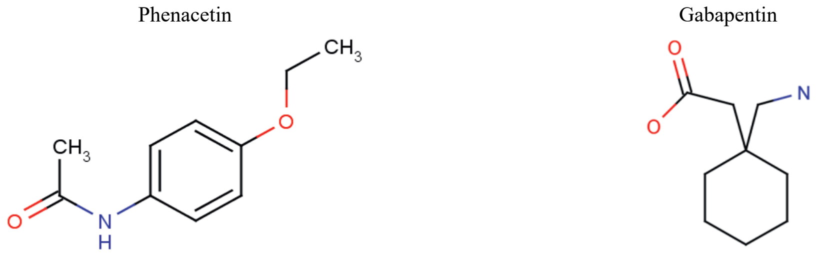
Descriptors	Best Value	Preferred Range	Phenacetin	Gabapentin
Number of Non-H atoms	22	17 – 27	13	12
Number of SSSR	3	1 – 3	1	1
Molecular Cyclized Degree (MCD)	11	7 – 13	19	20
Number of Non-H Rotatable Bonds	6	4 – 9	4	3
Number of Non-H Polar Bonds	6	5 - 10	5	3
Number of Terminal Methyl Groups	0	0 – 1	4	3
Number of N-H Donors	1	0 – 1	1	1
Number of O-H Donors	0	0 – 0	0	1
Number of Hydrogen Bond Donors	1	0 – 2	1	2
Number of Hydrogen Bond Acceptors	3	2 – 4	3	3
Number of O & N Atoms	4	3 – 6	3	4
Number of 2 Degree Acyclic Atoms	1	1 – 3	3	2
Number of 3 Degree Acyclic Atoms	0	0 – 1	1	1
Number of non substituted ring atoms	8	4 – 9	4	5
Number of substituted ring atoms	6	3 – 7	2	0
Number of one level bonding pattern	0	0-1	1	1
Number of two level bonding pattern	0	0-0	0	0
Number of three level bonding pattern	0	0-0	0	0
Number of Building Blocks	2	2 – 4	3	3
Number of Aromatic Systems	1	0 – 1	1	0
Number of Cyclic Building Blocks	1	1 – 1	1	1
Number of Linkers	0	0 – 0	0	0
Number of Caps	2	1 – 3	2	2
Maximum SSSR size	6	5 – 6	6	6
Maximum Cap Size	1	1 – 4	4	4
Druglike Index (DLI)			77.61	68.06

DRUGLIKE INDEX (DLI) FROM STRUCTURAL DESCRIPTORS

The first set of efforts to quantify druglikeness as a single number measure employed non-linear approaches such as neural networks [4-16, 18, 20-21] using molecular descriptors such as atom pair frequencies [8], whole molecule properties [9], ALOGP atom types [10], and ISIS keys [14]. Among the earliest methods of calculating DLI were the approaches of Sadowski and Kubinyi [10], and Ajay and Murcko [14], using neural networks. Sadowski and Kubinyi's [10] method employed ALOGP atom types [31,32] as descriptors, to construct a feed forward neural network, trained based on the back propagation with a momentum scheme. The method developed by Ajay and Murcko [14], used a similar procedure with ISIS keys as structural descriptors. In another of the earliest approaches, Xu and Stevenson [9] performed an analysis of structural diversity in drugs using selected descriptors, and computed their distributions in known drugs. Based on these distributions, a drug-like compound cluster center is formed. The cluster centers are used to rank compounds in any library in terms of their "drug-like" indices (DLI) The DLI was defined using the following equation, where n refers to number of descriptors.

$$DLI = \sqrt[n]{\prod_{i=1}^{n} Score(Descriptor(i))} \qquad (1)$$

The drug-like cluster center is developed from the distributions of 25 selected structural descriptors, identified in Table **1** [9]. The curve of relative population *versus* the DLI value is used as a simple means to assess the structural diversity and druglikeness of a library. Table **1** also shows the cluster center (best value) and preferred range (minimal range occupied by 50% of known drugs) calculated from the distribution of these properties in a recent drug database (Drugs_all database taken from reference 6).



Phenacetin

Gabapentin

Figure 2: Structures of Phenacetin and Gabapentin.

Hutter [8] developed a unique approach for DLI, based on the distribution of atom types and their pair-wise combinations in known drugs and non-drugs. A statistical analysis of the occurrence probabilities of atom types was used to derive a DLI score. Although any kind of fitting is not done, drugs were predicted with an accuracy of over 70%. This work highlighted the significance employing atom types and their pairs as descriptors for the assessment of druglikeness and DLI calculations. Furthermore, Hutter and coworkers also described ways to gradually filter molecules for druglikeness using a multiplicity of approaches [22].

STRUCTURAL DESCRIPTORS FOR THE ANALYSIS OF DRUGLIKENESS: ATOM TYPE DIVERSITY

Recent work on characterization of intrinsic structural diversity [6] was based on the concept that atom classification is hierarchical, with elemental types at the primary level. ALOGP [31, 32], and UALOGP [6] classifications constituted secondary and tertiary levels of finer differentiation. UALOGP [6] representation considered hydrogen atoms implicitly, *i.e.*, only heavy atoms were used for assessment. Three structural diversity measures were defined for a molecule with NHATS heavy atoms.

$$P_1\text{=Number of element types / NHATS} \quad (1)$$

$$P_2\text{=Number of heavy atom types / NHATS} \quad (2)$$

$$P_3\text{=Number of united atom types / NHATS} \quad (3)$$

Here, the P's define the atom type diversity based on elemental types (equation 2, P_1), ALOGP [23] heavy atom types (equation 3, P_2) and united atom (UALOGP [6]) types (equation 4, P_3). Atom type diversity (ATD) was defined as the product of P_1, P_2 and P_3, times 100 (a scale factor).

$$\text{ATD} = P_1.P_2.P_3 \text{ X } 10^2 \quad (4)$$

This definition ensured equal weight to each level of atom classification. Viswanadhan *et al.* [6] showed that the profiles of ATD are distinct for drugs and non-drugs. Their analysis indicates that drugs are seen to have higher ATD scores, though ~50% of drugs have ATD scores below 5. For non-drugs, this percentage is much higher (> 80%). For non-drugs, scores greater than 7 are much rarer (4 % for non-drugs *vs*. 25% for drugs). Thus, drugs are seen to be significantly richer with regard to atom type diversity [6]. The average ATD values for non-drugs and drugs are 3.59 and 5.78 respectively. Using this parameter, Viswanadhan *et al.* [6] also showed that leads are structurally more diverse than drugs. As examples, drugs shown in Fig (**2**) may be considered, which includes *Phenacetin* and *Gabapentin*. *Phenacetin*, one of the oldest drugs introduced in 1887, was used as an analgesic but later discontinued due to its potential for carcinogenicity [37] and it was replaced by a safer alternative *Paracetamol* which is a metabolite of this drug. The oral drug *Gabapentin* [38] was originally developed for the treatment of epilepsy, and currently it is widely used to relieve pain, and neuropathic pain. *Gabapentin* was originally approved by the U.S. Food and Drug Administration (FDA) in 1994 for use as an adjunctive medication to control partial seizures. In 2002, an indication was added for treating postherpetic neuralgia other painful neuropathies, and nerve-related pain. These molecules have relatively ATD values of 9.8 and 9.7.

ATOMIC LEVEL ASSESSMENT OF DRUGLIKENESS

In order to quantitatively elucidate druglikeness at the atomic level, Viswanadhan *et al.* [6] undertook a comprehensive analysis of atom types in known drugs, leads and a representative set of non-drugs. The starting point for this analysis was ALOGP, an atom type representation [31, 32], developed and validated for calculation molecular properties such as lipophilicity and molar refractivity, and also for QSAR applications. The latest version of this representation contains 44 carbon types, 10 hydrogen types, 9 oxygen types, 2 selenium types, 6 types for each of the halogens, 5 sulfur types, 1 type each for silicon and boron, and 6 phosphorus types. A more elaborate united atom representation (UALOGP), with implicit hydrogens was developed from this, for a detailed characterization of druglikeness [6].

Tables **2** and **3** show the distributions of atom types, for the drug and non-drug datasets of Viswanadhan *et al.* [6]. These types are defined in Table **4**. Table **2** (**a**) shows the distribution of ALOGP atom types in different fractions of the drug database. Table **2**(**b**) shows the distribution profile of UALOGP united atom types. Tables **3**(**a**) and **3**(**b**) show similar distribution profiles of atom types for the non-drug database. From these tables, it is easy to delineate what percentage range of a database contains a given atom type. For example, 15 atom types are not found in the drug database. Type 24 (benzene type carbon without an R-group attached) is found in 80% of the database, making it the most abundant heavy atom type in drugs, followed by the types 6 (methylene linker carbon attached to a carbon and a heteroatom), 25, 26 (benzene type carbon with an attached R group) and 58 (carbonyl oxygen) which are found in 60 to 70% of all drugs. Type 58 is the most abundant hetero atom type in the drug database. From Table **2**(**b**) it is seen that the type 24a (70 - 80%) is the most abundant carbon type followed by type 6a observed in 50 - 60 % of the drug database. Here type 6b (the alpha carbon subtype) is present in less than 10% of the drug database. This type 6a is one of the most abundant type observed in the non-drug database (6a is found in 80 - 90% while type 6b is found in 10-20% of the non-drug database).

Table 2(a): Distribution of atom types in the Drugs_all database. Atom types (original ALOGP) found in different database fractions or percentages

Percent of the Database	Atom Types Found
0%	64 65 92 93 98 101 102 103 104 112 113 114 115 116 119
0 - 10%	7 10 12 13 14 15 18 19 20 21 22 23 29 30 31 32 33 34 35 36 37 39 42 43 44 54 55 61 63 66 67 68 69 70 71 76 77 78 81 82 83 84 85 86 87 88 90 91 94 95 96 97 99 100 106 108 109 110 111 117 118 120
10 – 20%	4 9 11 16 17 27 28 38 41 49 53 57 59 73 74 89 107
20 – 30%	3 5 48 62 75
30 – 40%	56 60
40 – 50%	8 46 72 79
50 – 60%	1 2 40 51 52
60 – 70%	6 25 26 58
70 – 80%	24
80 – 90%	47 50
90 - 100%	

Table 2(b): United (hydrogen-filled) atom types found in different percentages (fractions) of the Drugs all drug database considered

Percent of the Database	Atom Types Found
0%	1e 1f 7b 8c 8d 8e 8f 16b 16c 16d 16e 16f 21b 24b 27c 33b 36a
0 - 10%	1b 1d 2e 2f 3b 3d 3e 3f 6b 7a 9a 9b 15a 15b 18a 18b 21a 27a 27b 33a 36b 36c 37a 37b 67a 67b 73a 73b 74a 79b 106a 106b
10 – 20%	2b 2d 3a 3c 8b 16a 57a 72b 74b
20 – 30%	1c 2a 5a

Table 2(b): contd…

30 – 40%	1a 2c 72a 79a
40 – 50%	8a 56a
50 – 60%	6a
60 – 70%	
70 – 80%	24a
80 – 90%	
90 - 100%	

Table 3(a): Distribution of atom types across the non-drug database: Atom types (original ALOGP) found in different database fractions or percentages

Percent of the Database	Atom Types Found
0%	23 63 64 65 77 78 80 86 87 88 91 92 93 95 96 97 98 99 100 101 102 103 104 106 109 111 112 115 116 117 118 119 120
0 - 10%	4 7 9 10 11 12 13 14 15 16 17 18 19 20 21 22 29 30 32 35 36 37 38 39 41 42 43 44 55 57 61 62 66 67 69 70 74 76 81 82 83 85 89 90 94 108 110
10 – 20%	31 33 34 54 56 68 84
20 – 30%	49 51 59 71 73 107
30 – 40%	3
40 – 50%	27 28 48
50 – 60%	1 5 8 46 53 60 79
60 – 70%	75
70 – 80%	26 40 52 72
80 – 90%	2 6 24 25 47 50 58
90 - 100%	

Table 3(b): United (hydrogen-filled) atom types found in different percentages (fractions) of the non-drug database considered

Percent of the Database	Atom Types Found
0%	1d 1e 1f 5b 7b 8b 8c 8d 8e 8f 9a 15b 16b 16c 16d 16e 16f 21b 24b 27c 30b 33b 36a 37b 67b 74a 106a 106b
0 - 10%	2e 2f 3a 3e 3f 7a 9b 15a 16a 18a 18b 21a 30a 36b 36c 37a 42a 57a 66a 67a 73a 74b 79b
10 – 20%	1b 3b 3c 3d 6b 8b 27a 33a 56a 73b
20 – 30%	1a 1c 2d

Table 3(b): contd…

30 – 40%	2a 2b 27b
40 – 50%	8a 72a
50 – 60%	5a 72b 79a
60 – 70%	2c
70 – 80%	
80 – 90%	6a 24a
90 - 100%	

Types 47 and 50 (hydrogen attached to a hetero atom) are highly abundant (80-90%) in both the databases. The most abundant donors are type 72 (carboxamide NH – 40-50%), and alcohol oxygen (type 56 - 30-40%), followed by phenolic oxygen (type 57 – 10-20%). Type 72 is more frequent (70-80%) in non-drugs relative to drugs (40-50%). The type 56 is less frequent (10-20%) in the non-drugs. Subtype 2c (C^0sp^3, with one hetero-atom attached to its next carbon) is found abundantly (60-70%) among non-drugs, whereas only 30-40% of drug database contains 2c. The subtype 72b is more frequent (50-60%) in non-drugs relative to drugs (10-20%), while the type 57 (oxygen as in phenolic hydroxyl) is less frequent (0-10%) in non-drugs. Among halogens, types 89(Cl attached to C^1Sp^2) is most abundant in drugs and type 84 (F attached to C^1Sp^2) is most abundant in non-drugs though these types occur only in 10-20% of the compounds studied. Understandably, Br and I are highly infrequent in drugs, as they significantly increase molecular weight, leading to undesirable characteristics such as poor intestinal absorption.

Analysis Drug Properties at the Atomic Level

The foregoing observations, led to the extension of the concept of druglikeness to atom types as well [6]. Viswanadhan *et al.* [6] defined Relative Druglikeness Potential (RDP_i) of each atom type as foll

$$RDP_i=p_{i,d} / p_{i,n} \qquad (5)$$

where *i* refers to the atom type, $p_{i,d}$ is the percentage occurrence of type *i* in the drug database and $p_{i,n}$ is the percentage occurrence of type *i* for the nondrug database (an approximation for expectation value, based on a typical distribution in commercial small molecule collections). Values of $RDP_i > 1$ indicate preferred types in drugs, while values <1 indicates the opposite. Viswanadhan *et al.* [6]

used a representative collection of 10000 molecules from Chembridge collection [36] to define non-drug set, that satisfies the preliminary filters used to define the set of small molecule drugs. To obtain more robust estimates of RDP_i values, we have calculated these values based on 4 different random samples of the Chembridge database satisfying the preliminary filters mentioned above.

Table 4: Statistical properties of atom types calculated for drugs. Distribution and description of atom types using all atom (ALOGP98) and united atom (UALOGP) representations are shown, along with mean occurrence per drug, percentage occurrence, and Relative Druglikeness Potential (RDP_i) values (equation 6)

Atom Type*	Description	Mean Occurrence Value (S.D.)	Percentage Occurrence X 10^2	RDP_i
C in				
1	:CH_3R,CH_4	1.0(1.3)	2.5	1.35
1a	C^0sp^3, having no X attached to next C	0.6(1.1)	1.4	1.61
1b	α – C	0.1(0.3)	0.2	1.00
1c	C^0sp^3, having 1 X attached to next C	0.4(0.8)	0.9	1.18
1d	C^0sp^3, having 2 X attached to next C	0.0(0.2)	0.0	0.0
1e	C^0sp^3, having 3 X attached to next C	0.0(0.1)	0.0	0.0
2	:CH_2R_2	1.8(2.3)	4.2	0.9
2a	C^0sp^3, having no X attached to next C	0.8(1.6)	1.8	1.2
2b	α – C	0.2(0.5)	0.5	0.8
2c	C^0sp^3, having 1 X attached to next C	0.6(1.0)	1.5	0.8
2d	C^0sp^3, having 2 X attached to next C	0.1(0.4)	0.3	0.6
2e	C^0sp^3, having 3 X attached to next C	0.0(0.1)	0.0	0.0
2f	C^0sp^3, having 4 X attached to next C	0.0(0.1)	0.0	0.0
3	:CHR_3	0.5(1.0)	1.2	1.5
3a	C^0sp^3 having no X attached to next C	0.3(0.8)	0.6	3.3
3b	α – C	0.1(0.3)	0.2	0.8
3c	C^0sp^3, having 1 X attached to next C	0.1(0.4)	0.3	1.3
3d	C^0sp^3, having 2 X attached to next C	0.0(0.2)	0.1	1.0
3e	C^0sp^3, having 3 X attached to next C	0.0(0.1)	0.0	0.0
3f	C^0sp^3, having 4 X attached to next C	0.0(0.1)	0.0	0.0
4	:CR_4	0.2 (0.5)	0.5	2.8
5	:CH_3X	0.4 (0.8)	1.1	0.9
6	:CH_2RX	1.4 (1.7)	3.4	0.7
6a	C^1sp^3, C^0sp^2	1.3 (1.7)	3.2	0.7

Table 4: contd…

6b	α – C	0.1 (0.3)	0.2	0.6
7	:CH_2X_2	0.0 (0.2)	0.1	0.3
7a	Any of C^2sp^3, C^1sp^2, C^0sp	0.0 (0.2)	0.1	0.3
8	:CHR_2X	0.9(1.3)	2.3	1.6
8a	C^1sp3, C^0sp2	0.7(1.2)	1.7	1.5
8b	α – C	0.2(0.5)	0.5	2.2
8c	C^0sp^3, having 1 X attached to next C	0.0(0.1)	0.0	0.0
8d	C^0sp^3, having 2 X attached to next C	0.0(0.1)	0.0	0.0
8e	C^0sp^3, having 3 X attached to next C	0.0(0.0)	0.0	0.0
8f	C^0sp^3, having 4 X attached to next C	0.0(0.0)	0.0	0.0
9	:$CHRX_2$	0.1 (0.3)	0.3	0.0
9a	C^2sp^3, C^1sp^2, C^0sp	0.1 (0.3)	0.3	0.0
9b	Attached H's when C is α – C	0.0 (0.1)	0.0	0.0
10	:CHX_3	0.0 (0.1)	0.0	0.0
11	:CR_3X	0.2 (0.4)	0.4	2.2
12	:CR_2X_2	0.0 (0.2)	0.1	0.0
13	:CRX_3	0.0 (0.2)	0.1	1.0
14	:CX_4	0.0 (0.0)	0.0	0.0
15	: =CH_2	0.0 (0.2)	0.1	0.0
15a	C^1sp^3, C^0sp^2	0.0 (0.2)	0.1	0.0
15b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.0 (0.0)	0.0	0.0
16	: =CHR	0.3(0.9)	0.8	2.7
16a	C^1sp^3, C^0sp^2	0.2(0.7)	0.5	1.9
16b	attached H's when C is an α – C	0.1(0.4)	0.3	2.3
16c	C^0sp^3, having 1 X attached to next C	0.0(0.0)	0.0	0.0
16d	C^0sp^3, having 2 X attached to next C	0.0(0.0)	0.0	0.0
16e	C^0sp^3, having 3 X attached to next C	0.0(0.0)	0.0	0.0
17	:=CR_2	0.2 (0.6)	0.5	3.2
18	:=CHX	0.1(0.2)	0.1	1.0
18a	C^2sp^3, C^1sp^2, C^0sp	0.0(0.2)	0.1	1.0
18b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.0(0.1)	0.0	0.0
19	:=CRX	0.1 (0.4)	0.2	1.7
20	:=CX_2	0.0 (0.1)	0.0	0.0
21	:≡CH	0.0 (0.1)	0.0	0.0
21a	C^2sp^3, C^1sp^2, C^0sp	0.0(0.1)	0.0	0.0
21b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.0(0.0)	0.0	0.0

Table 4: contd…

22	:≡CR,R=C=R	0.0 (0.2)	0.0	0.0
23	:≡CX	0.0 (0.0)	0.0	0.0
24	:R- -CH- -R	4.1 (3.6)	9.8	1.0
25	:R- -CR- -R	1.3 (1.3)	3.0	0.9
26	:R- -CX- -R	1.4 (1.5)	3.3	1.3
27	:R- -CH- -X	0.2(0.5)	0.4	0.5
27a	C^2sp^3, C^1sp^2, C^0sp	0.0(0.2)	0.1	0.5
27b	C^3sp^3, $C^{2-3}sp2$, $C^{1-3}sp$	0.1(0.5)	0.3	0.5
28	:R- -CR- -X	0.1 (0.4)	0.3	0.5
29	:R- -CX- -X	0.1 (0.3)	0.2	1.7
30	:X- -CH- -X	0.0 (0.2)	0.1	0.3
31	:X- -CR- -X	0.1 (0.3)	0.2	0.6
32	:X- -CX- -X	0.0 (0.2)	0.1	1.0
33	:R- -CH···X	0.0 (0.2)	0.1	0.5
34	:R- -CR···X	0.0 (0.2)	0.1	0.5
35	:R- -CX···X	0.0 (0.1)	0.0	0.0
36	:Al-CH=X			
36a	C^2sp^3, C^1sp^2, C^0sp	0.0 (0.1)	0.0	0.0
36b	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.0 (0.1)	0.0	0.0
36c	α- C	0.0 (0.0)	0.0	0.0
37	:Ar-CH=X	0.0 (0.1)	0.0	0.0
37a	C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.0 (0.1)	0.0	0.0
37b	α-C	0.0 (0.0)	0.0	0.0
38	:Al-C(=X)-Al	0.1 (0.5)	0.4	1.3
39	:Ar-C(=X)-R	0.1 (0.3)	0.2	1.7
40	:R-C(=X)-X, R-C≡X, X=C=X	0.8 (0.9)	2.0	1.0
41	:X-C(=X)-X	0.1 (0.4)	0.4	2.7
42	:X- -CH···X	0.1 (0.2)	0.1	0.0
43	:X- -CR···X	0.0 (0.2)	0.0	0.2
44	:X- -CX···X	0.0 (0.1)	0.0	0.0
H attached to				
46	:C^0sp^3 having no X attached to next C	3.5 (5.9)	8.5	1.4
47	:C^1sp^3,C^0sp^2	9.0(5.9)	21.5	0.8
48	:C^2sp^3, C^1sp^2, C^0sp	0.3(0.7)	0.7	0.9
49	:C^3sp^3, $C^{2-3}sp^2$, $C^{1-3}sp$	0.3(0.6)	0.6	0.7
50	:heteroatom	2.5(1.9)	6.0	2.0

Table 4: contd…

51	:α-C^d	1.3(1.7)	3.1	0.9
52	:C^0sp^3, having 1 X attached to next C	2.5(3.1)	5.9	0.9
53	:C^0sp^3, having 2 X attached to next C	0.3(1.0)	0.8	0.7
54	:C^0sp^3, having 3 X attached to next C	0.0(0.2)	0.1	0.5
55	:C^0sp^3, having 4 or more X attached to next C	0.0 (0.0)	0.0	0.0
O in				
56	:Aliphatic -OH	0.6 (1.1)	1.4	4.4
57	:phenol, enol, carboxyl OH	0.2(0.5)	0.4	1.3
58	: =O	1.1 (1.1)	2.6	1.1
59	:Al-O-Al	0.2 (0.5)	0.5	1.2
60	:Al-O-Ar, Ar_2O, R···O···R, R-O-C=X	0.5 (0.8)	1.2	0.9
61	:- -O	0.1 (0.4)	0.1	0.1
62	:O^-	0.5 (1.0)	1.3	0.0
63	:R-O-O-R	0.0 (0.1)	0.0	0.0
Se in				
64	:Any-Se-Any	0.0 (0.0)	0.0	0.0
65	:=Se	0.0 (0.0)	0.0	0.0
N in				
66	:Al-NH_2	0.1 (0.2)	0.1	0.0
67	:Al_2NH	0.1 (0.2)	0.1	0.3
68	:Al_3N	0.0 (0.2)	0.1	1.4
69	:Ar-NH_2, X-NH_2	0.1 (0.3)	0.2	0.0
70	:Ar-NH-Al	0.1 (0.2)	0.1	0.7
71	:Ar-NAl_2	0.1 (0.3)	0.1	0.4
72	:RCO-N<, >N-X=X	0.6 (0.8)	1.4	0.8
72a	H attached to heteroatom	0.4 (0.6)	0.9	1.1
72b	Without H	0.2 (0.5)	0.5	0.6
73	Ar_2NH, Ar_3N, Ar_2N-Al, R···N···R	0.2 (0.4)	0.4	1.0
73a	H attached to heteroatom	0.1 (0.3)	0.2	1.7
73b	Without H	0.1 (0.3)	0.2	0.8
74	R ≡ N, R= N-	0.1 (0.4)	0.3	0.3
74a	H attached to heteroatom	0.0 (0.1)	0	0.0
74b	Without H	0.1 (0.4)	0.3	0.3
75	:R- -N- -R,R- -N- -X	0.4 (0.8)	0.9	0.6
76	:Ar-NO_2. R- -N(- -R)- -O, RO-NO	0.0 (0.2)	0.1	0.3

Table 4: contd…

77	:Al-NO_2	0.0 (0.0)	0.0	0.0
78	:Ar-N=X, X-N=X	0.0 (0.1)	0.0	0.0
79	:N^+	0.5 (0.6)	1.1	0.4
79a	H attached to heteroatom	0.4 (0.6)	1.0	0.3
79b	Without H	0.0 (0.2)	0.1	0.0
F attached to				
81	:C^1sp^3	0.0 (0.1)	0.0	0.0
82	:C^2sp^3	0.0 (0.1)	0.0	0.0
83	:C^3sp^3	0.1 (0.5)	0.2	1.0
84	:C^1sp^2	0.1 (0.3)	0.2	0.6
85	:$C^{2-4}sp^2$, C^1sp, C^4sp^3, X	0.0 (0.1)	0.0	0.0
Cl attached to				
86	:C^1sp^3	0.0 (0.2)	0.0	0.0
87	:C^2sp^3	0.0 (0.1)	0.0	0.0
88	:C^3sp^3	0.0 (0.1)	0.0	0.0
89	:C^1sp^2	0.1 (0.5)	0.3	1.4
90	:$C^{2-4}sp^2$, C^1sp, C^4sp^3, X	0.0 (0.1)	0.0	0.0
Br attached to				
91	:C^1sp^3	0.0 (0.1)	0.0	0.0
93	:C^3sp^3	0.0 (0.0)	0.0	0.0
94	:C^1sp^2	0.0 (0.2)	0.1	0.3
95	:$C^{2-4}sp^2$, C^1sp, C^4sp^3, X	0.0 (0.0)	0.0	0.0
I attached to				
96	:C^1sp^3	0.0 (0.0)	0.0	0.0
97	:C^2sp^3	0.0 (0.0)	0.0	0.0
98	:C^3sp^3	0.0 (0.0)	0.0	0.0
99	:C^1sp^2	0.0 (0.3)	0.1	0.0
100	:$C^{2-4}sp^2$, C^1sp, C^4sp^3, X	0.0 (0.0)	0.0	0.0
halide ions				
101	:fluoride ion	0.0 (0.0)	0.0	0.0
102	:chloride ion	0.0 (0.0)	0.0	0.0
103	:bromide ion	0.0 (0.0)	0.0	0.0
104	:iodide ion	0.0 (0.0)	0.0	0.0
S in				
106	:R-SH	0.0 (0.1)	0.0	0.0
106a	H attached to heteroatom	0.0 (0.1)	0.0	0.0

Table 4: contd…

106b	Without H	0.0 (0.0)	0.0	0.0
107	:R$_2$S, RS-SR	0.1 (0.4)	0.3	0.9
108	:R=S	0.0 (0.1)	0.0	0.0
109	:R-SO-R	0.0 (0.1)	0.0	0.0
110	:R-SO$_2$-R	0.1 (0.3)	0.2	1.7
Si in				
111	:Si	0.0 (0.0)	0.0	0.0
B in				
112	:>B$^-$	0.0 (0.0)	0.0	0.0
P in				
115	:ylids	0.0 (0.0)	0.0	0.0
116	:R$_3$-P=X	0.0 (0.0)	0.0	0.0
117	:X$_3$-P=X (phosphate)	0.0 (0.1)	0.1	0.0
118	:PX$_3$ (Phosphite)	0.0 (0.0)	0.0	0.0
119	:PR$_3$ (Phosphine)	0.0 (0.0)	0.0	0.0
120	:C-P(X)$_2$=X (phosphonate)	0.0 (0.1)	0.0	0.0

The original ALOGP atom types [23] are identified, for each corresponding subset of UALOGP types. ^{b}R represents any group linked through carbon; X represents any heteroatom (O, N, S, P, Se, and halogens); Al and Ar represent aliphatic and aromatic groups, respectively; "=" represents a double bond; "≡" represents a triple bond; "- -" represents a aromatic bonds as in benzene or delocalized bonds such as the N_O bond in a nitro group; ".˙ " represents aromatic single bonds as the C–Nbond in pyrrole. The C--N bond order in pyridine may be considered as 2 while we have one such bond and 1.5 when we have two such bonds.

Table **4** shows the atom type distribution for the combined drug database, using both the united atom and all atom representations. The parameters given for each atom type are (i) mean occurrence per molecule, (ii) percent occurrence in the database, (iii) RDP$_i$ (relative druglikeness parameter) for each atom type.

Atom type druglikeness (RDP$_i$) analysis provides insight on those types which are more likely to be preferred in a drug molecule over a non-drug. Among heavy atoms, atom type 56 (hydroxyl oxygen) has the highest RDP$_i$ value of 7 reflecting its crucial role as a donor / acceptor in drug molecules. United atom type 3a (C^0sp^3 carbon having no heteroatom attached to its next C) is found to occur frequently among carbon atoms in drugs, with an RDP_i value 3.3. This saturated carbon type plays the role of bridgehead for many types of rings such as cyclohexane, piperidine, pyrrolidine *etc.*, which are quite common among drugs. Among other carbon types with high druglikeness values, 1a, 4, 8 (and subtypes 8a and 8b), 11, 16 (and subtypes 16a and 16b), 17, and 29 have higher values of $RDP_{i.}$ These results underscore the importance of non-aromatic, unsaturated carbon types in drugs relative to non-drugs. Though carbon atom type 4 is relatively infrequent, it is modestly preferred in drugs (RDP_i = 2.8). A number of atom types are highly infrequent in drugs and less preferred. These are assigned an RDP_i value of ~0. These represent tri-substituted nitrogen types with two or three attached aliphatic groups.

Among hetero atoms, types 56, 57 (hydroxyl oxygens), 58 (carbonyl oxygen) and 59 (ether oxygen) have higher RDP_i values, and among nitrogen types, 73a (tri-substituted nitrogen) stands out. Among halogen types, 83 and 89 appear more druglike. Among sulfur types, 110 (SO_2 type) stands out.

Recent studies from the lab of Bickerton and co-workers, describe a measure of drug-likeness based on the concept of desirability, termed as QED, the Quantitative Estimate of Drug-likeness [19]. The QED uses common physico-chemical properties to compare drugs like Lipinski's rule of five and facilitates the ranking of compounds in an intuitive and transparent way. However, in addition to the descriptors used by Lipinski, QED also considers the number of aromatic rings, the number of rotatable bonds, the polar surface area, and toxic groups. The desirability function has been derived based on a non-redundant data set comprising of 771 approved drugs from ChEMBL DrugStore database. QED method also highlights limitations of Lipinski's criteria [1] in effectively evaluating new drug candidates. Drugs that fail Ro5 [1] have drug-likeness measures that overlap with drugs that satisfy QED criteria.

FUTURE DIRECTION

Measures of druglikeness combine a number of features common to drugs, though the selection of those features is critical for success. Delineating these and employing them for a more readily usable and interpretable measure of druglikeness is an important research endeavor. Future work in this area will include development of better lead likeness and druglikeness scores as well as new parameters (such as Atom Type Diversity) to focus on specific features of compounds which are a result of optimization process that always precedes drug discovery.

ACKNOWLEDGEMENTS

We thank Jubilant Biosys Ltd for supporting the present research. Rajesh Hariharan thanks SASTRA University, Thanjavur, INDIA, for useful discussions and providing an opportunity to pursue Ph D.

CONFLICT OF INTEREST

The authors confirm that this chapter contents have no conflict of interest.

ABBREVIATIONS

PSA	=	Polar Surface area
MW	=	Molecular Weight
NATS	=	Number of Atoms
QSAR	=	Quantitative Structure Activity Relationship

REFERENCES

[1] (a) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3-25. (b) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharm. Tox. Meth.* **2000**, *44*, 235-249.

[2] Leeson, P. D. and Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry *Nat. Rev. Drug Discov.*, **2007**, *6*, 881-890.

[3] (a) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. *J. Comb. Chem.* **1999**, *1*, 55-68. (b) Viswanadhan, V. N.; Balan, C.; Hulme, C.; Cheetham, J. C.; Sun, Y. Knowledge-based Approaches in the Design and Selection of Compound Libraries for Drug Discovery. *Curr. Opin. Drug Discov Develop.* **2002,** *5*, 400-406.

[4] Viswanadhan, V. N.; Ghose, A. K.; Kiselyov, A.; Wendoloski, J. J.; Weinstein, J. N. Knowledge-based approaches for the design of small-molecule libraries for drug discovery. In *Combinatorial Library Design and Evaluation. Software Tools, and Applications in Drug Discovery.*; Ghose, A. K.; Viswanadhan, V. N.; Ed.; Marcel-Dekker, New York, USA, **2001**, pp. 267-289.

[5] Veber D. F, Johnson S. R, Cheng H. Y, Smith B. R, Ward K. W, Kopple K. D: Molecular properties that influence the oral bioavailability or drugs. *J. Med. Chem.* **2002**, *45,* 2615-2623.

[6] Viswanadhan, V. N.; Rajesh, H; Balaji, V. N.; Atom Type Preferences, Structural Diversity, and Property Profiles of Known Drugs, Leads, and Nondrugs: A Comparative Assessment *ACS Comb. Sci*. **2011**, *13*, 327–336.

[7] Ghose, A. K.; Herbertz, T.; Salvino, J. M.; Mallamo, J. P. Knowledge-based Chemoinformatics Approaches to Drug Discovery. *Drug Discov. Today*, **2007**, *11* (23-24), 1107-1114.

[8] Hutter, M. C. Separating Drugs from Nondrugs: A Statistical Approach Using Atom Pair Distributions. *J. Chem. Inf. Model*. **2007**, *47*, 186-194.

[9] Xu, J.; Stevenson, J. Drug-like Index: A New Approach to Measure Drug-Like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177-1187.

[10] Sadowski, J.; Kubinyi, H. A. Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325-3329.

[11] Murcia-Soler, M.; Perez-Gimenez, F.; Garcia-March, F. J.; Salabert- Salvador, M. T.; Diaz-Villanueva, W.; Castro-Bleda, M. J. Drugs and Nondrugs: An Effective Discrimination with Topological Methods and Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1688-1702.

[12] Givehchi, A.; Schneider, G. Impact of Descriptor Vector Scaling on the Classification of Drugs and Nondrugs with Artificial Neural Networks. *J. Mol. Model.* **2004**, *10*, 204-211.

[13] Biswas, D.; Roy, S.; Sen, S. A Simple Approach for Indexing the Oral Druglikeness of a Compound: Discriminating Druglike Compounds from Nondruglike Ones. *J. Chem. Inf. Model.* **2006**, *46*, 1394-1401.

[14] Ajay; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between "Drug-like" and "Nondrug-like" Molecules? *J. Med. Chem***. 1998**, *41*, 3314-3324.

[15] Walters, W. P.; Murcko, A. A; Murcko, M. A. Recognizing molecules with drug-like properties. *Curr. Opin. Chem. Biol.* **1999**, *3* (4), 384-387.

[16] eDrugSCAN, an online virtual screening tool, allowing stepwise search for drug-like compounds available at http://service.bioinformatik.uni-saarland.de/edrugscan/ (accessed July 25, 2011)

[17] (a) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887-2893. (b) Bemis, G. W.; Murcko, M. A. Properties of Known Drugs. 2. Side Chains. *J. Med. Chem.* **1999**, *42*, 5095-5099.

[18] Wang, J.; Ramnarayan, K. Towards Designing Drug-Like Libraries: A Novel Computational Approach for Prediction of Drug Feasibility of Compounds. *J. Comb. Chem.* **1999**, *1*, 524-533.

[19] G. Richard Bickerton; Gaia V. Paolini; Jeremy Besnard; Sorel Muresan; Andrew L. Hopkins, Quantifying the chemical beauty of drugs. Nature, **2012**, *4*, 90-98.

[20] Rishton, G. M. Reactive Compounds and *in vitro* False Positives in HTS. *Drug Discov. Today,* **1997**, *2*, 382-384.

[21] Kutchukian, P. S., David Lou, Shakhnovich, E. I. FOG: Fragment Optimized Growth Algorithm for the *de Novo* Generation of Molecules Occupying Druglike Chemical Space *J. Chem. Inf. Model.* **2009,** *49* (7), 1630-1642.

[22] Nadine Schneider,, Christine Jäckels,, Claudia Andres, and, Michael C. Hutter Gradual *in Silico* Filtering for Druglike Substances *J. Chem. Inf. Model.* **2008,** *48,* 613-628.

[23] Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 856-864.

[24] Wenlock, M. C.; Austin, R. P.; Barton, P.; Davis, A. M.; Leeson, P. D. A Comparison of Physiochemical Property Profiles of Development and Marketed Oral Drugs. *J. Med. Chem.* **2003**, *46*, 1250-1256.

[25] Ritchie, T. J.; Luscombe C. N.; Macdonald, S. J. F. Analysis of the Calculated Physicochemical Properties of Respiratory Drugs: Can We Design for Inhaled Drugs Yet? *J. Chem. Inf. Model.* **2009**, *49* (4), 1025–1032.

[26] Lajiness, M. S.; Vieth, M.; Erickson, J. Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discov. Devel*. **2004**, *7*, 470–477.

[27] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.,* **2002**, 42 (6), 1273–1280.

[28] Egan, W. J.; Merz, Jr. K. M.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics, *J. Med. Chem.* **2000,** *43,* 3867-3877.

[29] Palm K, Stenberg P, Luthman K, Artursson P: Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharm. Res.* **1997,** *14,* 568-571.

[30] Kelder J, Grootenhuis PDJ, Bayada DM, Delbressine LPC, Ploemen J-P: Polar molecular surface as a dominatingdeterminant for oral absorption and brain penetration of drugs. *Pharm. Res.* **1999**, *16*, 1514-1519.

[31] Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. *J. Phys. Chem. A.* **1998**, *102*, 3762-3772.

[32] Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163-172.

[33] Crippen, G. M.; Wildman, S. A.; Prediction of Physicochemical Parameters by Atomic Contributions *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868-873.

[34] Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177-182.

[35] Wishart, D. S.; Knox, C.; Guo, A. C.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for *in silico* drug discovery and exploration. *Nucleic Acids Res.* **2006**, *1*;*34* (Database issue):D668-72.

[36] ChemBridge Corporation. www.chembridge.com (accessed July 25, 2011)

[37] Cochran, A.J.; Lawson, D.H.; Linton, A.L.; Renal papillary necrosis following phenacetin excess. *Scott. Med. J.* **1967**, *12*, 246-250.

[38] Baillie, JK; Power, I The mechanism of action of gabapentin in neuropathic pain. *Curr Opin Investig Drugs*, **2006**, *7,* 33–9.

CHAPTER 13

Role of *In Silico* Stereoelectronic Properties and Pharmacophores in Aid of Discovery of Novel Antimalarials, Antileishmanials, and Insect Repellents

Apurba K. Bhattacharjee[*]

Department of Medicinal Chemistry, Division of Experimental Therapeutics, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, MD 20910-7500 (USA)

Abstract: Diseases caused by parasites have an overwhelming impact on public health throughout the world, particularly in the tropics and subtropics. Malaria and leishmaniasis are two such widely known neglected parasitic diseases. The current global situation indicates more than one million deaths from these two diseases every year despite several efforts by WHO to combat them. Vectors for carrying and transmitting these parasites are arthropods. Use of insect repellents is a vital countermeasure in reducing these arthropod-related diseases. However, despite access to many available drugs for treatment of these diseases, their growing resistance poses serious concerns and necessitates development of novel countermeasures. The present chapter discusses how the *in silico* methodologies can be utilized to develop pharmacophore models to identify novel antimalarials, antileishmanial, and insect repellents. The models presented in this chapter not only provided important molecular insights to better understand the "interaction pharmacophores" but also guided generation of templates for virtual screening of compound databases to identify novel bioactive agents. The pharmacophore models presented here demonstrated a new computational approach for organizing molecular characteristics that were both statistically and mechanistically significant for potent activity and useful for identification of novel analogues as well.

KEYWORDS: *In Silico* pharmacophore models, CATALYST methodology, parasites, malaria, leishmaniasis, arthropods, insect repellents, virtual screening, compound database, quantum chemical (QM) calculations, stereo-electronic properties, molecular electrostatic potentials (MEPs), drug design, drug discovery, novel compounds.

[*]**Corresponding author Apurba K. Bhattacharjee:** Department of Regulated Laboratories, Division of Regulated Laboratories, Walter Reed Army Institute of Research, 503 Robert Grant Avenue, Silver Spring, MD 20910, USA; Tel: 301-319-9043; Fax: 301-319-9449; E-mail: apurba1995@yahoo.com

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50013-8

INTRODUCTION

Diseases caused by protozoal parasites have an overwhelming impact on public health throughout the world, particularly in the tropics and subtropics. Malaria continues to be highlighted as the most severe of human parasitic diseases by the WHO, responsible for about one million deaths every year [1]. Malaria infection in humans is caused primarily by four parasitic species: *P. falciparum*, *P.vivax*, *P. ovale*, and *P. malariae*. Although *P. vivax* and *P. falciparum* are the two most widely distributed species, *P. falciparum* alone is responsible for over 95% of deaths worldwide. Development of curative antimalarial agents is difficult due to various developmental stages of the parasite within the host. Over the last few decades, the two mainstays of anti-malarial chemotherapy, CQ and pyrimethamine/sulfadoxine, have been significantly compromised in many regions of the world due to spread of drug-resistant parasites. Thus, efforts to control the disease met with decreasing success. To overcome the problem, a range of newer drugs and combinations have gradually been introduced that include, mefloquine (1984), artemisinins (1994), artemether/lumefantrine (1999), atovaquone/proguanil (1999), chlorproguanil/dapsone (2003), and more recently the general ACT but all of them have some issues for limiting their use [2]. Search for novel antimalarial drugs and drugs that can reverse resistance of the currently available drugs, particularly chloroquine continue to remain important goals for antimalarial discovery. Although artemisinin analogues such as artesunate and arteether were quite effective, particularly against the drug-resistant *P. falciparum,* observations of drug-induced and dose-related neurotoxicity in animals have raised concern about safety of these compounds for human use [3-5]. Thus, more extensive efforts for discovery of new less toxic and more affordable antimalarial drugs are clearly necessary. Although the genome sequence of *P. falciparum* was completed in 2002, eliminating many barriers for performing several state-of-the-art molecular and biological researches in malaria, new therapies have not yet resulted from the genome-dependent experiments, though a wealth of new information have been produced about the basic biology of the parasites [2]. These genome-dependent experiments are likely to aid discovery of new antimalarial therapeutics.

The leishmaniasis is another neglected tropical disease that shows a grim picture, identified by WHO as an increasing major health problem in the world [6]. Leishmaniasis represents a spectrum of diseases resulting from different species belonging to the genus *Leishmania*, a protozoal parasite transmitted by the bite of the phlebotomine sand fly. Clinical manifestations of the infection range from cutaneous and mucocutaneous to visceral leishmaniasis. An estimated 12 million people are currently afflicted worldwide with leishmaniasis and 1.5 to 2 million new cases added each year [6]. The visceral manifestation of the disease is often fatal if untreated which alone recently claimed an estimated 100,000 lives in Sudan outbreaks [7]. Non-availability of satisfactory chemotherapeutic agents and failure to develop an effective vaccine are considered to be two stumbling blocks in the combat of this disease [8]. The current chemotherapy of leishmaniasis relies heavily upon the use of pentavalent antimony compounds that require parenteral administration of high doses and a lengthy course of treatment resulting in marked increase of serious side effects and decreasing efficacy. Two pentavalent antimonial drugs, sodium stibogluconate (Pentostam) and meglumine antimonate (Glucantime) are the current choice of treatments for leishmaniases and had been the choice for past 50 years. Such heavy metal pharmacology is found to have severe side effects including nausea, diarrhea, convulsions, and even cardiotoxicity [9]. The treatments are not only not ideal due to these adverse side effects but also responsible for rapid development of clinical resistance within a few weeks and co-infections of leishmaniasis-AIDS together with high costs for long term treatments [10-13]. More importantly, prospects for antileishmanial vaccines remain unclear in the near future [11, 14-16]. Thus, there is clearly a need for discovery and development of less toxic drugs that are effective against all forms of leishmaniases.

Insects, broadly known as arthropods, are the vectors for both malaria and leishmaniasis and also responsible for many other lethal human diseases including African trypanosomiasis, dengue fever, filariasis, and viral encephalitis [17]. In terms of disease transmission, mosquitoes and sand flies are among the world’s most notorious insect vectors [17]. Since mosquitoes feed on blood, this insect species cause more human suffering than any other organism. Malaria results from infection carried by mosquitoes. Mosquitoes also can transmit the arboviruses responsible for yellow fever, dengue hemorrhagic fever, epidemic polyarthritis, and several forms of encephalitis. Although it took many years to search for the vector of leishmanaisis, it was finally established in 1921 that the transmission of this disease to humans occurs through sand flies belonging to the genus *Phlebotomus* [18].

Since international travel has grown enormously in recent years, the scope for transmission of these two diseases has also increased all over the world [1, 6]. To counter the transmission, quest to repel insects, particularly mosquitoes and sand flies, continued including research on mosquito behavior and control but still safe and effective insect repellents have not yet been found. Although DEET is the leading commercially available repellent for over fifty years now [19a], it has several disadvantages that include short duration in hot sultry climates, and strong plasticizer properties (softens or mars many plastic items or painted surfaces). Moreover, DEET is only effective against mosquitoes and has limited activity against flies [19b, 19c]. However, developing an ideal repellent agent that should repel multiple species, remain effective for 8-10 hours, does not cause irritation to the skin or mucous membranes, have no systemic toxicity, should be resistant to abrasion and rub-off, and be greaseless and odorless is still a distant dream. At present, no available insect repellent meets all of these criteria. Efforts to find a compound with such attributes face numerous challenges and variables that affect the inherent repellency of a potential chemical. Repellents do not all share a single mode of action, and surprisingly little is known about how repellents act on the target proteins of the insects. Furthermore, different species of mosquitoes react differently to the same repellent. Thus, understanding the physico-chemical requirements for repellent properties, how a repellent interacts with the target proteins is important for successful discovery new arthropod repellents. Because the biochemical steps leading to a desired repellent effect, especially interactions with the three-dimensional molecular structure of the receptor(s) are unclear, various efforts have been made to develop a general structural framework with high probability for repellent activity to guide the synthesis work [20]. The ability of the insect repellents to interact with the recognition sites in receptors results from a combination of steric and electronic properties. Study of stereoelectronic properties of insect repellents can provide valuable information not only to better understand the mechanism of repellent action but also provide the structural requirements for repellent activity to generate a reliable pharmacophore model to design of more effective repellents. In addition, three-dimensional (3D) pharmacophore model generations can be useful for identification of potential repellents through utilization of 3D database queries for search of compound databases.

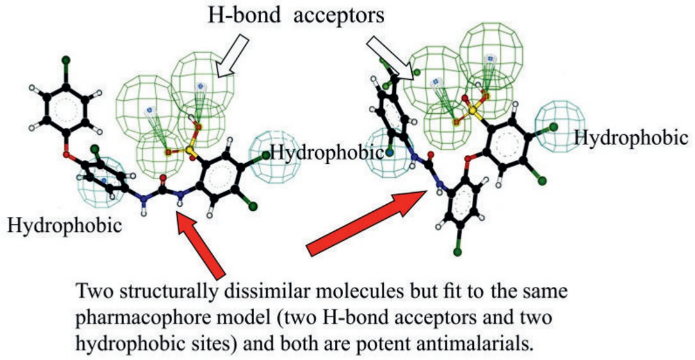
Discovery and development of new therapeutics are expensive and complex processes with ever changing technologies. On an average, it takes about 10 years and approximately five to six million U.S. dollars to bring a new effective chemical entity from the bench of discovery to the market [21, 22]. Therefore, any technology that can improve the efficiency of the process is considered highly valuable to the pharmaceutical industry.

With the advent of modern computers with high speed, astronomical memory and graphic tools, accomplishing computations and visualization of structures ranging from small to large bio-molecules including proteins have become more efficient with greater precision. The graphic tools in modern computers have not only made possible visualization of three-dimensional structures of large protein molecules, but allowed interactive virtual docking experiments between potential drug molecules and the binding sites of proteins. The current advances in these methodologies have direct applications ranging from accurate *ab initio* quantum chemical calculations of stereo-electronic properties, generation of three-dimensional pharmacophores, and performance of database searches to identify bioactive agents [22].

Increasing costs for pharmaceutical development have resulted in the emergence of *in silico* screening or virtual screening of databases to identify potential new compounds in recent years [22-24]. Virtual screening is a process of intelligent use of computing to analyze large databases of chemical compounds to identify potential drug candidates. The process can serve as a complimentary tool to HTS for rapid and effective experimental assay of large pool of compounds. Screening compounds by this method is essentially a knowledge-based approach and thus implicitly requires certain information about the nature of the receptor binding site or the nature of ligand that is expected to bind effectively at the active site. However, the type of procedure followed in virtual screening for compound databases depends upon availability of information as input and requirement for the output. Thus, if three-dimensional structure of the target enzyme or the protein is available, small molecule docking procedures can be adopted to perform structure-based virtual screening for identification of an ideal ligand. If the three-dimensional structure of the target protein is unknown, feature based pharmacophore models can be constructed from activity data of known compounds and the developed model template can be used for virtual screening to identify potential new hits. Pharmacophores may also be developed from other molecular properties, such as the ADME properties, toxicity data, lipophilicity, and drug-related properties. Identification of new active compounds using *in silico* pharmacophores has shown remarkable success in recent years [25-29].

In this chapter, recent *in silico* stereo-electronic and pharmacophore modeling studies of antimalarials, antileishmanials, and insect repellents are reviewed along with our efforts to identify novel active compounds. The goal is to provide a perspective for how information on three dimensional electronic profiles can further facilitate identification, design and synthesis of new lead antimalarial, antileishmanial and insect repellent compounds.

Concept of pharmacophores is one of the most important steps for understanding the interaction between a receptor and its ligand. *In silico* "three dimensional pharmacophore is as an ensemble of steric and electronic features those are necessary for optimal interaction with a specific receptor to trigger or inhibit its biological response" [25]. Literature survey reveals that Paul Ehrlich probably first offered the definition of a pharmacophore in early 1900s as "a molecular framework that carries (*phoros*) the essential features responsible for a drug's (*pharmacon*) biological activity" [30]. This definition remained almost the same until Peter Gund provided a remarkably similar definition as "a set of structural features in a molecule that is recognized at a receptor site and is responsible for that molecule's biological activity" in 1977 [27]. Peter Gund is one of the pioneers in pattern searching based on functional features (pharmacophores) for compound databases to identify new compounds that may share the same functional features and developed the first 3D searching software, Molpad [27]. A more modern definition of a three dimensional pharmacophore is a geometric distribution of chemical features, such as hydrogen bond acceptor, hydrogen bond donor, aliphatic and aromatic hydrophobic functions, and ring aromatic hydrophobicity in the three dimensional space surrounding a molecule which can define its specific biological activity [26, 28, 29, 31, 32]. For example, two antimalarial compounds having different chemical structures may share the same pharmacophore (Fig. (**1**)).



Two structurally dissimilar molecules but fit to the same pharmacophore model (two H-bond acceptors and two hydrophobic sites) and both are potent antimalarials.

Figure 1: Example of 3D pharmacophore model: Defining feature requirements in a molecular structure for antimalarial activity.

RESULTS AND DISCUSSIONS

Discussions on stereo-electronic profiles and pharmacophore models for antimalarials, antileishmanials, and insect repellents presented here are based on the following two considerations:

(a) Quantum chemically calculated stereo-electronic properties and quantitative structure-activity relationships for mechanistic insights and guidance for generation of feature-based three-dimensional molecular "interaction pharmacophores".

(b) Chemical features and activity relationships for generation of three-dimensional pharmacophore models as tools for virtual screening of compound databases in order to identify potential new compounds**.**

The implicit assumptions [25, 26] for both the above two considerations are:

a) The structures used in the model are responsible for the biological activity, not its metabolite.

b) The conformation of the model is the bioactive conformation.

c) The binding site is the same for all molecules in the proposed model.

d) Biological activity is accounted only in terms of thermodynamic equilibrium, particularly by enthalpic energy considerations, assuming entropy for the molecules to be similar.

e) Kinetics of the processes are ignored.

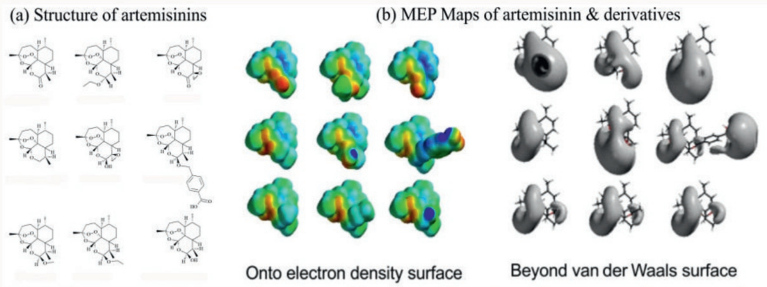
f) Transport properties, diffusion and solvent effects are largely avoided.

Stereo-Electronic Considerations

Since the ability of a bioactive molecule to interact with recognition sites of receptors results from a combination of steric and electronic properties [25], study of their stereo-electronic properties can provide valuable information not only to better understand the mechanism of action but also enable viewing intrinsic “interaction pharmacophore” profiles for aiding design and synthesis of more potent analogues. Quantum chemical methods can provide accurate estimate of stereo-electronic properties as well as assessment of interactions between bioactive molecules and receptors [31-33].

Antimalarial Compounds

Despite non-availability of affordable, safe and effective therapeutics for treatment of severe malaria, importance of the disease and efforts to eradicate it have led to a huge inventory of antimalarial compounds [34] though largely ineffective. Most of the earlier quantum chemical studies, particularly in the 1990s on antimalarials were focused on artemisinin and artemisinin-like compounds for understanding mainly the mechanism of action in order to find an effective alternative to the drug resistant CQ [35-38] and to understand its mechanism of action at the molecular level to address the concern about its observed neurotoxicity in animals [39]. However, the good news is that no human neurotoxicity was observed so far, it was found to be confined only in dogs, rats, and monkeys with dose dependent intramuscularly injected derivatives of artemisinins, artemether and arteether [39]. In continuation of efforts for understanding the molecular electronic nature of the artemisinins, Fig. (**2a**), and the role of electronic properties toward neurotoxicity, we evaluated the stereo-electronic discriminators that differentiate between analogues with higher and lower experimental neurotoxicities [40] by performing quantum chemical calculations on artemisinin and eight of its derivatives [5]. We observed the least neurotoxic compounds to be more polar with an electric field pointing away from the endoperoxide bond, have a higher positive potential on the electron density surface (van der Waals surface) of all the carbon-containing ring C, a more stable peroxide bond to cleavage, a less negative electrostatic potential by the endoperoxide, and a single negative potential region extending beyond the electron density surface of the molecule, Fig. (**2b**). In general, the observed stereo-electronic attributes, Fig. (**2b**), related to the peroxide bond such as dipole moment and electric field, lower energy requirement for breaking the peroxide bond, more intrinsic nucleophilicity of the peroxide bond, and less electrophillic ring "C" showed links toward neurotoxicity of artemisinins [5].


(a) Structure of artemisinins
(b) MEP Maps of artemisinin & derivatives
Onto electron density surface
Beyond van der Waals surface

compd	relative neurotoxicity	dipole moment (D)	maximum positive MEP over ring C (kcal/mol)	maximum negative MEP over peroxide bond (kcal/mol)	energy required for scission of peroxide bond (kcal/mol)	aqueous solvent polarization energy (kcal/mol)	antimalarial activity IC_{50} (nM)
1	1.0	6.50	37.9	-34.0	685.7	2.7	10.5
2	1.0	4.07	32.5	-40.0	676.5	0.9	4.2
3	1.0	7.87	39.8	-38.4	685.2	4.9	NA
4	1.2	3.49	28.7	-44.5	674.8	0.56	5.7
5	1.5	6.04	31.4	-41.0	677.9	2.0	9.5
6	1.5	4.98	27.4	-40.0	672.6	0.94	15.1
7	2.1	2.31	25.8	-44.5	671.6	0.06	4.5
8	8.9	2.38	25.3	-45.7	672.2	0.12	4.1
9	213.0	2.11	21.5	-44.4	671.6	0	1.8

Figure 2: (a) Chemical structure of artemisinin and its eight derivatives. (b) Molecular electrostatic potential maps of artemisinin and the eight derivatives. (c) Table showing experimental neurotoxicity and several calculated stereo-electronic properties of artemisinin and its eight derivatives.

In 2000, Girones *et al*. [41] calculated kinetic energy based molecular quantum similarity measures to correlate the antimalarial activity of various artemisinin derivatives. Interaction of artemisinin with hydroxypropyl-β-cyclodextrin (HPBCD) was investigated by Illapakurthy *et al*. [42] who reported significant increase in phase solubility of these compounds in HPBCD. In our laboratory, we demonstrated [43] by a combined NMR and molecular modeling study that both artelinic acid and artesunic acid form complexes with natural cyclodextrin and thus, a possible alternative formulation scheme for these compounds with increased aqueous solubility while retaining its antimalarial activity. Quantum chemical calculations and automated docking simulations by Tonmunphean *et al*. [44] indicated significant effects of stereoisomer on the binding mode and activity of these compounds. In another study to reflect on the mechanism of action of stereoisomerism of these compounds, we reported an analysis of β-artelinic acid using a combination of NMR, *ab inito* quantum chemical (HF-6-31G**), and cyclic voltammetry methods and compared with two other artemisinin analogs, α-artelinic acid and β-arteether [45]. Our results indicated the importance of non-bonded interactions between specific protons and the ether oxygen atom in the neighborhood of the anomeric carbon atom in the two isomers to be responsible for different efficacies. In addition, we also explored the stereo-electronic and pharmacophore properties of several peroxide containing antimalarial trioxanes [46] and tetraoxanes [47]. The "interaction- pharmacophores" observed in the two above studies indicated the crucial presence of at least one hydrogen bond acceptor region in the trioxane or tetraoxane moiety for potent activity. Docking calculations with heme were found to be consistent with the above observation as the proximity of the heme iron to the oxygen atom of the trioxane or the tetraoxane moiety favored potent activity. Electron transfer from the oxygen of trioxane or the tetraoxane moiety was documented to be crucial for activity [46, 47]. The computed stereo-electronic properties of peroxy ketals showed a negative electrostatic potential region beyond van der Waals surface away from the peroxide moiety suggesting the compounds to be less toxic and should be safer [46]. Although these features suggested less likelihood for neurotoxicity of the peroxy ketals, the compounds showed poor antimalarial efficacy compared to the artemisinins, indicating a possible tradeoff between neurotoxicity and antimalarial efficacy in peroxide containing compounds [46].

With the completion of *Plasmodium falciparum* genome project and emergence of structure-based drug design methodologies, drug development efforts have largely shifted to targeting specific proteins in the parasite that are unique yet critical for its growth and survival [48]. Since *P. falciparum* is of prokaryotic origin, its apicoplast contains metabolic pathways that differ significantly from those found in the human host and thus, targeting the fatty acid biosynthesis pathways of malaria parasite has been a focus for antimalarial therapeutics research in recent years. Eukaryotes rely on type I FAS to survive whereas, the prokaryotes do not contain type I FAS but rely on a type II FAS for the *de novo* production of fatty acids [48]. The initiating steps of *P. falciparum* type II FAS depend upon the acyl carrier protein (*Pf*ACP) and two other enzymes, malonyl coenzyme A: ACP trans-acylase (*Pf*MCAT) catalyzing the formation of malonyl-ACP from malonyl-coenzyme A (malonyl-CoA) and ß-ketoacyl-ACP synthase III (*Pf*KASIII) catalyzing the condensation of malonyl-ACP and acetyl-CoA, forming a ß-ketoacyl-ACP product. This reaction is identical to that catalyzed by the bacterial FabH, an orthologue of the malarial *Pf*KASIII, being pursued for antimicrobial targets in recent years [49]. TLM and its analogues were the first compounds studied as inhibitors of the type II FAS of *Mycobacterium tuberculosis, Staphlococcus aure*us, and *Pastuerella multocida* [50]. Since TLM was documented to selectively target type II FAS both *in vitro* and *in vivo* and the analogues have little or no toxicity [51], these compounds were used as starters for evaluation against malaria [48]. We developed the first "interaction-pharmacophore" model for inhibition of KASIII from TLM, Fig. (**3a**), by performing sequentially the semi-empirical (AM1) and *ab initio* HF self-consistent field (Hartree - Fock SCF) quantum chemical calculations. Initially, a conformational search analysis on TLM was performed using the systematic conformational search techniques in SPARTAN [52] at the AM1 [53] single point level to obtain the population of low energy conformers. Next, we performed Monte Carlo "simulated annealing" approach as implemented in SPARTAN to generate trial conformations by way of random bond and ring torsions. To begin with the simulation, the molecule was considered to be in a high temperature system *i.e.*, it had sufficient energy to move from low to high energy conformations. This is important because often the global minimum conformation remains hidden by many local minima. As more conformations are explored, the temperature is decreased, making the molecule less able to move out of low energy conformations. Thus, when the search is completed, the molecule is most likely to be in the lowest energy conformation found up to that point. The global minimum-energy of the conformers identified by both the above methods are compared and assessed. Following this procedure, the lowest and the most abundant (highest population density) energy conformer was selected for complete geometry optimization by using HF/6-31G** basis sets comprising both d & p orbital polarized functions as the optimal choice in Gaussian98 [54] running on a SGI Octane workstation. The electronic properties such as molecular electrostatic potentials and orbital energies were calculated on the optimized geometry of the molecule.

Molecular electrostatic potentials (MEPs) were sampled over the entire accessible surface of the molecule (surface of a constant 0.002 e/au3 electron density corresponding roughly to a van der Waals contact surface), providing a measure of charge distribution from the point of view of an approaching reagent. The regions of positive electrostatic potential indicate excess positive charge, *i.e.*, repulsion for the positively charged test probe, while regions of negative potential indicate areas of excess negative charge, *i.e.*, attraction of the positively charged test probe. These iso surface values provide an indication of overall molecular size and of location of negative or positive electrostatic potentials. For example, in the present study, the MEPs encoded onto a surface of constant electron density (0.002 e/au3) portrays both steric and MEP characteristics of the molecules. This encoding is done by the use of color, colors toward the blue representing one extreme value of a property (most electrophilic being deepest blue) and colors toward the red representing the other extreme (deepest red being the most nucleophilic). Isopotential surfaces extending outward from the van der Waals surface of each molecule at -20.0, -10.0, and -5.0 kcal/mol were also generated to indicate the electron density profiles beyond molecular surface.

The three-dimensional MEP map of the 6-31G** optimized TLM superimposed onto total electron density, Fig. (**3b**), reveals that the center for the most negative potential (red region) lies in the vicinity of the sulfur and the carbonyl oxygen atoms of the molecule, whereas the center for most positive potential (blue region) lies by the adjacent methyl groups of the sulfur atom and to a lesser extent by the ethylene hydrogen atoms. However, on examination of the three-dimensional MEP profiles of TLM beyond the edge of van der Waals surface at -20.0, -10.0, and -5.0 kcal/mol, Fig. (**3b**), roughly corresponding to 1.45 A (-5.0 kcal/mol) to 1.35 A (-20.0 kcal/mol) away from the edge of the molecular surface, indicates a progressively large negative potential region extending laterally from the carbonyl oxygen to the sulfur atom and a small localized negative potential region by the hydroxyl oxygen atom. The large extended negative potential region from the carbonyl oxygen to the sulfur atom in TLM at -5.0 kcal/mol, Fig. (**3b**), may be regarded as a nucleophilic suction-pump acting as a magnet for the electrophilic part of the receptor. Since this potential is approximately 1.45 A away from the molecular surface, this feature is likely to be recognized first by the receptor to promote long range interactions between them. Electrostatic potential characteristics are considered to be key features of molecules through which it recognizes its receptor at longer distances to promote interaction between them [55]. The electrostatic interactions are considered to be the driving force toward formation of non-covalent Michaelis type of complexes with the receptor.

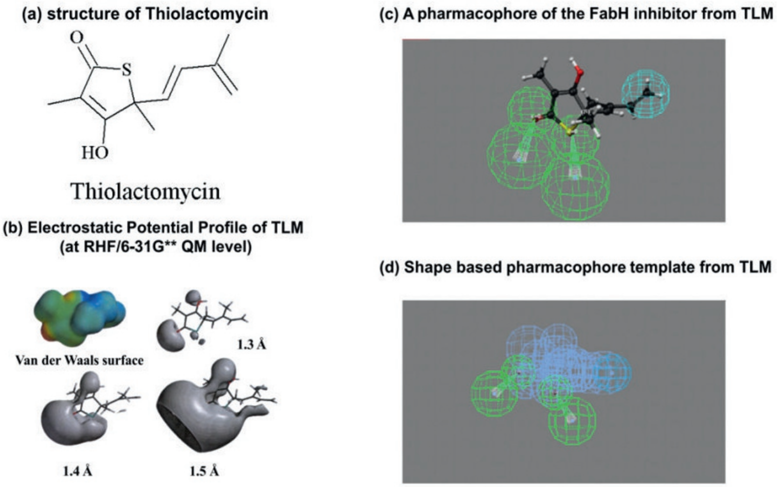

(a) structure of Thiolactomycin
O
S
HO
Thiolactomycin
(b) Electrostatic Potential Profile of TLM
(at RHF/6-31G** QM level)
Van der Waals surface
1.3 Å
1.4 Å
1.5 Å
(c) A pharmacophore of the FabH inhibitor from TLM
(d) Shape based pharmacophore template from TLM

Figure 3: Shows how the pharmacophore model was developed from the optimized (a) structure of thiolactomycin (TLM), (b) electrostatic potential profiles, (c) feature based pharmacophore of TLM, and (d) 3D shape-based template of TLM pharmacophore for virtual screening of database to identify new inhibitors and antimalarials.

The positive potential appears to be spread over a large region by the ethylene hydrogen atoms in TLM rendering this portion of the molecule to be hydrophobic. Widely distributed weak positive electrostatic field regions on the accessible molecular surface are believed to be an indication of hydrophobicity of a molecule [56]. Therefore, these electrostatic potential features provided the intrinsic reactivity profile or the 'interaction pharmacophore model' of TLM which later guided us to develop a feature-based pharmacophore model for virtual screening of databases to identify new *Pf*KASIII inhibitors as potential candidates for antimalarial therapeutics [48]. Fig. (**3d**) shows how we have developed the pharmacophore model from the optimized structure of TLM and eventually used it as a 3D shape template for virtual screening of database to identify new inhibitors and antimalarials.

Antileishmanial Compounds

Amongst earlier studies to account for the role of molecular electronic properties toward antileishmanial activity, Werbovetz *et al*. [57] reported that electrophilic compounds such as phenyl arsenoxide and 4-chloro-3,5-dinitro-α,α,α-trifluorotoluene (chloralin) can inhibit the assembly of leishmanial tubulin. In other studies, diospyrin, was shown to have significant inhibitory effect on the growth of leishmania donovani promastigotes due to catalytic activity of DNA topoisomerase I of the parasite [58]. Mukhopadhyay *et al*. [59] and Croft *et al*. [60] reported S-adenosylmethionine decarboxylase to demonstrate the inhibition of growth of leishmania donovani promastigotes (strain UR6) in a dose dependent manner. However, no stereo-electronic studies were reported in published literatures on these compounds. One of the earlier attempts to design new antileishmanials based on structure-activity relationships was by Bell *et al*. [61] who identified several aromatic diamidines and di-imidazolines having potent activity with reduced toxicity relative to pentamidine against leishmania mexicana amazonensis.

Thus, although efforts were made in the past to identify, design and synthesize new antileishmanial compounds, no reports were found in the published literature which devoted on stereo-electronic properties and "interaction-pharmacophore" profiles of known antileishmanials for understanding the molecular mechanism of activity to aid discovery of novel antileishmanial compounds. Toward this effort, we published a few studies focusing on the stereo-electronic properties of antileishmanial macrocyclic bisbenylisoquinoline derivatives [62] and camptothecins [63]. The first study was designed to assess the role of calculated stereo-electronic properties of five bisbenzylisoquinoline, gyrocarpine, daphnandrine, obaberine, pheanthine, and malekulatine toward antileishmanial activity using *ab initio* (3-21G*/4-31G*-HF) quantum chemical methods. The results on the antileishmanial macrocyclic bisbenylisoquinoline derivatives indicated [62] that the ability to form a cavity at the macrocyclic ring, preference for a specific orientation of lone-pair electrons of the ether-oxygen atoms, electrostatic potential profiles by the oxygen atoms, and similarity of the lowest unoccupied molecular orbital (LUMO) isosurface at the cavity were associated with potent antileishmanial activity [62]. Although no single stereo-electronic property could account for all the experimental activity data of three different strains of leishmania in the study, specific conformational preference for the ether-oxygen atoms in the cavity of macrocyclic rings and the resulting electronic property arising out of it probably played an important role in the complex mechanism of antileishmanial action in these alkaloids [62]. In the camptothecin study [63], we made attempts to rationalize the potent antileishmanial activity of methylenedioxy camptothecins which were known to act specifically against the pathogenic protozoan leishmania donovani *in vitro*, and also for generation of cleavable complexes in the presence of DNA and purified mamamalian topoisomerase I. We investigated the role of molecular electronic properties of camptothecin and four of its 10,11-methylenedioxy analogues [63] and observed a difference in the delocalization of positive potential between methylenedioxy camptothecins and camptothecin, and attributed the difference to increased affinity of the compounds for DNA in addition to both geometric and electronic differences of the E ring [63]. We concluded that one or both of these factors may contribute to the superior biological activity of the methylenedioxy camptothecin analogues.

Insect Repellents

Lack of literature information on molecular electronic structures of known insect repellents and efforts to better understand the mechanism of insect repellency properties prompted us to study the stereo-electronic properties of DEET to start with followed by thirty of its analogs [64]. In continuation of this study, we further investigated the molecular similarity and differences of stereo-electronic properties of DEET together with its analogues, natural insect juvenile hormone, and a synthetic insect juvenile hormone mimic, undecen-2-yl carbamate) [20], results of which later guided us to develop a pharmacophore model used for virtual screening of compound databases [65].

The stereo-electronic property study on known repellents involved quantum chemical calculations ranging from AM1 semi-empirical calculations to conformational search for the lowest and most abundant energy conformer of JH, JH-mimic, and fifteen DEET compounds [20]. We performed complete geometry optimization for each of the lowest and most abundant energy conformer using *ab initio* quantum chemical methods. Similarity analyses of stereo-electronic properties including the structural parameters, atomic charges, dipole moments, molecular electrostatic potentials, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) energies were performed on all the above molecules. Similarity of stereo-electronic profiles of the amide/ester moiety, negative electrostatic potential regions beyond the van der Waals surface, and a large distribution of hydrophobic regions in the compounds [20] were observed to be the three important similarity features that probably have similar interactions with the JH receptor. The similarity of electrostatic profiles beyond the van der Waals surface was attributed to the molecular recognition process with the JH receptor at a distance [20]. This feature similarity between the compounds was suggested by us to be a display of electrostatic bio-isosterism of the amide group in the DEET compounds, JH, and the JH-mimic and we hypothesized it as a model for molecular recognition at the JH receptor [20]. The insect repellent property of DEET and its analogues was proposed to be a probable conflict for complementary binding interactions with the JH- receptor binding sites [20].

In summary, the above computed stereo-electronic property studies, particularly the electrostatic potential profiles beyond van der Waals surfaces for the antimalarial, antileishmanial, and the insect repellents were found to be reasonably well correlated with the observed experimental activity of the agents. More importantly, the calculated stereo-electronic profiles provided the foundation to guide our later development of feature based pharmacophore models for identification of new compounds as described in the following section.

Chemical Feature Based Pharmacophores

Despite many efforts for discovery of improved therapeutics for malaria, leishmaniasis, and insect repellents, little success has been made to discover truly effective non-toxic compounds based on alternative structure theory that will not likely to develop rapid resistance. For past several decades, efforts have only led to the development of derivatives of preexisting chemical structures. Compounds from new chemical classes have barely been explored. In pursuit of these objectives, we adopted an *in silico* strategy to develop pharmacophore models from published literature data and use the generated models to identify potentially active compounds of novel chemical classes through virtual screening of compound databases. The advantage of the pharmacophore is that it transcends the structural class and captures features those are responsible for the intrinsic activity of potential therapeutics of new chemical classes or chemo-types from searches of compound databases [28, 29, 31]. Quantitative methods that attempt to identify arrangements of atom features in space in relation to an experimental biological activity are commonly referred to as 3D – QSAR (three dimensional quantitative structure-activity relationships) pharmacophore generation methods.

There are several approaches used for developing 3D-QSAR models of bioactive compounds that include 3D-QSAR molecular conformation based alignment rule, statistical techniques such as partial least squares (PLS) to identify relationships between structural descriptors and biological activity [66], 3D-QSAR-CoMFA [67], 3D-QSAR-VolSurf/Grid [68], and 3D-QSAR-CATALYST [69] methods. In this section, the focus of discussion will be the efforts for developing feature based pharmacophores utilizing the 3D-QSAR- CATALYST procedure [69] and utilization of the models for virtual screening of compound databases to identify potential antimalarial, antileishmanial, and insect repellent agents.

In recent years, virtual screening of databases using pharmacophores has been successfully applied to identify many novel ligands for a variety of proteins and enzymes [22]. The novel inhibitors discovered have very little similarity with other known inhibitors and a majority of the leads is found to have potent activities in low micromolar level [22, 23]. Successful pharmacophore based virtual screening of databases using the CATALYST methodology resulted in identification of many novel potent compounds against several targets such as, serine protease chymase [70], antigen α4β1 [71], EDG3 [72] mesangial cell proliferation [73], and rat 5α-reductase [74]. Even without the knowledge of 3D structure of the biological target, this methodology has been quite successful for example, identification of α1- adrenoreceptor antagonists, LTD4 receptor antagonists, corticotrophin-releasing hormone antagonists, Na+/bile acid co-transpoters [74], mesangial cell proliferation inhibitors [73], discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors [75] and new P450 inhibitors [76] to name a few.

However, similar efforts for identification of new antimalarials, antileishmanials and insect repellents were not found in literature [28, 29, 65].

Antimalarial Compounds

Since crystal structures of majority of target proteins or enzymes of antimalarial drugs were unknown, we focused on *in silico* approaches to develop pharmacophore models from known active antimalarial compounds. The strategy not only helped us to identify many new antimalarial agents but also provided insights for possible interactions with the unknown target receptors at the active site. One of our first successful virtual screening efforts was the identification of new antimalarial agents by developing a pharmacophore from a set of known CQ resistance reversal agents [77, 78].

First, we developed a pharmacophore model from the known CQ-reversal agents and then cross-validated it by mapping its features on a series of other CQ-reversal agents such as, chloropheniramine, cyproheptadine, ketotefin, pizotyline, azatadine, loratadine, verapamil, and penfluridol. Mapping of the pharmacophore onto six well known CQ-resistance reversal agents showed excellent consistency [77, 78]. Next, we performed a database search using the pharmacophore for potential new CQ-reversal agents from our in-house WRAIR - Chemical Information System [63] database of over 290,000 compounds. The search resulted in identification of several 2,4-diamino-3*,4*-dichloro-6-quinazolinesulfonanilide analogues as promising candidates for further studies. The lead identified compound was observed to be a potent antimalarial in the RP mouse malaria presumptive causal prophylactic test as well as in MM *in vivo* mouse malaria test [77].

In our next effort, we have applied the methodology to develop a pharmacophore for proton-pump inhibitors from four benzimidazoles, namely, omeprazole, lansoprazole, rabeprazole and pantoprazole which are clinically used as proton pump inhibitors [79]. The generated pharmacophore model was used for search of new compounds from our in-house database and identified 128 compounds that have similar features. Three of these compounds were observed to have efficacious antimalarial properties in mouse malaria *in vivo* [79].

In continuation of these efforts, we developed another pharmacophore model, Fig. (**4a**) from a series of indolo[2,1-b]quinazoline-6,12-diones (tryptanthrins) which exhibited remarkable *in vitro* antimalarial activity (below 100 ng/mL) and low cytotoxicity against sensitive and multidrug-resistant *Plasmodium falciparum* malaria [28]. However, although these compounds possessed outstanding *in vitro* activity and reasonably well tolerated toxicity for promising antimalarial candidates, the compounds did not display *in vivo* activity, probably due to poor bioavailability and aqueous solubility. Nonetheless, the pharmacophore model that we developed for this series of compounds was found to be very useful for identification of a variety of different classes of antimalarials and provided a fairly reliable foundation for 3D database searches, Fig. (**4a**). The pharmacophore was found to map well onto many well-known antimalarial drugs such as, quinine, hydroxychloroquine, Fig. (**4b**), and also rhodamine dyes, and chalcones [28]. Interestingly, the observed mapping of this pharmacophore model onto quinine, Fig. (**4b-A**) led us to believe that like quinine, the tryptanthrins may target heme polymerase from the *P. falciparum* tropozoites. Since the target protein for antimalarial activity of the tryptanthrins was unknown, we evaluated six substituted 4-azaindolo[2,1-*b*]quinazoline-6,12-dione analogues of the tryptanthrins for hemin binding affinity by 1H NMR methods, x-ray crystallography, and *ab initio* quantum chemical calculations [80] and found the evidence for heme-tryptanthrin stacking organization in all these analogues. The observation was also consistent for the proposed interactions with hemin determined separately by NMR experiments [80].

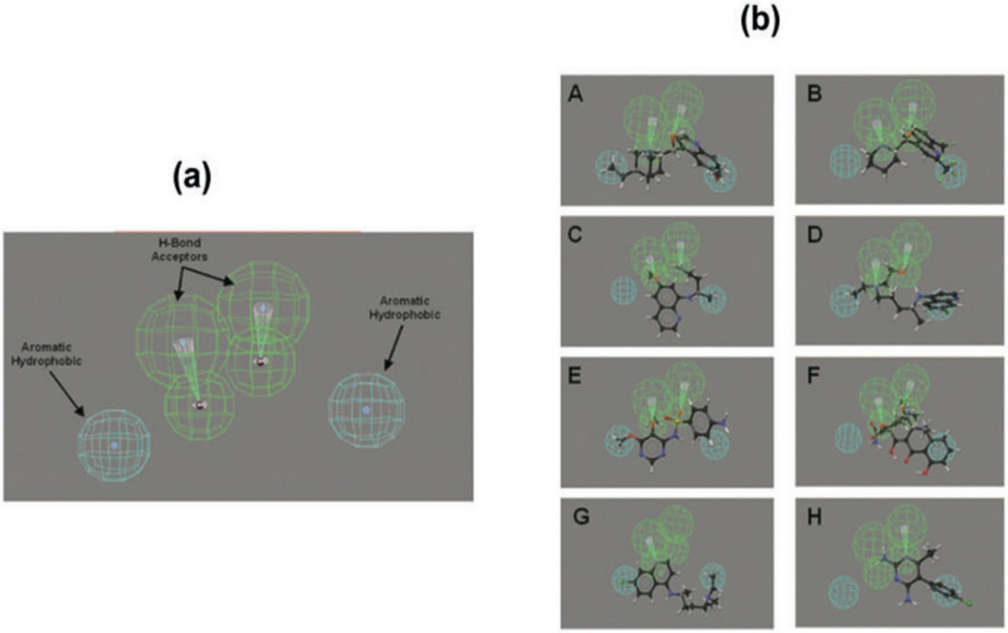


Figure 4: (a) Pharmacophore for antimalarial activity of the tryptanthrins. (b) Mapping of the pharmacophore onto eight commonly used antimalarial drugs in the United States: (A) quinine, (B) mefloquine, (C) primaquine, (D) hydroxychloroquine, (E) sulfadoxine, (F) doxycycline, (G) chloroquine, and (H) pyrimethamine.

Using the pharmacophore as a search template for virtual screening of the in-house database led to successful identification of five new aminoquinazoline derivatives as promising candidates for further study, as these compounds were found to be potent both *in vitro* and *in vivo* in mouse malaria screening tests [28]. Thus, the pharmacophore model that we developed from the tryptanthrins was not only useful for identification of novel class of antimalarials but also provided a possible mechanism of antimalarial action with the target, the heme protein.

Our first effort of application of structure-based drug design methodologies following the completion of the *P. falciparum* genome project was focused on the specific proteins in parasites that are unique yet critical for cellular growth and survival. With a direct role in the regulation of cellular proliferation, the cyclin-dependent proteins kinases (CDKs) were attractive drug targets for discovery of new antimalarial chemotherapies and therefore, we targeted the malarial CDK. Primarily, three plasmodial CDKs (PfPK5, PfPK6 and Pfmrk) were being investigated. There are several inhibitors for CDKs that were reported to possess antiparasitic activity when assayed with the malarial parasites *in vitro* [81]. We developed a new pharmacophore model from known inhibitors targeting specifically the malarial *Pf*mrk [29] and used the model template for searching the in-house chemical database to identify new potential inhibitors. The procedure resulted in the discovery of sixteen potent *Pf*mrk inhibitors [29]. The predicted inhibitory activity of some of these *Pf*mrk inhibitors from the molecular model agree exceptionally well with the experimental inhibitory values from the *in vitro* CDK assay [29]. Statistically, the most significant model obtained by us was found to contain two hydrogen-bond acceptor functions and two hydrophobic sites including one aromatic-ring hydrophobic site [29]. Although the model was not developed from X-ray structural analysis of the known CDK2 structure, it was found to be consistent with the structure-functional requirements for binding of the CDK inhibitors in the ATP binding pocket. Mapping of the pharmacophore on known CDK inhibitors that were tested in our *Pf*mrk assay such as, (a) indirubin, (b) staurosporine, (c) kenpaullone, (d) WR032428, and (e) WHI-P180 were observed to be consistent with the model [29]. Despite complexity of the malarial CDK activity, predicted *Pf*mrk inhibition from the developed pharmacophore model was quite robust and can be useful for further design of selective *Pf*mrk inhibitors and assessing the subtle differences in structure-function information between *Pf*mrk and other CDKs [82].

In another recent effort on structure-based drug discovery of antimalarials, we developed a phamacophore model for malarial *Pf*KASIII inhibitory activity and successfully utilized it to identify several *Pf*KASIII inhibitors [48] from calculated stereo-electronic profiles of one known related inhibitor, TLM. We utilized the electrostatic potential profile, Fig. (**3b**) of TLM for developing the model. The large extended negative electrostatic potential regions by the carbonyl oxygen atom and sulfur atom in TLM were considered as centers for two hydrogen bond acceptors and the region by the ethylene hydrogen atoms (weak electrostatic potential region) as the hydrophobic site in the molecule. A preliminary model for inhibition of KASIII was constructed, Fig. (**3c**) using these features on the optimized geometry of TLM and converting the molecular structure of TLM into a 3D shape with the features, a combined template was generated, Fig. (**3d**). This shape based pharmacophore template was used to conduct the *in silico* screen of our in-house multi-conformer chemical database in an iterative manner which resulted in the identification of several new *Pf*KASIII inhibitors [48]. Thus, a combined approach of stereo-electronic and pharmacophore profile generations from known inhibitors could be useful for discovery novel inhibitors.

In another recent effort, we developed a pharmacophore model for chalcones from the data of an in-house chalcone project and identified several new antimalarial agents through database searches, which was also helpful in the design of a few new potent antimalarials [83-85]. Chalcones are known to rapidly metabolize by liver microsomes but chalcones analogues with modified enone linker were found to have significant improvement in metabolic stability. Our goal was to identify compounds that share the antimalarial properties of the chalcones, but lacking the enone structure. However, despite understanding the structural basis for antimalarial activity of the chalcones, its pharmacophore for activity remained unknown. We developed the first pharmacophore model for chalcones to obtain both structural and functional requirements for antimalarial activity. The model enabled identification of several new antimalarial agents and facilitated the design of novel analogues [83]. The generated pharmacophore contained an aromatic and an aliphatic hydrophobic site, one hydrogen bond donor site, and a ring aromatic feature distributed over a three dimensional space [84]. The activity of the compounds estimated by the pharmacophore was found to correlate well with those determined experimentally. Two of the identified compounds were found to be highly potent *in vitro* against all five strains of *P. falciparum* tested. Moreover, one compound showed significant potency in a malaria-infected mouse model [84]. The model was also reported to be useful for design of novel antimalarials [85]. The study therefore demonstrated how the chemical features of a set of diverse chalcones and chalcone-like compounds could be organized to develop a pharmacophore for antimalarial activity and be utilized for discovery and design of novel antimalarials.

Yet, in another recent study, we [86] developed a pharmacohore model of the antimalarials, 4(1*H*)-quinolones, known to be highly effective in inhibiting the replication of *P. falciparum* along with synergism to the well known antimalarial atovaquone (Malarone). We not only developed a model for antimalarial activity of the quinolones consisting of two aliphatic hydrophobic functions and one aromatic ring hydrophobic function but went beyond the pharmacophore to calculate mathematical descriptors directly from the identified molecular structures and reasonably well predicted the antimalarial activity of the compounds [86].

Antileishmanial Compounds

Several antimalarial indolo[2,1-b]quinazoline-6,12-dione (tryptanthrin) derivatives which were originally screened for our malaria study also exhibited antileishmanial activity at concentrations below 100 ng/ml when tested against *Leishmania donovani* amastigotes *in vitro* in our laboratory [87]. We reported a quantitative structure-activity relationship study between the *in vitro* antileishmanial activity, a 3D phramacophore for antileishmanial activity and molecular electronic properties of 27 analogs of indolo [2,1-b]quinazoline-6,12-dione (tryptanthrins) [87]. The procedure adopted in the study was a combination of semi-empirical AM1 quantum chemical, cyclic voltammetry and pharmacophore generation based on 3D-QSAR-CATALYST methods. A modest to a fairly accurate correlation was observed between activity and the calculated molecular properties such as the molecular density, octanol-water partition coefficient, lowest unoccupied molecular orbital energies, and redox potentials measured by cyclic voltammetry experiments. The generated pharmacophore was a reasonably well predictive model for antileishmanial activity of the tryptanthrins [87]. The carbonyl group of the 5-member ring in the indolo[2,1-b]quinazoline-6,12-dione skeleton and the electron transfer ability to this oxygen atom were attributed to be crucial for antileishmanial activity of these compounds. The validity of the model could be extended to structurally different class of potent antileishmanial compounds through virtual screening of databases and *in vitro* toxicity studies on the identified compounds in both macrophage and neuronal lines for favorable properties thus, opening new chapters for further antileshmanial chemotherapeutic study [87].

Werbovetz *et al*. previously demonstrated antileishmanial activity of several dinitroaniline sulfonamides against *Leishmania* parasites [88-90]. In continuation of the efforts and to further explore the functional features responsible for antileishmanial activity of dinitroaniline sulfonamides, we reported a three-dimensional pharmacophore model for antileishmanial activity of the compounds [14]. The pharmacophore contained an aliphatic hydrophobic group, an aromatic hydrophobic group, an aromatic functionality and a hydrogen-bond acceptor in specific regions of space [14]. It was used for search of databases of drug-like compounds, particularly commercial databases. From a search of 55,000-compound Maybridge database, we found several compounds that fit to the pharmacophore. Nineteen of the most promising compounds were tested for antileishmanial activity. Two compounds were found to be highly potent (IC_{50} values under 5 μM) and another five compounds were moderately active (IC_{50} values between 20 and 40 μM) [14]. Unlike the dinitroaniline sulfonamides, the active compounds were not found to display antimitotic effects against Leishmania [14]. However, despite not possessing the expected mechanism of action, the active compounds were found to be potently antileishmanial *in vitro* and found to affect the integrity of the parasite mitochondrion. Thus, our pharmacophore based screening approach could provide novel active compounds to open up new scope for further study for chemotherapy of leishmaniasis.

Insect Repellents

In pursuit of our goal to discover novel arthropod repellents and to better understand the mechanism of insect repellency of DEET and DEET-like repellents, we performed a three-dimensional quantitative structure-activity (QSAR) study and developed a pharmacophore model for potent repellent activity from a set of eleven known diverse insect repellents using the CATALYST methodology [65]. The generated model contained three hydrophobic sites and a hydrogen-bond acceptor site in specific locations around the three dimensional space of the compounds which were found to be crucial for potent repellent activity [65].

The pharmacophore showed an excellent correlation (correlation = 0.9) between the experimental protection time afforded by the compounds in the training set and their predicted protection time. The validity of the pharmacophore model goes beyond the list in the training set and is found to map quite well onto a variety of other insect repellents, including a highly potent repellent compound that was extracted from the hair of Gaur, an animal frequently seen in South East Asia, Fig. (**6b**). By mapping this model onto one of the identified potent analogue, we generated a three-dimensional shape based template that allowed a search our in-house compound database to discover four new potential insect repellent candidates [65]. A U.S. patent (# 7,897,162) on the model and discovery of new arthropod repellents was issued recently [91].

CONCLUDING REMARKS

The calculated stereo-electronic property of the antimalarial, antileishmanial, and the insect repellent agents presented in this chapter provided important molecular electronic insights to guide our understanding of the “interaction pharmacophores” and generation of the phramacophore models which were crucial for identification of new bioactive agents.

The 3D-QSAR pharmacophores on known antimalarial, antileishmanial, and insect repellent compounds demonstrated a new computational approach for organizing the molecular characteristics from a set of structurally diverse compounds to a model that were both statistically and mechanistically significant for potent activity and useful for identification of novel analogues. The models were also useful for unraveling the possible rationale for target-specificity of the compounds. Because the target proteins for many these known agents may remain unknown, developing pharmacophores could be very useful not only to identify new potent compounds but also to obtain insights about the possible mode of interaction at the active site. Furthermore, the *in silico* models can also be very useful for design of more efficacious novel therapeutic agents. Overall, the *in silico* approaches presented here can maximize the efficiency for discovery of these agents.

ACKNOWLEDGEMENTS

Material has been reviewed by the Walter Reed Army Institute of Research. There is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the author and are not to be construed as official or reflecting true views of the Department of the Army or the Department of Defense. The author gratefully acknowledges valuable insights in improving the manuscript from COL James F. Cummings, Director, Division of Regulated Activities, WRAIR.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

ABBREVIATIONS

CQ	=	Chloroquine
DDT	=	Dichlorodiphenyltrichloroethane
AIDS	=	Acquired immune deficiency syndrome
DEET	=	*N,N*-Diethyl-*m*-toluamide
WHO	=	World Health Organization
ACT	=	Artemisinin Combination Therapy

HTS	=	High Throughput Screening
ADME	=	Absorption, distribution, metabolism and excretion
HPBCD	=	Hydroxypropyl-β-cyclodextrin
NMR	=	Nuclear magnetic resonance
FAS	=	Fatty Acid Synthase
*Pf*ACP	=	*Plasmodium falciparum* acyl carrier protein
ACP	=	Acyl carrier protein
*Pf*KASIII	=	ß-ketoacyl-ACP synthase III
TLM	=	Thiolactomycin

REFERENCES

[1] World Malaria Report **2009**; World Health Organization: Geneva, **2008**. http://malaria.who.int/whosis/whostat/EN_WHS09_Full.pdf

[2] (a) Wells, T.N.C.; Alonso, P.L.; Gutteridge, W.E. New medicines to improve control and contribute to the eradication of malaria. *Nat. Rev. Drug Discov.,* **2009,** *8*, 879-891. (b) Pink, R.; Hudson, A.; Mouries, M.A.; Bendig, M. Opportunities and challenges in antiparasitic drug discovery. *Nat. Rev. Drug Discov*., **2005,** *4*, 727-740. (c) Trigg, P.I.; Kondrachine, A.V. In *Malaria Parasite Biology, Pathogenesis and Protection: The current global malaria situation*; Sherman, I.W., Eds; ASM Press: Washington, D.C., **1998**; Chapter 2, pp 11-22.

[3] Vroman, J.A.; Gaston, M.A.; Avery, M.A. Current progress in the chemistry, medicinal chemistry and drug design of artemisinin based antimalarials. *Curr. Pharm. Design.,* **1999,** *5,* 101-138.

[4] Brewer, T.G.; Grate, S.J.; Peggins, J.O.; Weina, P.J.; Petras, J.M.; Levine, B.S.; Heiffer, M.H.; Schuster, B.G. Fatal neurotoxicity of arteether and artemether. *Am. J. Trop. Med. Hyg.* **1994,** *51*, 251-259.

[5] Bhattacharjee, A.K.; Karle, J. M. Stereoelectronic properties of antimalarial artemisinin analogues in relation to neurotoxicity. *Chem. Res. Toxicol.* **1999,** *12*, 422-428.

[6] World Health Organization. http://www.who.int/emc/diseases/leish/leisdis1.html. http://www.who.int/leishmaniasis

[7] Seaman, J.; Mercer, A.; Sondorp, E. The epidemic of visceral leishmaniasis in Western Upper Nile, southern Sudan: course and impact from 1984 to 1994. *Int. J. Epidemiol*. **1996,** *25,* 862-871.

[8] Werbovetz, K. Target-based drug discovery for malaria, leishmaniasis, and trypanosomiasis, *Current Medicinal Chemistry*, **2000,** *7*, 835-860.

[9] (a) Cook, G.C. Leishmaniasis: some recent developments in chemotherapy. *J. Antimicrob. Chemother.* **1993,** *31*, 327-330. (b) Olliaro, P.L.; Bryceson, A.D.M. Practical progress and new drugs for changing patterns of Leishmaniasis. *Parasitol. Today* **1993,** *9*, 323-328.

[10] Singh, S.; Sivakumar, R. Challenges and New Discoveries in the Treatment of Leishmaniasis. *J. Infect. Chemother.* **2004,** *10*, 317-315.

[11] Davis, A. J.; Kedzierski, L. Recent Advances in Antileishmanial Drug Development. *Curr Opin Investig Drugs.* **2005,** *6*, 163-169.

[12] Ouellette, M.; Drummelsmith, J.; Papadopoulou, B. Drugs in the Clinic, Resistance and New Developments. *Drug Resist Updat.* **2004,** *7*, 257-266.

[13] Croft, S. L.; Vivas, L.; Brooker, S. Recent Advances in Research and Control of Malaria, Leishmaniasis, trypanosomiasis and schistomiasis. *East Mediterr Health J.* **2003,** *9*, 518-533.

[14] Delfín, D.A.; Bhattacharjee, A.K.; Yakovich, A.;Werbovetz, K.A. Identification and Evaluation of Novel Antileishmanial Compounds through *In Silico* Three-Dimensional Pharmacophore Development and Database Searching. *J. Med. Chem.,* **2006,** *49*, 4196-4207.

[15] Magill, A. J.; *Leishmaniasis*. In *Tropical Medicine and Emerging Infectious Diseases*, 8th ed.; Strickland, G.T., Ed.; W.B. Saunders Co.: Philadelphia, PA, **2000**; pp. 665-687.

[16] Croft, S. L.; Sunder, S; Fairlamb, A.H. Drug resistance in leishmaniasis. *Clin. Microb. Rev.,* **2006,** *19*, 111-126.

[17] Eldridge, B.F.; Edman, J.D., Eds., *Medical Entomology: A Textbook on Public Health and Veterinary Problems Caused by Arthropods,* PUBLISHER, CITY, 2000.

[18] Samant, M; Dube, A. Leishmaniasis: an overview. Drugs and pharmaceuticals Current R & D Highlights (Leishmaniasis), CENTRAL DRUG RESEARCH INSTITUTE, Documentation and Library Services Division (2008).

[19] (a) McCabe, E.T.; Barthel, W.F.; Gertler, S.I.; Hall, S.A. Insect repellents. III. N,N-diethylamides, *J. Org. Chem.*, **1954,***19*, 493-498. (b) Fradin, M.S. Mosquitoes and mosquito repellents: a clinician's guide. *Ann Intern Med.,* **1998,** *128*, 931. (c) Gilbert, I.H.; Gouck, H.K.; Smith, C.N. New mosquito repellents, *J. Econ. Entomol.,* **1955,** *48*, 741.

[20] Bhattacharjee, A.K.; Ma, D.; Karle, J.M.; Gupta, R.K. Molecular similarity analysis between insect juvenile hormone and N,N-diethyl-m-toluamide (DEET) analogs may aid design of novel insect repellents. *J. Mol. Recognit.,* **2000,** *13*, 213-220.

[21] Janseen, D. The power of prediction. *Drug Disc*., ISSUE, **2002,** 38.

[22] Podlogar, B.L.; Muegge, I.; Brice, L.J. Computational methods to estimate drug development parameters, *Curr. Opin. Drug Disc.* **2001,** *12*, 102.

[23] Walters, W. P.; Stahl, M. T.; Murko, M. A. Virtual screening - an overview. *Drug Discov. Today* **1998,** *3*, 160-178.

[24] Lyne, P. D. Structure-based virtual screening - a review. *Drug Discov. Today* **2002,** *7*, 1047-1055.

[25] Leach, A.R.; Gillet, V.J.; Lewis, R.A.; Taylor, R. Three-Dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010,** *53*, 539-558.

[26] Guner, O. In *Pharmacophore Perception, Development and Use in Drug Design* IUL Publishers, Biotechnology Series, **2000**.

[27] P. Gund. Three dimensional pharmacophore pattern searching, *Prog Mol Subcell Biol.,* **1977,** *5*, 117-143.

[28] Bhattacharjee, A.K.; Hartell, M.G.; Nichols, D.A.; Hicks, R.P.; Stanton, B.; van Hamont, J.E.; Milhous, W.K. Structure-activity relationship study of antimalarial indolo [2,1-b]quinazoline-6,12-diones (tryptanthrins). Three dimensional pharmacophore modeling and identification of new antimalarial candidates. *European J. Med. Chem.*, **2004,** *39*, 59-67.

[29] Bhattacharjee, A.K.; Geyer, J. A.; Woodard, C.L.; Kathcart, A.K.; Nichols, D.A.; Prigge, S.T.; Li, Z., Mott, B.T.; Waters, N.C. A Three Dimensional *In Silico* Pharmacophore Model for Inhibition of *Plasmodium Falciparum* Cyclin Dependent Kinases and Discovery of Different Classes of Novel Pfmrk Specific Inhibitors. *J. Med. Chem.,* **2004,** *47*, 5418-5426.

[30] Ehrlich P. Über den jetzigen Stand der Chemotherapie. *Chem. Ber.* **1909,** *42*,17. Quoted by Ariens, E.J. Molecular pharmacology, a basis for drug design. *Prog. Drug Res.,* **1966,** *10*, 429.

[31] Bhattacharjee, A. K.; Kuča, K.; Musilek, K.; Gordon, R.K. *In Silico* Pharmacophore Model for Tabun-inhibited Acetylcholinesterase (AChE) Reactivators: a Study of their Stereoelectronic Properties. *Chem. Res. Toxicol.,* **2010,** *23*, 26-36.

[32] Bhattacharjee, A. K.; Gordon, J. A.; Marek, E.; Campbell, A.; Gordon, R.K. 3D-QSAR studies of 2,2-diphenylpropionates to aid discovery of novel potent muscarinic antagonists. *Bioorg. & Med. Chem.* **2009,** 17, 3999–4012.

[33] Buchwald, P.; Bodor, N. Computer-aided drug design: the role of quantitative structure-property, structure-activity and structure-metabolism relationships (QSPR, QSAR, QSMR), *Drug Future***, 2002,** *27*, 577-588**.**

[34] Wernsdorfer, W.; McGregor, I. (Eds.) Malaria: principles and Practice of malariology; Churchill Livingstone, New York, **1988**, pp 1818.

[35] Thomson, C.; Cory, M.; Zerner M. Theoretical studies of some new antimalarial drugs. *Int. J. Qunat. Biol Symp*., **1991,** *18*, 231-245.

[36] Bernardinelli, G.; Jefford, C.W.; Maric, D.; Thomson, C.; Weber, J. Computational studies of the structures and properties of potential antimalarial compounds based on the 1,2,4-trioxane ring structure. I. Artemisinin-like molecules. *Int. J. Qunat. Chem. Quant. Biol Symp.* **1994,** *21*, 117-131.

[37] Shukla, K.L.; Gund, T.M.; Meshnick, S.R. Molecular modeling studies of the artemisinin (qinghaosu) – hemin interaction. Docking between the antimalarial agent and its putative receptor. *J. Mol. Graphics,* **1995,** *13,* 215-222.

[38] Cheng, F.; Shen, J.; Luo, X; Zhu, W.; Gu, J.; Ji, R.; Jiang, H.; Chen, K. Molecular docking and 3-D-QSAR studies on the possible antimalarial mechanism of artemisinin analogues. *Bio. Org. Med. Chem.* **2002,** *10*, 2883-2891.

[39] Brewer, T.G.; Grate, S.J.; Peggins, J.O.; Weina, P.J.; Petras, J.M.; Levine, B.S.; Heiffer, M.H.; Schuster, B.G. Fatal neurotoxicity of arteether and artemether. *Am. J. Trop. Med. Hyg.* **1994,** *51*, 251-259.

[40] Wesche, D.L.; DeCoster, M.A.; Tortella, F.C.; Brewer, T.G. Neurotoxicity of artemisinin analogs *in vitro*. *Antimicrob. Agents Chemother*. **1994,** *38*, 1813-1819.

[41] Girones, X.; Gallegos, A.; Carbo-Dorca, R. Modeling antimalarial activity: application of kinetic energy density quantum similarity measures as descriptors in QSAR. *J. Chem. Inf. Comp. Sci.,* **2000,** *40,* 1400-1407.

[42] Illapakurthy, A.C.; Sabnis, Y.A.; Avery, B.A.; Avery, M.A.; Wyandt, C.M. Interaction of artemisinin and its related compounds with hydroxypropyl-beta-cyclodextrin in solution state: experimental and molecular modeling studies. *J Pharm Sci.* **2003,** *92*, 649-655.

[43] Hartell, M.G.; Hicks, R.; Bhattacharjee, A.K.; Koser, B.W.; Carvalho, K.; Van Hamont, J.E. NMR and molecular modeling analysis of the interaction of the antimalarial drugs artelinic acid and artesunic acid with -cyclodextrin. *J. Pharm. Sci.* **2004,** *93,* 2076-2089.

[44] Tonmunphean, S.; Wijitkosoom, A.; Tanitirungrotechai, Y. Influence of stereoisomer of dispiro-1,2,4,5-tetraoxanes on their binding mode with heme and on antimalarial activity: molecular docking studies. *Bio Org & Med Chem.* **2004,** *12*, 2005-2012.

[45] Bhattacharjee, A.K.; Skanchy, D.J.; Hicks, R.P.; Carvalho, K.A.; Chmurny, G.N.; Klose, J.R.; Scovill, J.P. Structure of β-Artelinic acid clarified using NMR analysis, molecular modeling & cyclic voltammetry, and comparison with α-artelinic acid and β-arteether. *Internet Elec. J. Mol. Design.* **2004,** *3*, 55-72.

[46] Bhattacharjee**,** A.K.; Karle, J.M**.** Role of molecular electronic properties of some novel antimalarial cyclic peroxy ketals in relation to potency and potential toxicity. *Molecular Engineering.* **1999,** *8*, 391-402.

[47] Bhattacharjee, A. K.; Carvalho, K.A.; Opsenica, D.; Šolaja, B. A. Structure-activity relationship study of steroidal 1,2,4,5-tetraoxane antimalarials using computational procedures. *J. Serb. Chem. Soc.*, **2005,** *70*, 329-345.

[48] Lee, P.J.; Bhonsle, J.B.; Gaona, H.W.; Huddler, D.P.; Heady, T.N.; Kreishman- Deitrick, M.; Bhattacharjee, A. K.; McCalmont, W.F.; Gerena, L.; Lopez-Sanchez, M.; Roncal, N.E.; Hudson, T.H.; Johnson, J.D.; Prigge, S.T.; Waters; N.C. Targeting the fatty acid biosynthesis enzyme, ß-ketoacyl –acyl carrier protein synthase III (*Pf*KASIII) in the identification of novel antimalarial agents. *J. Med. Chem.* **2009,** *52*, 952-963.

[49] Heath, R.J.; White, S.W.; Rock, C.O. Inhibitors of fatty acid synthesis as antimicrobial chemotherapeutics. *Appl. Microbiol. Biotechnol.* **2002,** *58*, 695-703.

[50] Sakya, S.M.; Suarez-Contreras, M.; Dirlam, J.P.; O'Connell, T.N.; Hayashi, S.F. Syntheis and structure-activity relationships of thiotetronic acid analogues of thiolactomycin. *Bioorg. Med. Chem Lett*, **2001,** *11*, 2751-2754.

[51] Jackowski, S.; Murphy, C.M.; Cronan, J.E. Jr.; Rock, C.O. Acetoacetyl-acyl carrier protein synthase. A target for the antibiotic thiolactomycin. *J Biol Chem* 1989. 264, 7624-7629. Miyakawa, S.; Suzuki, K.; Noto, T.; Harada, Y.; Okazaki, H. Thiolactomycin, a new antibiotic. IV. Biological properties and chemotherapeutic activity in mice. *J Antibiot (Tokyo)* **1982,** 35, 411-419.

[52] SPARTAN, version 5.0, **2001,** Wavefunction, Inc., Irvine, CA.

[53] Dewar, M.J.S.; Zoebisch, *E.G.*; Horsley, E.F.; Stewart, J.J.P. New general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.,* **1985,** *107*, 3902-3909.

[54] Frisch, M. J.; Trucks, G.W.; Schlegel, H.B.; Gill, P.M.W.; Johnson, B.G.; Robb, M.A.; Cheeseman, J.R.; Keith, T.A.; Petersson, G.A.; Montgomery, J.A.; Raghavachari, K.; Al-Laham, M.A.; Zakrzewski, V.G.; Ortiz, J.V.; Foresman, J.B.; Cioslowski, J.; Stefanov, B.B.; Nanayakkara, A.; Challacombe, M.; Peng, C.Y.; Ayala, P.Y.; Chen, W.; Wong, M.W.; Andres, J.L.; Replogle, E.S.; Gomperts, R.; Martin, R.L.; Fox, D.J.; Binkley, J.S.; Defrees, D.J.; Baker, J.; Stewart, J.P.; Head-Gordon, M.; Gonzalez, C.; Pople, J.A. Gaussian 94 (Revision A.1), Gaussian, Inc., Pittsburgh PA. **1995**.

[55] Murray, J.S.; Zilles, B.A.; Jayasuriya, K.; Politzer, P. Comparative analysis of the electrostatic potentials of dibenzofuran and some dibenzo-p-dioxins. *J. Am Chem. Soc.,* **1986,** *108*, 915-918.

[56] Du, Q.; Arteca, G.A. Modeling lipophilicity from the distribution of electrostatic potential on a molecular surface, *J. Comput. Aided Mol. Des*., **1996,** *10*, 133-144.

[57] Werbovetz, K.A.; Brendle, J.J.; Sackett, D.L. Purification, characterization and drug susceptibility of tubulin from leishmania. *Mol. and Biochem. Parasitology.* **1999,** *98*, 53-65.

[58] Ray, S.; Hazra, B.; Mittra, B.; Das, A.; Majumder, H.K. Diospyrin, a bisnaphthoquinone: a novel inhibitor of type I DNA topoisomerase of leishmania donovani. *Mol. Pharmacol.* **1998,** *54*, 994-999.

[59] Mukhopadhyay, R.; Kapoor, P.; Madhubala, R. Antileishmanial effect of a potent s-adenosylmethionine decarboxylase inhibitor: CGP 40215A. *Pharmacol. Res.* **1996,** *33*, 67-70.

[60] Croft, S.L.; Snowdon, D.; Yardley, V. The activities four anticancer alkyllysophospholipids against leishmania donovani, trypanosome cruzi and trypanosome brucei. *J. Antimicrob. Chemother.* **1996,** *38* 1041-1047.

[61] Bell, C.A.; Hall, J.E.; Kyle, D.E.; Grogl, M.; Ohemeng, K.A.; Allen, M.A.; Tidwell, R.R. Structure-activity relationships of analogs of pentamidine against Plasmodium falciparum and leishmania. *Antimicrob. Agents. Chemother.* **1990,** *34*, 1381-1386.

[62] Bhattacharjee, A.K. *In vitro* antileishmanial activity of some natural bisbenzylisoquinoline alkaloids could be correlated with their calculated molecular electronic properties. *Intl. J. Qunat. Chem.* **1999,** *75*, 995-1002.

[63] Werbovetz, K.A.; Bhattacharjee, A.K.; Brendle, J.J.; Scovill, J.P. Stereoelectronic Features Modulating the Biological Activity of Camptothecin Analogs. *Bioorg. & Med. Chem.* **2000,** *8*, 1741-1747.

[64] Ma, D.; Bhattacharjee, A.K.; Gupta, R.K.; Karle, J.M. Predicting mosquito repellent potency of DEET analogs from molecular electronic properties. *Am. J. Trop. Med. Hyg.,* **1999,** *60*, 1-6.

[65] Bhattacharjee**,** A. K.; Dheranetra, W.; Nichols, D.A.; Gupta, R.K. 3D pharmacophore model for insect repellent activity and discovery of new repellent candidates. *QSAR Comb. Sci.,* **2005,** *24*, 593-602.

[66] Geladi, P.; Kowalski, B.R. Partial least squares regression (PLS): a tutorial. *Analytica. Chimica Acta.* **1986,** *185,* 1-17.

[67] Cramer, 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.* **1988,** *43*, 5959-5967.

[68] Bobbyer, D.N.A.; Goodford, P.J.; McWhinnie, P.M. New hydrogen-bond potentials for use in determining energetically favorable binding sites on molecules of known structures. *J. Med. Chem.* **1989,** *32*, 1083-1094.

[69] CATALYST version 4.9, Accelrys, San Diego, CA (http://www.accelrys.com).

[70] Koide, Y.; Tatasui, A.; Hasegawa, T.; Nurakami, A.; Satoh, S.; Yamada, H.; Kazayama, S.; Takahashi, A. Identification of a stable chymase inhibitor using a pharmacophore based database search. *Bioorg Med Chem Lett.* **2003,** *13*, 25-29.

[71] Singh, J.; van Vlijimen, H.; Liao, Y.S.; Lee, W.C.; Cornebise, M.; Harris, M.; Shu, I.H.; Gill, A.; Cuervo, J.H.; Abraham, W.M.; Adams, S.P. Identification of potent and novel alpha4beta1 antagonists using *in silico* screening. *J. Med. Chem.* **2002,** *45*, 2988-93.

[72] Koide, Y.; Hasegawa, T.; Takahashi, A.; Endo, A.; Mochizudi, N.; Nakagawa, M.; Nishida, A. Development of novel EDG3 antagonists using a 3D database search and their structure-activity relationships. *J. Med. Chem.* **2002,** *45*, 4629-4638.

[73] Kurogi, Y.; Miyata, K.; Okamura, T.; Hashimoto, K.; Tsutsumi, K.; Nasu, M.; Moriyasu, M. Discovery of novel mesangial cell proliferation inhibitors using a three-dimensional database searching method. *J. Med. Chem.*, **2001,** *44*, 2304-2307.

[74] Chen, G.S.; Chang, C.S.; Kan, W.M.; Chang, C.L.; Wang, K.C.; Chern, J.W. Novel lead generation through hypothetical pharmacophore three-dimensional searching: discovery of isoflavonoids as nonsteroidal inhibitors of rat 5 alpha-reductase. *J. Med. Chem.* **2001,** *44*, 3759-3763.

[75] Schuster, D.; Maurer, E.M.; Laggner, C.; Nashev, L.G.; Wilckens, T.; Langer, T.; Odermatt, A. The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening. *J. Med. Chem.,* **2006,** *49*(12), 3454-66.

[76] Schuster, D.; Laggner, C.; Steindl, T.M.; Palusczak, A.; Hartmann, R.W.; Langer, T. Pharmacophore modeling and *in silico* screening for new P450 19 (aromatase) inhibitors. *J. Chem. Inf. Model.,* **2006,** *46*, 1301-11.

[77] Bhattacharjee, A.K.; Kyle, D.E.; Vennerstrom, J.L.; Milhous, W.K. A 3D QSAR pharmacophore model and quantum chemical structure activity analysis of chloroquine(CQ)-resistance reversal. *J. Chem. Info. Comput. Sci.*, **2002,** *42*, 1212-1220.

[78] Bhattacharjee, A.K.; Kyle, D.; Vennerstrom, J. Structural Analysis of Chloroquine-Resistance Reversal by Imipramine Analogs. *Antimicrob. Agents. Chemother.*, **2001,** *45,* 2655-2657.

[79] Riel, M.A.; Kyle, D.E.; Bhattacharjee, A.K.; Milhous, W.K. The efficacy of proton pump inhibitor drugs against Plasmodium falciparum *in vitro* and their probable pharmacophores. *Antimicrob. Agents. Chemother..* **2002,** *46*, 2627-2632.

[80] Hicks, R.P.; Nichols, D.A.; DiTusa, C.A.; Sullivan, D.J.; Hartell, M.G.; Koser, B.W.; Bhattacharjee, A.K. Evaluation of 4-azaindolo[2,1-*b*]quinazoline-6,12-diones' interaction with hemin and hemozoin: a spectroscopic, x-ray crystallographic and molecular modeling study. *Internet Elec. J. Mol. Design*, **2005,** *4*, 751-764.

[81] Bhattacharjee, A.K. *In silico* 3D pharmacophores for aiding discovery of the *Pfmrk* (*Plasmodium* Cyclin-dependent protein kinases) specific inhibitors for therapeutic treatment of malaria. *Expert Opin Drug Discov.* **2007,** *2(8)*, 1115-1127.

[82] Bhattacharjee, A.K. Antimalarial drugs. www.cen-online.org, **2010,** July 19, pp 3.

[83] Gutteridge, C.E.; Nichols, D.A.; Curtis, S.M.; Thota, D.S.; Vo, J.V.; Gerena, L.; Montip, G.; Asher, C.O.; Diaz, D.S.; DiTusa, C.A.; Smith K.S.; Bhattacharjee, A. K. *In vitro* and *in vivo* efficacy against *Plasmodium falciparum* and *in vitro* metabolism of 1-phenyl-3-aryl-2-propen-1-ones. *Bioorg. Med. Chem. Lett.*, **2006,** *16*, 5682.

[84] Bhattacharjee, A.K.; Nichols, D.A.; Gerena, L.; Roncal, N.; Gutteridge, C.E. An *in silico* 3D pharmacophore model of chalcones useful in the design of novel antimalarial agents. *Medicinal Chemistry*, **2007,** *3*, 317-326.

[85] Gutteridge, C.E.; Hoffman, M.M.; Bhattacharjee, A.K.; Milhous, W.K.; Gerena, L. *In vitro* efficacy of 7-benzylamino-1-isoquinolinamine against *plasmodium falciparum* related to the efficacy of chalcone. *Bioorg. Med. Chem. Lett.* **2011,** *21*, 786-789.

[86] Basak, S. C.; Mills, D.; Hawkins, D.M.; Bhattacharjee, A.K. Quantitative structure-activity relationship (QSAR) studies of antimalarial compounds from their calculated mathematical descriptors. *SAR and QSAR in Environ. Res*. **2010,** *21*, 103-125.

[87] Bhattacharjee, A.K.; Skanchy, D.J.; Jennings, B.; Hudson, T.H.; Brendle, J.J.; Werbovetz, K.A. Analysis of Stereoelectronic Properties, Mechanism of Action and Pharmacophore of Synthetic Indolo[2,1-b] quinazoline-6,12-dione Derivatives in Relation to Antileishmanial

Activity Using Quantum Chemical, Cyclic Voltammetry, and 3D-QSAR CATALYST Procedures. *Bioorg. Med. Chem.* **2002,** *10*, 1979-1989.

[88] Werbovetz, K. A.; Sackett, D. L.; Delfín, D.; Bhattacharya, G.; Salem, M.; Obrzut, T.; Rattendi, D.; Bacchi, C. *Mol. Pharmacol.,* **2003,** *64*, 1325-1333.

[89] Bhattacharya, G.; Salem, M.; Werbovetz, K. A. Antileishmanial dinitroaniline sulfonamides with activity against parasite tubulin. *Bioorg. Med. Chem. Lett.* **2002,** *12*, 2395-2398.

[90] Bhattacharya, G.; Herman, J.; Delfín, D.; Salem, M. M.; Barszcz, T.; Mollet, M.; Riccio, G.; Brun, R.; Werbovetz, K. A. Synthesis and antitubulin activity of N1- and N4-substituted 3,5-dinitroaniline sulfanilamides against African trypanosomes and leishmania. *J. Med. Chem.* **2004,** *47*, 1823-1832.

[91] Gupta, R.K.; Bhattacharjee, A.K.; Lee, D.M. Arthropod repellent pharmacophore models, compounds identified as fitting the pharmacophore models, and methods of making and using thereof. **2011,** March 1, *U.S. Patent ID* # 7,897,162.

CHAPTER 14

Molecular Taxonomy

Ray Hefferlin*

Physics Department, Southern Adventist University, Collegedale, Tennessee 37315, USA

Abstract: This chapter is for those in the field of mathematical chemistry who would like to practice their skills on other than normal molecules; for colleagues in the physics community with a curiosity about periodicities of particles from molecules to strings; and for specialists in informatics. Similarities are shown to exist in the constructions of periodic systems for five orders of particles.

Keywords: Mathematical chemistry, molecules, periodicity, periodic systems, periodic tables, atoms, sub-atomic particles, fundamental particles, strings, mesons, baryons, quarks, photons, isotopic spin, strangeness, nuclei, nuclear molecules, magic numbers, Pascal's triangle.

INTRODUCTION

This chapter is written to pique curiosity about developments at the intersection of chemistry, physics, and informatics — periodic systems of molecules, where molecules are understood to mean objects composed of atoms or of sub-atomic particles. We consider fundamental particles (manifestations of strings); mesons and baryons (formed of quarks); nuclei (structured from nucleons); and molecules, as commonly understood, (constituted of atoms). In each case, a periodic system emerges. The periodic chart of the elements is well-known; structurally simpler charts, in non-science use, are calendars, tables of conjugations and declensions, and multi-verse hymns and folk songs. The underlying designs of periodic systems, and some of their successes, are featured.

***Corresponding author Ray Hefferlin:** Physics Department, Southern Adventist University, Collegedale, P.O. Box 1817, Collegedale, TN 37315, USA; Tel: 001 + 423-236-2869; Fax: 001 + 423-236-1669; E-mail: hefferln@southern.edu

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50014-X

STRINGS AND THEIR PERIODICITIES

> *There's plenty of room at the bottom*
>
> *(prophetically speaking of microscopic machines,*
>
> *but also applicable to strings)*
>
> *Richard Feynman*

What is String Theory?

String theory postulates that quarks, leptons, neutrinos, mesons, baryons, and the forces between them (mediated by gauge bosons) are all manifestations of the frequencies, or energies, of individual strings which vibrate and move about [1]. They do so in four-dimensional space-time, or on curled-up spaces having any one of several dimensions beyond the fourth. These spaces are located at *every* event in space-time. The two kinds of motions imply a spectrum of string frequencies, and therefore of energies, and hence of masses. Many of the masses can be associated with the scores of fundamental particles that have been found from analysis of experimental data. The strings are incredibly small, somewhere near to 10^{-35} meters.

What is Chung's Classification of the Possible String Energies?

Chung [2-4] has presented us with though-provoking, and elegant, periodic tables of quarks, leptons, neutrinos, mesons, baryons, and gauge bosons. His work is not what is known as the Standard Model, but just the same, we might consider it as a part of Fundamental Particle Chemistry. A short, approximate, version of his presentations will now be given.

He begins with the photon (which mediates the electromagnetic interaction), one-half the pion (the pion mediates the strong interaction), the Z_L^o (which mediates the weak left-handed non-conservation), the X_R and X_L (which mediate the weak CP right-handed and left-handed non-conservations, and the Z_R^o (which mediates the weak right-handed non-conservation). The photon has dimension 5, and the energies of the other, dimension 6 through 10, bosons are each a constant factor $(1/\alpha^2)$ greater than the one of the dimensionality below it.

Chung posits that strings may be on a torus of dimensionality 5, a smaller torus of dimensionality 6 looping the torus of dimension 5, a still smaller torus of dimensionality 7 looping the torus of dimension 6, and so on. The strings may also have "excited" states when the string is (on) a loop around the torus of a given dimension. The excited states are designated with the loop number. The higher the dimensionality and the higher the loop number, the higher the energy. Sub particles (*e.g*., sub quarks) with given masses exist on the loops. There may be more than one manifestation of any given particle, and a number indicates the dimension in (on) which each is found.

Some elementary particles and hadrons live on the tori of various dimensionalities but have loop numbers zero. The other particle masses were derived from sums and products of the sub particle masses by some sort of fitting algorithm; this is the only role that sub particles play. The results are enshrined in two periodic systems. The first is for strings manifested as gauge bosons, quarks, leptons and neutrinos, and it is shown in Table **1**. In the table, the force-mediating bosons, described two paragraphs above; the u and d quarks; and the e, υ_e, υ_μ, and υ_τ leptons (all are in boldface) have accepted energies; the *numbered* d, u, s, c, b, and t sub quarks and the *numbered* sub leptons are used to compute other quark and lepton energies. The charge superscripts, such as those on e^- and e^+, are omitted as the two masses are approximately equal. The energies of some entries in the quark column are related to energies of other particles, *e.g.* 3μ. The energies of the muon, tauon, and hidden muon are calculated by the equations:

$$\mu = e + \mu_7 \ldots \tag{1}$$

$$\tau = e + \tau_7 \ldots \tag{2}$$

$$\mu' = e + \mu_7 + \mu_8 \ldots \tag{3}$$

e is the electron (or, to this degree of precision, positron) mass/energy. The hidden muon is called such to avoid disturbing the standard model, which allows for just three leptons; according to Chung, it has just recently been discovered. According to Chung, this table is periodic because of similarities of entities in the same *rows* of various *columns*. The second periodic table, Table **2**, is of strings manifested as hadrons. The mesons are π, K, and η'; D, formed from one c quark and one or another of d, u, or s; η_c(1s) formed from two c quarks; B, formed from one b quark and one or one b quark and one or another of d, u, or s; and K, formed from two b quarks. The baryons are the proton, Ω, Ξ, and Λ_b. The charge superscripts are omitted and p is the same, to this precision in mass, as the neutron.

Table 1: The periodic table of string manifestations as quarks, leptons, and gauge bosons in terms of increasing dimensionality and hence mass-energy

Torus Dimension	**Quark/Lepton Loop Number**	**Gauge Boson**	**Quark**	**Lepton**
5	0	**Photon**	**u**	$\boldsymbol{\upsilon_e}$
6	0	$(\frac{1}{2})\boldsymbol{\pi_{1/2}}$	**d**	**e**
7	0	$\mathbf{Z_L^o}$	**3μ**	$\boldsymbol{\upsilon_\mu}$
7	1		d_7 and u_7	μ_7
7	2		s_7	τ_7
7	3		c_7	
7	4		b_7	
7	5		t_7	
8	0	$\mathbf{X_R}$	**μ’**	$\boldsymbol{\upsilon_\tau}$
8	1		b_8	μ_8
8	2		t_8	
9		$\mathbf{X_L}$		
10		$\mathbf{Z_R^o}$		

Table 2: The periodic table of string manifestations as mesons and baryons in terms of increasing dimensionality and mass-energy

Torus Dimension	**Baryon Loop Number**	**Meson**	**Baryon**
5	0		
6	0	**π**	
6	1	K	
7	0		
6	2	η’	p
6	3	D	Ω
6	4	η_c(1s)	Ξ
6	5	B	
6	6		Λ_b
6	7	K	

QUARKS AND THEIR PERIODICITIES

At an early stage of quark theory there were three quarks: d, u, and s. These three quarks were reverse-engineered from data for the particles shown in the meson and baryon diagrams shown above — they were *not* observed in any experiment. Two quarks (d for down and u for up) have charges $Q = -1/3$ and $+2/3$, in units of the electron charge e. Their average charge is $<Q> = +1/6$; their hypercharge Y is $2<Q> = +1/3$; and their isotopic spin components are $I_3 = -1/2$ and $+1/2$. The third quark, s (for strange), has a charge of $-1/3$, so $<Q> = -1/3$, $Y = -2/3$, and $I_3 = 0$. Plotting the appropriate values on I_3,Y coordinates gives the elegant d, u, and s quark triangle. Reversing the charges for antiquarks gives their triangle (Fig. (**1**)).

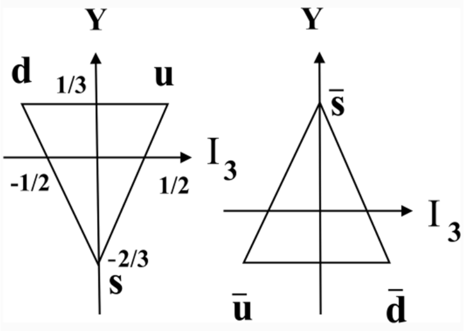


Figure 1: Quarks and antiquarks plotted on the isotopic spin component (horizontal axis) and strangeness (vertical axis). Antiquarks are designated with a bar above their names. With permission of PERIODIC SYSTEMS OF MOLECULES AND THEIR RELATION TO THE SYSTEMATIC ANALYSIS OF MOLECULAR DATA, Edwin Mellin Press [5].

Continuing particle discoveries forced the realization that more quarks are necessary to explain them, so the c (charm), t (top), and b (bottom) quarks were invented. The c quark can be drawn above Fig. (**1**) on its own axis, forming a tetrahedron. A t or b quark can be substituted for it, or in principle for a d or u quark, and that is why Fig. (**1**) is a valid periodic system.

HADRON PERIODICITIES

In a simple view of the the standard model, the left side of Fig. (**1**) is subjected to a double iterative affine transformation to produce the periodic system of baryons, Fig. (**2**). The procedure is:

1. Draw the quark triangle at the left side of Fig. (**1**),
2. Superimpose an identical triangle with its center on vertex d,
3. Identify the vertices of the new triangle as dd at top left, du at top right, and ds at the bottom,
4. Repeat the process for the u and s vertices of the original triangle, forming a larger triangle concentric with the original triangle,
5. At vertex dd of the triangle formed in step 4, again draw the original triangle of step 1 so that its *center* is on vertex dd,
6. Repeat the process for the du (which is the same as ud), uu, ds (same as sd), us (same as su), and ss vertices of the larger triangle, forming a still larger concentric triangle,
7. Identify the vertices of the new triangle as ddd at top left; ddu, dud, and udd next at right, and so on.

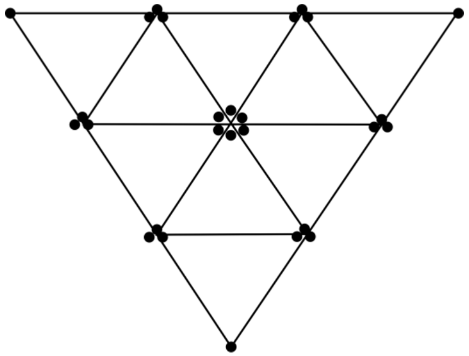


Figure 2: Schematic of the three quarks d, u, and s, taken three at a time as explained in the text. The vertices are ddd, uuu, and sss clockwise from upper left. Violation of the uncertainty principle is avoided by the introduction of a news quantum number, color. Only if each quark has one of the three primary colors, resulting in the combination being white, can it be observed as a hadron. The six hadrons in the center are dus, dsu, uds, usd, sdu, and sud.

Group theory dictates that this 27-particle diagram be decomposed into a decuplet with 10 particles, two octets with eight particles each, and a singlet with one particle**.** Adding the c quark to Fig. (**1**) creates more complex geometrical shapes for new baryons [6]. It is possible to substitute still heavier quarks into these shapes, thus rendering them true periodic systems [7].

Superposing the right side of Fig. (**1**) onto each vertex of the left side produces a hexagon, and more than one such hexagon is used to represent the numerous mesons that have been found experimentally [8]. It awes the author that usage of the triangles in Fig. (**1**) can so effectively reproduce the previously constructed meson and baryon diagrams.

NUCLEAR PERIODICITY

Given the nuclear magic numbers, we can ask "Can we construct a periodic system based on them?" Indeed, it can be done by cutting the Segré chart (Fig. (**3**)) in both axial directions, just after the magic numbers (because each nucleus occupies a tiny square on the chart). We obtain large squares and rectangles of various sizes and shapes. Recognizing that magic-number nuclei are special in that they are more stable than their neighbors, we orient the pieces of paper such that the edges containing magic number nuclei are on the front right side, on the front left side, or both, of each piece. Finally, we stack the squares and rectangles in such a way that those sides are aligned, as shown in Fig. (**4**). This is the "cut and stack" construction, which will be encountered later as one way to form a periodic system of chemical molecules. The procedure ignores the valley of stability, which tracks its way through some of the layers. Though aesthetically pleasing, this system has yet to facilitate any forecasts of nuclear properties.

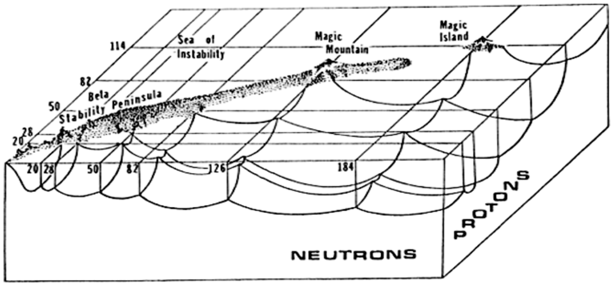


Figure 3: The Segré chart seen in three dimensions and at an angle. The negative binding energy increases in the *upward* direction; the valley of stability becomes the stability peninsula. From [9], with permission.

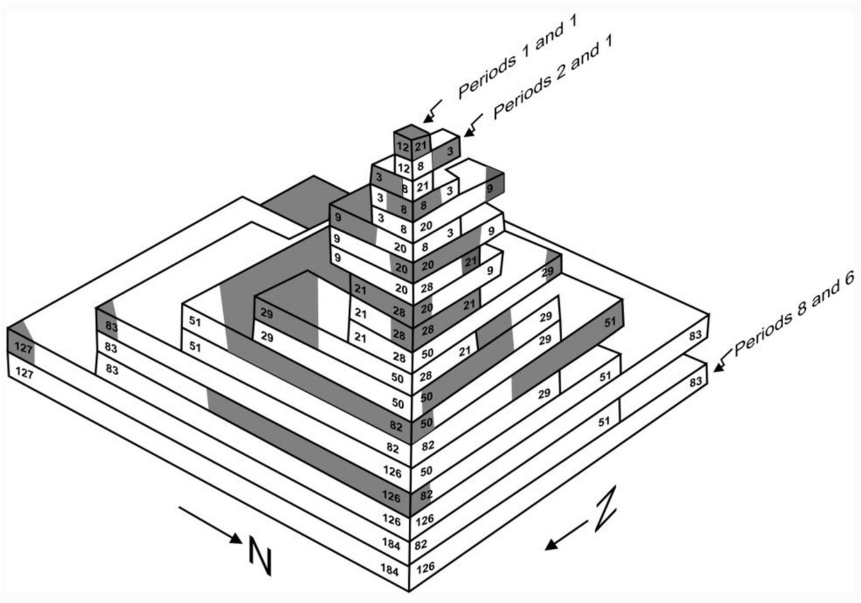


Figure 4: The periodic system for nuclei. Nuclei with magic numbers of protons or neutrons close nucleon shells, just as rare-gas atoms close atomic shells. The spreads of protons and neutrons are shown by numbers on the edges of the blocks. Passing a magic number causes the nuclei to enter a new period, just as is the case with atoms in the chart of the elements. The shaded regions show where the valley (or peninsula) of stability crosses the layers.

Nuclear molecules, where the nuclei are in close proximity, have been observed. Examples are ^{6}Be+^{6}Be [10] and ^{12}C+^{12}C [11]. These species have such strong force fields that pair production takes place between the two nuclei.

ATOMIC PERIODICITIES

Fig. (**5**), a popular cartoon by Harris, imagines an early attempt to classify elements. The heroic efforts made since that hypothetical conversation have culminated in over a hundred two- or three-dimensional periodic tables [12,13]. They include chemotopology [14], artificial intelligence [15], reduced potential curves [16], quantum computation (outside the scope of this chapter), information theory [17], and group dynamics [18,19]. It must be kept in mind that quantum computational usually produces data one atom at a time, and so the approach is equivalent to constructing the periodic system from experimentation. Group theory seems to be the only hope for producing a complete chart of the elements based on pure theory.

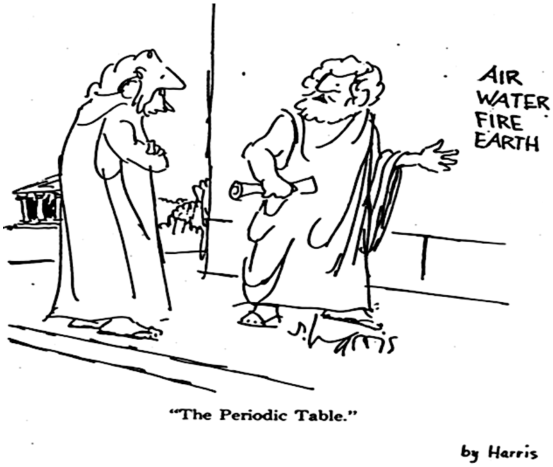


Figure 5: Two Greek philosophers discuss the first periodic chart of the elements, by Sydney Harris (used with permission).

Extrapolations of the chart to still heavier atoms than those of period 7 have been proposed [20,21], one speculative proposal going on to element 2,022 [22]. Contrariwise, it has been claimed on two grounds (not related to quantum computation) that there is an *upper limit* to *Z*, the atomic number [23, 24].

Parenthetically, it is of interest that these magic numbers have been found, *ex post facto*, to exist in Pascal's triangle [25]. They have been recognized as arc lengths between formula-derived radii of the golden-rectangle logarithmic spiral [26] and also as turning points of the boundaries of the domain of stable isotopes on a plot of proton-neutron ratio *vs.* atomic number [26].

MOLECULAR PERIODICITIES

Fig. (**6**), the companion to Fig. (**5**), shows what could have been the first perception of a diatomic molecule. Heroic and more or less successful efforts have been made since that supposed conversation to classify molecular compounds through identifying those that are similar by:

a) Using correlation methods [27].

b) Associating them with chosen independent variables such as the period and group numbers of the constituent atoms [28, 29].

c) Plotting them on a graph with “donors” and “acceptors” [30], or atom symbols such as C and H for alkanes [31,32], as axes.

d) Cutting the (Z_1,Z_2) plane of diatomic molecules parallel to each axis, just after atomic numbers, and stacking the resulting squares and rectangles (the “cut and stack” method) so that the magic-number molecules lie on one or two of two faces of the stack [33].

e) Considering the atomic chart as a matrix and forming the Kronecker product to produce the periodic system of diatomic molecules [34].

f) Partitioning them by the use of topological indices [35].

g) Classifying them with notations such as that of CAS.

h) Applying group theory [36,37].

Most of these proposals have been bolstered with efforts to predict properties of molecules. Complete reviews of molecular periodic systems can be found in [38, 39 (Fig. **7** of the chapter is in error and the correct figure will be sent upon request.)].

Figure 6: Two later Greek philosophers, fluent in English, discuss the beginning of an attempt to construct a periodic system of diatomic molecules formed from the chart shown in Fig. (**5**). Used by permission of Melissa Hefferlin.

SUMMARY

Since this chapter was intended to showcase periodic systems, it is proper to list here the various protocols for constructing such systems:

1. The "cut and stack" method begins with a two-dimensional map of objects, cuts this map at (really, just after) the magic numbers, and stacks the resulting quadrangles. This method has no *theoretical* basis:

    - A system of nuclei (See Section on Nuclear Periodicity).
    - Mollecular periodic system (Molecular Periodicity, item d).

2. The superposition scheme begins with a two dimensional map of objects, superposes the center points of a copy of this map (or its inverse) upon each object, and identifies the resulting larger map with combinations of the original objects. It was *not* the intention of the proposing investigator to use this scheme, though afterwards it serves as an excellent tool for additional understanding. The investigators, using their own schemes, have made very successful predictions:

    - Meson octets, and the baryon decuplet and octet, begin with quark maps (See Hadron Periodicities).
    - Kong's periodic systems of molecules begins with the element chart (Molecular Periodicities, item b).

3. The Kronecker-product procedure considers a map of object as a matrix, performs an outer product of the matrix with itself *n* times, and recognizes the product matrix as the projection of a (2*n*+2*)*-dimensional structure. This procedure encompasses several previously constructed periodic systems and has allowed some successful forecasts of spectroscopic constants with precisions and accuracies of less than 10 percent:

    - The Kronecker product periodic system of diatomic (triatomic) molecules (Molecular Periodicities, item e).

4. a: The group-dynamic protocol uses a group (chain) which conforms to some known behaviors of objects and creates multiplets of them; the ensemble of similar multiplets constitutes a system:

- The triangles and tetrahedra for three and four quarks (See Quarks and their Periodicitie); any one quark in a triangle or tetrahedron may be replaced by a heavier one.

- The periodic chart of the elements (Atomic Periodicities, last sentence of first paragraph).

4. b: This protocol allows moving up into periodic systems in the next higher space(s):

- Meson octets, and the baryon decuplet and octet (Hadron Periodicities).

    - any one diagram may be repeated for excited states of the baryons.

    - if a heavier quark is substituted for a lighter one, different hadrons appear.

- Periodic systems of diatomic (triatomic) molecules (Molecular Periodicities, item h).

ACKNOWLEDGEMENTS

The author sincerely thanks Ding-Yu Chung, former participant at SLAC and at SPIRES-HEP, and funding from an endowment made to Southern Adventist University.

CONFLICT OF INTEREST

The author confirms that this chapter contents have no conflict of interest.

REFERENCES

[1] Green, B.R. *The Elegant Universe*; W.W. Norton: New York, **1999**.

[2] Chung, D.Y. The periodic table of elementary particles. *arXiv:physics/0003023v1*, **2000**, 1-41.

[3] Chung, D.Y. The masses of elementary particles and hadrons. *arXiv:hep-ph/0003237v4*, **2001**, 1-25.

[4] Chung, D.Y. The periodic table of elementary particles and the composition of hadrons. *arXiv:physics/0111147v5*, **2004**, 1-32.

[5] Hefferlin, R. *Periodic Systems of Molecules and their Relation to the Systematic Analysis of Molecular Data*; Edwin Mellin Press: Lewiston, New York, **1989**.

[6] Halzen, F.; Martin, A.D. *Quarks and Leptons: an Introductory Course in Modern Particle Physics*. John Wiley and Sons: New York, **1984**.

[7] Abazov, V.M.; Abbott, B.; Abolins, M.; Acharia, B.S.; Adams, M.; Adams, T.; Aguilo, E; Ahn, S.H.; Absan, M.; Alexeev, G.D.; *et al.* (the CDF collaboration). Measurements of the lambda b lifetime in the exclusive decay lambda b → J/psi lambda. *Phys. Rev. Lett*. **2007**, *99*(14), 142001-1214007.

[8] Wikipedia. List of Mesons. http://en.wikipedia.org/wiki/List_of_mesons [Accessed 25th March 2011]

[9] Thompson, S.G.; Tsang, C.F. Superheavy elements, *Science*, **1972**, *178*, 1047-1055.

[10] Freer, M.; Collaboration. Exotic Molecular States in ^{12}Be. *Phys. Rev. Lett*. **1999,** 82, 1383–1386.

[11] Konnerth, D.; Dünnweber, W.; Hering, W.; Trautmann, W. Trombik, W.; Zipper, W.; Habs, D.; Hennerici, W.; Hennrich, H.J.; Kroth, R.; Lazzarini, A.; Rpnow, R.; Metag, V.; Simon, R.S. Correlated spin orientations in ^{12}C + ^{12}C molecular reonances, *Phys. Rev. Lett.* **1985**, *55*(6), 588-591.

[12] Mazurs, *E.G. Craphic Representations of the Periodic System during One Hundred Years*, 2nd ed.; University of Alabama Press: University, Alabama, **1974.**

[13] Van Spronsen, J.W. *The Periodic System of Chemical Elements*; Elsevier, Amsterdam, **1969**.

[14] Restrepo, G.; Mesa, H.; Llanos, E.J.; Villaveces, J.L. Topological study of the periodic system. *J. Chem. Inf. Comput*. **2004**, *44*, 68-75.

[15] Fayos, J. Atomic similarity through a neural network: self-associative periodic table of elements. In Carbo-Dorca, R.; Mezey, P.G. *Advances in Molecular Similarity*. JAI Press: Stamford, CT, **1998**; pp 205-214.

[16] Jenč, J.; Brandt, B.A., Ground-state reduced-potential curves and estimation of the dissociation energy of alkali-metal diatomic molecules. *Phys. Rev. A*. **1987**, *35*, 3784-3792.

[17] Bonchev, D. Periodicity of the chemical elements and nuclides: an information-theoretic analysis. In *The Mathematics of the Periodic Table*; Rouvray, D.H.; King, R.B., Eds.; Nova: New York, **2006**; pp. 161-188.

[18] Rumer, Y.B.; Fet, A.I. The group *spin*(4) and the Mendeleev system, *Theor. Maht. Phys.***1972**, *9*, 1081-1085.

[19] A. O. Barut, On the Group Structure of the Periodic Table of the elements. In *Structure of matter (Proceedings of the Rutherford Centennary Symposium, 1971*; Wybourne, B., Ed.; University of Canterbury Press: Canterbury, **1972**, pp. 126-136.

[20] Chaikhorsky, A.A. On some regularities of the periodic system and the chmical properties of the transuranium elements. J. Inorg. Nucl. Chem. **1976**, *Supplement*, 147-150.

[21] Wikipedia. Extended Periodic Table. http://en.wikipedia.org/wiki/Extended_periodic_table. [accessed 21st March 2011].

[22] Jueneman, F.B. Beyond superheavy elements. *Indust. Res. Devel.* **1979**, *September*, 17.

[23] Khazan, A. Upper limit in the periodic table of elements. *Progress Phys*. **2007**, *1*, 38-86.

[24] Khazan, A., Upper limit in the periodic table and synthesis of superheavy elements. *Progress Phys*. **2007**, *2*, 104-109.

[25] Weise, D. A Phythagorean approach to problems of periodicity in chemical and nuclear physics. In *Advanced topics in theoretical chemical physics*; Maruani, J.; Lefebre, R.; Brändas, E., Eds.; Springer: Dodrecht, **2003**; pp. 459-474.

[26] Boeyens, J.C.A. A molecular-structure hypothesis. *Int. J. Mol. Sci*. **2010**, *11*, 4267-4284.

[27] Johnson, B.A.; Maggiora, G.M. Concepts and Applications of Molecular Similarity.; Wiley: New York, **1990**.

[28] Górski, A.; Morphological classification of chemical structural units. *Polish J. Chem*. **2001**, *75*, 159-207.

[29] Carlson, C.M.; Gilkeson, J.; Linderman, K.; LeBlanc, S.; Hefferlin, R. Global forecasting of data using least-squares methods and molecular databases: a feasibility study using triatomic molecules. *Croat. Chem. Acta*. **1997**, *70*, 479-508.

[30] Hall, H.K. Jr. Toward an organic chemist's periodic table. *J. Chem. Ed*. **1980**, *57*(1), 49-51.

[31] Dias, J.R.; Setting the benzenoids to order, *Chem. Britain*, **1994,** *May*, 384 - 386.

[32] Dias, J.R.; Formula periodic tables – their construction and related symmetries. *J. Chem. Inf. Comp. Sci*. **1996**, *36*, 361-366.

[33] Hefferlin, R.; Campbell, R.; Gimbel, D.; Kuhlman, H.; Cayton, T.; The periodic table of diatomic molecules – an algorithm for retrieval and predication of spectrophysical properties. *J. Quant. Spectrosc. Radiat. Transfer*. **1979**, *21*, 315-336.

[34] Hefferlin, R. Matrix-product periodic systems of molecules. *J. Quant. Spectrosc. Radiat. Transfer*, **1994**, *34*, 314-317.

[35] Basak, S.B.; Mills, D.; Gute, B.D.; Natarajan, R. Predicting pharmacological and toxicological activity of heterocyclic compounds using QSAR and molecular modeling. *Top. Heterocycl. Chem.* **2006,** 39-80. DOI 10.1007/7081_025

[36] Zhuvikin, G.V. and R. Hefferlin, 1983. Periodicheskaya sistema dvukhatomnykh molekul: teoretiko-gruppovoi podkhod, *Vestnik Leningradskovo Universiteta*, **1983**, *16*, 10 – 16.

[37] Zhuvikin, G.V.; Hefferlin, R. Bosonic Symmetry and Periodic Systems of Molecules. In *Anales de física, Monographias 2,Group Theoretical Methods in Physics* (*Proceedings of the XIX International Colloquium: Salamanca*); del Olmo, M.A.; Santander, M.; Guilarte, J.M., Eds.; Real Sociedad Española de Física: Madrid, **1992:** pp. 358-361.

[38] Hefferlin, R.; Burdick, G.W. Periodic systems of molecules: physical and chemical. *Russ. J. Gen. Chem.* **1994**, *64*, 1659-1674.

[39] Hefferlin, R. Periodic Systems of Molecules. Presuppositions, Problems, and Prospects. In *Philosophy of chemistry*; Baird, D.; Scerri, E.; McIntyre, L., Eds.; Springer: Dordrecht, **2006**; pp. 221-243.

Subject Index

2-methoxyresorcinol 185, 186
10-hydroxycamptothecin 185, 186
1,1,2-trimethylcyclopropane 78, 79, 81

A

Active compounds 182, 183, 308, 319, 326
Active database structures 108, 109
Active molecules mean similarity 97, 98
Acyclic graphs 52, 56, 57, 58, 65, 67, 70
Adjacency matrix 3, 6, 43, 45, 163, 167, 172, 227, 230, 231, 251, 257, 259, 260, 261
Alkane data sets 76
Alkane molecules 56, 57, 60
Alkanes 58, 63, 76, 83, 85, 89, 122, 238, 345
Alloxantin 173, 175
Antileishmanial activity 316, 317, 325, 326
Antileishmanial compounds 316, 325
Antimalarial activity 309, 312, 322, 324, 325
Antimalarial compounds 309, 310, 320
Antimalarial drugs 304, 320, 322
Antimalarials 163, 188, 190, 308, 309, 310, 315, 316, 319, 321, 323, 324, 325, 327
Anti-tuberculosis drugs 222
Applications of molecular topology 161
Aromatic atoms 120, 134
Aromatic bonds 116, 118, 120, 121, 123, 131, 134, 298
Aromaticity detection 113
Artemisinins 304, 310, 311, 312
Atom classification 282, 288
Atom type distribution 287, 290, 291, 298
Atom type diversity (ATD) 282, 288, 300

B

Backtracking 113, 116, 129, 131, 132
Balaban index 3, 148, 155, 228
Based pharmacophore models 308, 319
Benzene's structures 24, 29, 127
Benzyne 126, 127
Best five-parameter models 61, 62
Binary weighting 100, 105
Bioactive molecules 310
Bio-concentration factor 222
Biodescriptors 4, 8, 13, 14, 15, 16, 267, 268
Biological activities 5, 43, 46, 182, 234, 254, 255, 308, 309, 310, 319
Boiling point values 84, 85
Bond multiplicities 120, 123, 132
Bourgas Indices 42, 68
Branched alkanes 76, 83, 84

C

CAESAR model 238, 239, 240
Caesar programs 222, 242
Caffeic acids 173, 175
Calculated chemodescriptors 13, 15
Calculated descriptors 8, 225, 253, 272
Calculated molecular descriptors 252, 263, 273
Calculated stereo-electronic properties 309, 311, 317, 327
Calculation of descriptors 223, 234
Carcinogenicity 222, 226, 237, 238, 239, 244, 289
Catalyst methodology 303, 320, 327
CDKs 323, 324
Chalcones 173, 174, 322, 324
Chembridge database 282, 285, 293
Chemical classification 239, 240
Chemical compounds 43, 46, 64, 267, 307
Chemical databases 92, 94, 323, 324
Chemical descriptors 8, 222, 251, 255
Chemical graph theory 6, 43, 143, 256
Chemical indices 5, 8, 258, 272
Chemical informatics 225, 226
Chemical nature 48, 49, 52
Chemical potential 196, 200, 202, 203, 204, 207, 211, 212, 214, 215, 217
Chemical reactions 28, 30, 210, 211, 215
Chemical reactivity 196, 203, 206, 207, 211, 216
Chemical regulation 223, 235, 237, 239
Chemical softness 196
Chemical structures 6, 13, 164, 165, 189, 190, 222, 223, 225, 226, 234, 236, 253, 255, 309
Chemodescriptors 8, 13, 14, 15, 16, 267, 268
Chemoinformatics 3, 16, 92, 95, 153
Chlorophenols 32, 33, 34
Chordal graphs 141, 142
Circular substructures 97, 103
Classical molecular connectivity indices 63
Classification functions 173, 174, 179, 180
Cluster analysis 149, 151, 152, 154
Cluster centers 287
Collection of comparability graphs 144, 153
Combinatorial chemistry 3, 5, 115, 187
Comparative correlational analysis 76
Comparative molecular field analysis (CoMFA) 162, 191, 275
Comparative molecular similarity indices analysis (CoMSIA) 162, 191

Subhash C. Basak, Guillermo Restrepo & José L. Villaveces (Eds)

10.1016/B978-1-68108-198-4.50019-9

Comparison of performances of model 270
Complete representation of molecular topology 47, 67
Complexity indices 54, 65, 68, 69
Complex properties 166, 253
Compound databases 303, 307, 308, 309, 319, 320, 327
Compounds, inactive 179, 182, 183
Computed chemodescriptors 15
Congeneric 5, 267, 272, 273, 275
Congenericity principle 251, 273
Connected components 123, 129, 132, 133, 134
Connected molecular graphs 129
Connectivity index 51, 52, 53, 61, 76, 77, 78, 79, 81, 82, 83, 85, 89, 259, 260
Connectivity indices 3, 7, 76, 82, 84, 85, 89, 161, 166, 167, 169, 171, 228, 251, 253
Connectivity isomers 113, 137
Connectivity matrix 76, 77, 78, 79, 80, 81, 82, 88, 89, 156
Connectivity matrix polynomial 76
Connectivity matrix spectra 76
Constant number 57, 60
Correct classification 179, 270, 272
Correlated descriptors 149, 153
Correlation coefficients 84, 85, 224, 238, 242
Correlation matrix 139, 146, 149, 150, 151, 152, 154, 155, 239
Corresponding comparability graphs 149, 150
Covalent bonds 116, 118, 120, 257
Critical volume 60, 61, 64
Cromolyn sodium 185, 186
Cubane 32, 123
Cyclic molecules 59
Cyclohexanes 126, 127, 299
Cyclomatic number 228, 238

D

Database structure fingerprints 107, 109
DEET compounds 318
Degeneracy 44, 46, 49, 69
Degree acyclic atoms 286
Degrees of vertices 78, 81, 89
Dendograms 148, 150, 151
Depicting carbon skeleton 78, 79, 81
Deriving molecular descriptors 77
Describing molecular structure 115
Developmental toxicity 222, 237, 238, 239, 240, 242, 246
Developmental toxicity (DT) 222, 237, 238, 239, 240, 242, 246
Diagonal forms 86, 87
Differential QSAR 3, 9
Diflubenzuron 177, 179
Dimensionality 337, 338
Diophantine equation 113, 133
Directed comparability graphs 142, 143
Discrete mathematical chemistry 3, 17, 251, 256
Discriminant function 171, 175, 178, 179, 182, 183
Diverse chemical classes 238, 239, 240
Diverse compounds 8, 327
Diverse sets of chemicals 273, 275
Diversity begets diversity principle 251, 272, 273
DNA 12, 317
DRAGON 7, 155, 175, 234, 235, 240
Drug database 282, 283, 287, 289, 290, 292
Drug discovery 3, 43, 162, 163, 233, 234, 284, 285, 300, 303
Druglike descriptors 282
Drug-like filters (DLFs) 282, 283, 284
Druglike Indices (DLI) 282, 283, 286, 287
Druglikeness 282, 283, 286, 287, 288, 289, 292, 299, 300
Drug molecules 298, 299
Drugs database 282

E

ECFC4 fingerprints 92, 97, 99, 103, 105
Edge-connectivity indices 82, 83, 85
Edge-product 76, 77, 78, 79, 80, 81, 83, 85, 88, 89
Edge-product-connectivity matrix 76, 77, 78, 79, 80, 89
Edge-sum-connectivity matrix 76, 77, 78, 82, 89
Electron density 198, 200, 202, 205, 206, 217, 231, 232, 314
Electronegativity 184, 196, 203, 206, 210, 218, 230
Electronegativity equalization principle (EEP) 196, 209, 210
Electronic properties 198, 306, 310, 311, 314
Electronic structure principles 209, 213
Electrophilicity 203, 211, 212
Electrophilicity equalization principle 196, 212
Electrophilicity index 196, 205, 206, 209, 218
Electrotopological descriptors 229, 234
Element symbol 116, 117, 125, 129
Entropy measures 147, 148, 150
Envelope models 251
Environmental sciences 222
Erlangen programmes 24, 26
E-state indices 3, 229, 240, 251
Ether-oxygen atoms 317
Ethylene hydrogen atoms 315, 316, 324
Exact molecular formulas 116, 121, 131, 133, 136
Exchange-correlation energy functionals 199, 214, 216
Excited states 211, 215, 338, 347
Existing compounds 135
Extended connectivities 42, 44, 45, 64, 65, 66

Extended connectivity fingerprint 92, 97, 110
Extended connectivity indices 65, 66

F

False negative (FN) 225, 238, 242
False positive (FP) 187, 225, 242
Flavones 173, 174
Flavonoids 173, 174, 175
Fluoroquinolones 222, 234
Formal concept analysis 24, 35
Fourth Indo-U 16, 17
Fragment weighting scheme 92, 99
Fundamental particles 336, 337
Fuzzy formula 132, 135

G

Gabapentin 287, 289
Gauge bosons 337, 338, 339
Given molecular formulas 113, 114
GNLSE 214, 217
Graph edit distance (GED) 139, 146, 147, 149, 151, 152, 154, 155
Graph entropies 139, 147
Graphical interaction models 120
Graph invariants 3, 6, 7, 12, 64
Graph non-adjacency numbers 42
Graph similarity measures 145, 154, 155
Graph-theoretical matrices 77, 147
Graph theoretical representation of molecules 6
Graph theoretic matrices 3, 6, 251
Graph theory 3, 42, 43, 46, 132, 139, 155, 251
Graph vertices 49, 50, 55, 58, 61, 65, 69
Ground state 198, 199, 201, 215
Group theory 115, 120, 341, 343

H

Halocarbons 8, 14, 15, 274
HAMLET 10
Hard-soft-acid-base (HSAB) 209, 210, 217
Hasse diagram technique (HDT) 32, 33
Heat of atomization 60, 61
Heteroatom 168, 169, 172, 178, 285, 290, 295, 296, 297, 298, 299
Heterogeneous groups 268, 269
Hexagon 30, 33, 342
Hexagonal symmetry 29, 30
Hexane substructures 126
Hierarchical quantitative structure-activity relationship 3, 251
Highest occupied molecular orbital (HOMO) 207, 216, 217, 232, 262, 318
Highest occupied molecular orbitals 217, 232, 262, 318
High throughput screening (HTS) 5, 307, 329
Homogeneous activity classes 101, 102
Hosoya index 3, 42, 51, 57
HSAB principle 196, 211, 215
Hydrogen atoms 6, 165, 169, 230, 231, 257, 262, 288
Hydrogen-filled graphs 6, 257, 260
Hydrogens 67, 116, 120, 121, 123, 134, 178, 292
Hydrogen-suppressed graphs 6, 260
Hydrophilic compounds 284, 285
Hydroprene 177, 179

I

IDF weighting 92, 95
Increasing complexity 50, 56, 58, 60, 65, 67
Induced subgraphs 125, 141, 142
Information content indices 228, 229
Information retrieval 92, 94, 110
Information theoretic indices 3, 7, 251, 258
Information-theoretic measures 139
Insect repellent compounds 308, 327
In silico pharmacophore models 303
Internal rotations 211, 212
Interrelated two-way clustering 251, 268, 273
Inverse frequency weighting 92, 96, 98, 102, 103, 107
Inverse frequency weighting experiments 99, 105
Ionization potential (IP) 203, 217, 232
Isobutane molecules 67
Isomeric molecules 144, 211
Isomers 29, 30, 31, 44, 64, 83, 114, 117, 123, 128, 312
Isopentane 7, 163, 166, 167
Isotopic spin 336, 340

J

JH receptor 318
Juvenile hormone active (JHAs) 177, 191

K

Kinetic energy 199, 200, 214, 218

L

Large number of descriptors 233, 235, 264
Lead optimization 9, 282, 284
Leave one out proper cross validation 3, 251
Leishmaniasis 303, 305, 319
Leptons 337, 338, 339
Level bonding pattern 286
Le Verrier-Fadeev-Frame method 76, 87
Ligand-based virtual screening 92
Linear discriminant analysis 173, 174, 176, 178, 191, 251, 269
Linear discriminant analysis (LDA) 173, 174, 176, 178, 180, 181, 182, 191, 251, 269

Linear extensions 33, 37
Linear free energy related (LFER) 4, 5, 254, 255
Lipid peroxidation (LP) 14, 15
Lowest unoccupied molecular orbital (LUMO) 207, 216, 217, 232, 262, 317, 318, 326

M

Magic-number molecules 345
Magnetizability 205, 212
Mass spectra 135, 137
Mathematical chemodescriptor 3
Mathematical way of thinking 24, 26, 27, 29, 31, 32, 34, 36
Matrices GedM 149, 151, 152, 154
Matrix polynomial computation 76
Maximum hardness principle (MHP) 196, 209, 211, 215, 217
MDDR 96, 97, 101, 103, 106, 110
MDDR activity classes 101, 102
MDL Drug Data Report/database 92
Methoprene 177, 179
Mgen 120, 121, 122, 125, 135, 136
Minimum electrophilicity principle 196, 211
Minimum magnetizability principle (MMP) 196, 209, 212
Minimum polarizability principle (MPP) 196, 209, 211, 215, 217
Model building 8, 255
Model development 223, 272
Model objects 6
Model quality 8, 15, 269, 272, 273, 274
Molar heat capacity 76, 84
Molar refractivity 283, 284, 289
Molecular Bioactivity 92, 96, 110
Molecular complexity 48, 52, 66, 274
Molecular connectivity (MC) 46, 52, 62, 161, 171
Molecular connectivity approach 46, 47, 52
Molecular connectivity indices 42, 61, 62, 63, 274
Molecular connectivity models 47, 61
Molecular cyclized degree (MCD) 286
Molecular descriptors 8, 10, 26, 34, 77, 139, 143, 161, 162, 163, 164, 237, 240, 255, 258
Molecular design of new structures 187
Molecular electronic properties 316, 317, 325
Molecular electronic structures 318
Molecular electrostatic potentials (MEPs) 303, 311, 314, 318
Molecular formulas 113, 114, 115, 116, 117, 121, 129, 131, 132, 133, 134, 135, 136, 137
Molecular frameworks 35, 36, 308
Molecular graph vertices 52, 165, 258
Molecular masses 132, 135
Molecular modeling study 312
Molecular models 323
Molecular periodicities 344, 346, 347
Molecular refraction 60, 61
Molecular refraction OC 62, 63
Molecular representation 34, 103
Molecular shapes 47, 166, 175
Molecular similarity 10, 50, 318
Molecular skeleton 30, 42
Molecular structural complexity 42, 49
Molecular structure 5, 6, 26, 34, 42, 48, 118, 122, 144, 161, 222, 223, 226, 233, 254
Molecular structure conundrum 6
Molecular structure elucidation 113, 114, 137
Molecular structure generation 137
Molecular structures form 223
Molecular subgraphs 46, 62
Molecular topology (MT) 42, 46, 48, 50, 52, 60, 161, 162, 163, 181, 182, 185, 189, 190, 191
Molecular vibrations 211
Molecule formation 210, 212
MOLGEN–MS 114, 137
Morgan 44, 45, 64
Multicriteria approaches 24
Multilinear regression analysis (MLRA) 175, 176, 179, 191
Mutagenicity 4, 24, 26, 36, 222, 226, 237, 238, 242, 243, 251, 272, 273

N

Naïve q2 3, 251
Naphthalene 120, 121, 123
Neurotoxicity 311, 312
Neutrinos 337, 338
New active compounds 180, 182, 308
New antimalarial agents 320, 324
New chemical classes 319
NHATS 288
Nitrobenzene 124, 125, 127, 128, 129, 168
Non-branched alkanes 76, 84
Non-drug databases 282, 283, 284, 289, 290, 291
Non–H atoms 116, 118, 119, 133, 134
Non-hydrogen atoms 6, 165, 167, 257, 261
Novel applications 3, 251
Novel compounds 303
Nuclear molecules 336, 343

O

Octane models 63, 64
Orbital energies 231, 232, 233, 314
Orbitals, molecular 231, 232
Ordering of substances 27, 29
Order neighborhood of vertices 260

Order theory 26, 27, 35, 37
Organic compounds 223, 254
Overall Zagreb indices 42, 51, 57, 61, 62

P

Pair of vertices 44, 172, 257, 259
Parasites 303, 304, 313, 316, 323
Partial least square (PLS) 3, 223, 235, 251, 273, 319
Partially ordered sets 24
Partial order 24, 27, 37, 140, 141, 142, 143, 144
Particular atoms 232
Pascal's triangle 336, 344
Patterns of increasing complexity 56, 58, 60, 65
Pc sub-graphs 168
Periodicities 336, 337, 340, 347
Periodic system of diatomic molecules 345
Periodic systems 336, 338, 340, 342, 343, 346, 347
Periodic systems of molecules 336, 340, 346
P. falciparum 304, 313, 322, 323, 325
Pharmacological distribution diagram (PDDs) 182, 183, 184, 185, 191
Pharmacophore models 303, 309, 315, 316, 318, 319, 320, 321, 322, 323, 324, 327
Phenacetin 287, 289
Physical organic chemistry 4, 254, 255
Physicochemical properties 5, 42, 43, 63, 64, 76, 89, 211, 226, 234, 236, 254, 257, 284, 299
Piperidines 126, 299
Plasma protein binding (PPB) 166, 191
P. malariae 304
Polarizability 196, 205, 211, 214, 215, 216
Polar Surface Area (PSA) 283, 285, 299, 300
POLONIUS 10
Posetic predictive methods 24, 31
P. ovale 304
Prediction of bioactivity 13
Predictions 4, 5, 8, 162, 167, 176, 236, 237, 242, 264, 265, 266, 267, 272, 275
Principal components regression (PCR) 3, 251, 273
Product-connectivity matrix 76, 77
Products of vertex degrees 46
Property-property relationships (PPRs) 5, 255
Protein kinase 98, 189, 192
Protein kinase C (PKC) 98, 189, 192
Proteomics 12, 14, 15, 17
Protons 215, 312, 339, 343
P. vivax 304
Pyridin-2 amine 241, 243, 244, 245, 246

Q

QC descriptors 233, 267, 269
QED 299
QSAR model building 267, 273
QSAR model development 7, 263
QSAR modeling 14, 77, 222, 223, 227, 234
QSAR models 42, 46, 222, 223, 224, 232, 234, 235, 236, 239, 263, 265, 266
Quantitative molecular similarity analysis 3, 11
Quantitative network analysis 139
Quantitative structure-activity relationships (QSAR) 4, 9, 10, 34, 43, 89, 155, 161, 162, 163, 217, 222, 235, 253, 263, 268, 300
Quantum 3, 4, 5, 8, 230, 231, 232, 233, 234, 258, 309, 310, 312, 313, 317
Quantum chemical (QC) 3, 5, 8, 164, 192, 231, 232, 233, 255, 258, 262, 267, 310, 312, 318
Quantum chemical descriptors 8, 231, 258
Quantum fluid density functional theory (QFDFT) 196, 213
Quantum fluid dynamics (QFD) 196, 213, 217
Quantum mechanics 197, 230, 231
Quaternary compounds, structure of 4, 254
Quinine 168, 188, 322
Quinolones 180, 181, 325

R

Randić index 166, 167
RDPi values 292, 293, 299
Reactive oxygen species (ROS) 14, 15
Reactivity descriptors 213, 216
Reference structure 92, 93, 94, 96, 97, 98, 100, 101, 105, 106, 107, 108, 109
Reference structure's fingerprints 100, 105
Refractive index 76, 84
Relative drug-likeness potential 282
Relative druglikeness Potential (RDPi) 292, 293, 298, 299
Relevant descriptors 223, 233, 234, 238
Relevant molecular descriptors 263
Restrictions on molecular graphs 123
Ridge regression (RR) 3, 15, 223, 235, 251, 270, 273
Rosmarinic acids 185, 186

S

SAR models 236, 237
See structure-activity relationship SAR
Selected structural descriptors 144, 287
Self organizing molecular field analysis 162, 192
Series MOLGEN 114
Shannon information content 139
Similarity coefficient 92, 108
Similarity matrix 139, 146, 149, 150, 155
Similarity measures 93, 105, 139, 143, 145, 155, 312
Similarity searching 92, 95, 100

Single bonds 116, 118, 119, 120, 123, 129, 133, 134, 172, 258, 298
Single molecular formulas 115, 136
Skeletal graph models 258
Skin sensitization 222, 237, 240, 242, 245
SMILES code of molecular structures 237
Soft molecules 205, 211, 212
Spectrum 4, 14, 15, 79, 81, 82, 87, 137, 233, 305, 337
Standard Gibbs energy of formation 76, 84
Standard models 337, 338, 340
Statistical models 265, 266, 275
Stereo-electronic properties 303, 307, 310, 316, 317, 318
Stereo-electronic properties of DEET 318
String manifestations 336, 339
Structural alerts 226, 238
Structural chemistry 6, 139, 140
Structural complexity 42, 139, 143, 144
Structural descriptors 4, 180, 182, 223, 225, 242, 282, 283, 286, 287, 288, 319
Structural diversity 108, 282, 287
Structural graph descriptors 140
Structural information 43, 147
Structural properties 147, 283
Structural similarity 139, 144, 145, 153
Structure-activity correlations 48, 60
Structure-activity relationship (SAR) 3, 4, 34, 43, 89, 161, 217, 222, 229, 251, 252, 253, 309, 316, 319, 325
Structure generation 113, 137
Structure generation techniques 137
Structure-property models 42
Structure-property relationship 76, 85, 89, 161, 217
Structures, three-dimensional 162, 223, 232, 307, 308
Sub-atomic particles 336
Subgraphs of increasing size 46, 47, 50, 64
Substituted ring atoms 121, 286
Substituted sites 30, 31
Substructural fragments 93, 187
Substructure restriction 113, 125
Sulfur atoms 314, 315, 324
Sum-connectivity matrix 77
Sunset keys 103, 105
Sunset molecular discovery LLC 96, 103
Symmetric square matrix 78, 79, 81, 82, 86

T

Tailored similarity 3, 11
Tanimoto coefficient 96, 98, 106, 107, 108, 109
Target proteins 306, 308, 320, 322, 328
TC descriptors 272
TC indices 258, 273, 274
Teratogen Information System (TERIS) 240
Terminal vertices 70
TF weighting 92, 95
Three-dimensional descriptors 231, 232
Three-dimensional molecular structure 232, 306
Time dependent (TD) 196, 213, 215, 218
Time dependent density functional theory (TDDFT) 196
Time dependent Schrödinger equation (TDSE) 213
Topochemical indices 164, 251
Topological charge indices (TCIs) 171, 172, 173, 192
Topological complexity 48, 51, 52, 56, 59
Topological descriptors 43, 47, 62, 63, 85, 144, 163, 164, 175, 176, 187, 235
Topological distance 165, 172, 173, 178, 227, 228, 230, 259
Topological indices 42, 43, 44, 50, 51, 54, 56, 61, 161, 187, 227, 228, 233, 258, 273
Topostructural indices 164, 251
Total adjacency 44, 51, 56, 62, 69
Toxicology 15, 17
Transitive orientation 142, 143
Triangular chord 141, 142
Triplet index 259, 260, 261
Tryptanthrins 321, 322, 323, 325, 326
TS descriptors 258, 267, 270, 272
Two-dimensional descriptors 227, 229
Two-edge subgraphs 47, 49, 51, 61

U

UALOGP 282, 288, 289, 293
Ulcerative colitis 182, 184, 185
Underlying graphs 145, 148

V

Valence connectivity indices 3, 251
Values of graph-invariants 52, 56
Vaporization enthalpy 76, 84
Version MOLGEN 113, 115
Vertex 45, 46, 47, 58, 64, 65, 66, 68, 69, 70, 145, 163, 228, 256, 257
Vertex centrality 42
Vertex chromatic number 141, 155
Vertex dd 341
Vertex degrees 42, 43, 44, 45, 51, 53, 78, 79, 80, 256, 260, 261
Vertex-labeled graph G1 78, 79
Vertex-product 76, 85
Vertex-product-connectivity index 76, 77, 83, 89
Vertex-product-connectivity matrix 76, 77, 78, 80, 89
Vertex-product-connectivity polynomial 87
Vertex set 140, 143, 146, 257
Vertex-sum-connectivity index 76, 77, 83, 89
Vertex-sum-connectivity matrix 76, 78, 81, 89

Virtual screening 92, 93, 94, 99, 107, 108, 182, 187, 188, 303, 307, 308, 309, 320, 323
Virtual screening of databases 307, 315, 316, 320, 326

W

Waals surface 311, 312, 314, 315, 318, 319
Weighted pseudographs 3, 6, 251, 257
Weighting schemes 92, 94, 95, 97, 98, 99, 101, 102, 104, 105, 106, 108
Wiener index 3, 7, 42, 51, 57, 59, 60, 68, 147, 156, 228, 259
Wiener number 44, 262
Wiener number values 56
WOMBAT activity classes 101