Chemometric Methods in Molecular Design

edited by Han van de Waterbeemd

VCH

Methods and Principles in Medicinal Chemistry

Edited by
R. Mannhold
P. Krogsgaard-Larsen
H. Timmerman

Chemometric Methods in Molecular Design

edited by Han van de Waterbeemd

VCH

Weinheim • New York
Basel • Cambridge • Tokyo

Volume editor:
Dr. Han van de Waterbeemd
F. Hoffmann - La Roche Ltd.
Pharma Research New Technologies
CH-4002 Basel
Switzerland

Editors:
Prof. Raimund Mannhold
Biomedical Research Center
Molecular Drug Research Group
Heinrich-Heine-Universität
Universitätsstraße 1
D-40225 Düsseldorf
Germany

Prof. Povl Krogsgaard-Larsen
Dept. of Organic Chemistry
Royal Danish School of Pharmacy
DK-2100 Copenhagen
Denmark

Prof. Hendrik Timmerman
Faculty of Chemistry
Dept. of Pharmacochemistry
Free University of Amsterdam
De Boelelaan 1083
NL-1081 HV Amsterdam
The Netherlands

This book was carefully produced. Nevertheless, authors, editors and publisher do not warrant the information contained therein to be free of errors. Readers are advised to keep in mind that statements, data, illustrations, procedural details or other items may inadvertently be inaccurate.

Published jointly by
VCH Verlagsgesellschaft mbH, Weinheim (Federal Republic of Germany)
VCH Publishers, Inc., New York, NY (USA)
Editorial Director: Dr. Thomas Mager
Production Manager: Dipl.-Wirt.-Ing. (FH) Hans-Jochen Schmitt

Library of Congress Card No. applied for.
British Library Cataloguing-in-Publication Data: A catalogue record for this book is available from the British Library.

Deutsche Bibliothek Cataloguing-in-Publication Data:
Chemometric methods in molecular design / ed. by Han van de Waterbeemd. - Weinheim ; New York ; Basel ; Cambridge ; Tokyo : VCH, 1995
(Methods and principles in medicinal chemistry ; Vol. 2)
ISBN 3-527-30044-9
NE: Waterbeemd, Han van de [Hrsg.]; GT

Printed on acid-free and chlorine-free paper.

Composition, Printing and Bookbinding: Druckhaus ,,Thomas Müntzer“ GmbH, D-99947 Bad Langensalza.
Printed in the Federal Republic of Germany.

Distribution:
VCH, P.O. Box 10 11 61, D-69451 Weinheim (Federal Republic of Germany)
Switzerland: VCH, P.O. Box, CH-4020 Basel (Switzerland)
United Kingdom and Ireland: VCH (UK) Ltd., 8 Wellington Court, Cambridge CB1 1HZ (England)
USA and Canada: VCH, 220 East 23rd Street, New York, NY 10010-4606 (USA)
Japan: VCH, Eikow Building, 10-9 Hongo 1-chome, Bunkyo-ku, Tokyo 113 (Japan)

Preface

One of the main objectives of the series *Methods and Principles in Medicinal Chemistry* is to provide practitioners and newcomers in the field with practice-oriented information on methodological aspects of QSAR.

After Hugo Kubinyi's volume on Hansch analysis and related approaches, the present handbook treats chemometric methods in molecular design. An introductory chapter by the volume editor, Han van de Waterbeemd, is followed by a section on molecular descriptors covering the classical physicochemical parameters, but also descriptors derived from solvent-accessible surface area and topological indices.

Experimental design in synthesis planning and structure-property correlations are the focus of the second section. Both methods for direct optimization of lead compounds as well as approaches for the systematic investigation of a parameter space are discussed. In continuation of this topic a strategy for QSAR development based on statistical experimental design and multivariate data analysis is outlined. Other topics in this section are optimization procedures in the case of non-linear structure-activity data applying topological descriptors, and an illustration how disjoint principal properties of organic substituents can be used for test series design.

In the central section of this volume the various approaches for multivariate data analysis are described; both established (principal component and factor analysis, SIMCA, PLS, discriminant analysis, cluster significance analysis) as well as new, emerging techniques (graphical analysis, spectral mapping, nonlinear mapping and canonical correlation analysis) are comprehensively described and exemplified in detail.

In the last section, leading experts treat a topic that has attracted increasing interest in recent time: the statistical validation of QSAR results. The first chapter describes the main tools for assessing the validity of QSAR models, and the second chapter gives the rules for choosing the proper statistical method for model validation.

Taken together, the chapters in this volume give a practice-oriented introduction to the continuously developing field of chemometrics in molecular design and provides the reader with recipes for a proper application of these tools.

Düsseldorf
Kopenhagen
Amsterdam

Raimund Mannhold
Povl Krogsgaard-Larsen
Hendrik Timmerman
Winter 1994

A Personal Foreword

It was at the first Noordwijkerhout Symposium in 1977 that I first came into contact with applications of quantitative structure-activity relationships. It was Corwin Hansch, Roelof Rekker and Hugo Kubinyi who inspired me at the time in those early days of my career to continue in the rather new field of QSAR. Much as I learned during my "postdoc" with Bernard Testa at the University of Lausanne, the real challenge of using chemometric methods in molecular design and discovery really only began in 1988 when I moved to industry. Indeed, in this environment one sees what can be best achieved with these methods and how they are used, misused, or not used at all by medicinal chemists. I would like to thank Klaus Müller for creating a stimulating environment and for having given me the freedom to work on this book.

When I was approached about writing this book, I felt immediately that this could not be the task of one person only. As can be seen in the diverse chapters in this volume, there are many methods which can be used in the process of molecular design, requiring the expertise and experience of other researchers. Therefore, I would like to warmly thank all the contributors to this volume.

The scope of this series is to offer many practical examples of interest to the medicinal chemist. In this volume, we have collected a large variety of different techniques. Several of these have reached some degree of maturity and many examples can be found in the literature as well as in the present volume. However, certain other methods, described here, are rather new and still under development. For these approaches, we have chosen to include some more details on the algorithms and their preliminary evaluations. As a medicinal chemist, we hope that you appreciate the efforts of the data analysis "experts" in developing new methods for extracting information or "mining" data from complex and incomplete biological and chemical data. We trust that this volume shows that the present chemometric methods used in structure-property correlations, and in particular, in QSAR studies, go far beyond the classical Hansch approach. It is hoped that this book fulfills two missions. First it should offer the medicinal chemist an insight into the diversity of multivariate chemometric methods and their applications in the design of bioactive molecules. Secondly, for the specialist it should provide an update of current and newly emerging techniques.

This book is a rather peculiar kind of Christmas present to me. I really hope that you, as medicinal chemist or any other kind of researcher involved in molecular design, will find this book a valuable gift too and will enjoy playing with these chemometric approaches to the benefit of your daily work.

Finally, I would like to thank my wife Kitty and daughter Marion for sharing their enthusiasm with me and for bringing me sufficient coffee, when I was proofreading all the chapters during many long autumn evenings. It was a great experience!

December 1994,
Basel

Han van de Waterbeemd

List of Contributors

Dr. Volkhard Austel
Dr. Karl Thomae GmbH
Chemical Department
Birkendorfer Straße 65
Postfach 1755
D-88397 Biberach an der Riss 1,
Germany
Tel.: +49 7351 54 2974
Fax: +49 7351 54 2165

Dr. Daniel Chessel
Université Lyon I
URA CNRS 1451
F-69622 Villeurbanne, France

Prof. Sergio Clementi
Laboratorio di Chemiometria
Dipartimento di Chimica
Università di Perugia
Via Elce di Sotto 8
I-06123 Perugia, Italy
Tel. and Fax: +39 75 45646
gabri@chemiome.chm.unipg.it

Dr. Gabriele Costantino
Istituto di Chimica Farmaceutica
e Tecnica Farmaceutica
Università di Perugia
Via del Liceo 1
I-06123 Perugia, Italy
Tel.: +39 75 58 55131
Fax: +39 75 58 55124

Dr. Gabriele Cruciani
Laboratorio di Chemiometria
Dipartimento di Chimica
Università di Perugia
Via Elce di Sotto 8
I-06123 Perugia, Italy
Tel. and Fax: +39 75 45646

Dr. James Devillers
Centre de Traitement de
l'Information Scientifique (CTIS)
21, rue de la Bannière
F-69003 Lyon, France
Tel.: +33 78 628499
Fax: +33 78 629912

Dr. Steven Dixon
PennState University
Department of Chemistry,
College of Science
152 Davey Laboratory
University Park, PA 16802, USA
Tel.: +1 814 865-3739
Fax: +1 814 865-3314

Prof. William J. Dunn III
Department of Medicinal Chemistry
and Pharmacognosy
University of Illinois at Chicago
833 S. Wood, m/c 781
Chicago, Illinois 60680, USA
Tel.: +1 312 996 5232
Fax: +1 312 996 3272

Dr. Leanne Egolf
PennState University
Department of Chemistry
College of Science
152 Davey Laboratory
University Park, PA 16802, USA
Tel.: +1 814 865 3739
Fax: +1 814 865 3314

Dr. Lennart Eriksson
Department of Chemistry
Research Group for Chemometrics
University of Umea
S-90187 Umea, Sweden
Tel.: +46 90 165119
Fax: +46 90 138885

Prof. Dr. Martyn G. Ford
University of Portsmouth
School of Biological Sciences
King Henry Building
King Henry I Street,
Portsmouth, Hants PO1 2DY, UK
Tel.: +44 705 842036
Fax: +44 705 842070

Dr. Rainer Franke
Consulting in Drug Design
Gartenstraße 14
D-16352 Basdorf, Germany
Tel. and Fax: +49 33397 62396

Daniel J. Gans
Bristol-Myers Squibb
Princeton, USA

Dr. Andreas Gruska
Consulting in Drug Design
Gartenstraße 14
D-16352 Basdorf, Germany
Tel. and Fax: +49 33397 62396

Dr. Lowell H. Hall
Eastern Nazarene College
Department of Chemistry
23 East Elm Avenue
Quincy
Massachusetts 02170, USA
Tel.: +1 617 773-8350
Fax: +1 617 773-4833

Prof. Peter Jurs
PennState University
Department of Chemistry
College of Science
152 Davey Laboratory
University Park, PA 16802, USA
Tel.: +1 814 865 3739
Fax: +1 814 865 3314
pcj@psuvm.psu.edu

Prof. Lemont B. Kier
Virginia Commonwealth University
Department of Medicinal Chemistry
Richmond, VA 23298, USA
Tel.: +1 804 786 8488
Fax: +1 804 225 3299
kier@vcuvax

Dr. Paul J. Lewi
Information Science Dept.
Janssen Research Foundation
Janssen Pharmaceutica NV
Turnhoutseweg 30
B-2340 Beerse, Belgium
Tel.: +32 14 602 111
Fax: +32 14 602 841

Prof. Dr. Raimund Mannhold
Biomedical Research Center
Molecular Drug Research Group
Universitätsstraße 1
D-40225 Düsseldorf, Germany
Tel.: +49 211 311 2759
Fax: +49 211 312 631

Dr. Jim W. McFarland
Pfizer Inc.
Central Research Division
Eastern Point Road
Groton, CT 06340, USA
Tel.: +1 203 441 3692
Fax: +1 203 441 4111
mcfarland@pfizer.com

Dr. David W. Salt
University of Portsmouth
School of Biological Sciences
King Henry Building
King Henry I Street
Portsmouth, PO1 2DY, UK
Tel.: +44 705 84 20 36
Fax: +44 705 84 20 70

Prof. Dr. Michael Sjöström
Department of Chemistry
Research Group for Chemometrics
University of Umea
S-90187 Umea, Sweden
Tel.: +46 90 165119
Fax: +46 90 138885
sjoe@biovax.umdc.umu.se

Dr. Roberta Valigi
Laboratoria di Chemiometria
Dipartimento di Chimica
Università di Perugia
Via Elce di Sotto 8
I-06123 Perugia, Italy
Tel. and Fax: +39 75 45646

Dr. Han van de Waterbeemd
F. Hoffmann-La Roche Ltd.
Pharma Research New Technologies
CH-4002 Basel, Switzerland
Tel.: +41 61 688 8421
Fax: +41 61 688 1075
johannes.van_de_waterbeemd@
roche.com

Prof. Dr. Svante Wold
MDS Inc.
371 Highland Ave
Winchester, MA 01890, USA
Tel.: +1 617 7295446
Fax: +1 617 7212652
and
Umea University
Department of Chemistry
Research Group for Chemometrics
S-90187 Umea, Sweden
Tel.: +46 90 165358
Fax: +46 935 44039 and
+46 90 138885

Contents

Chemometric Methods in Molecular Design
edited by Han van de Waterbeemd
© VCH Verlagsgesellschaft mbH, 1995

1 Introduction

Han van de Waterbeemd

Abbreviations

ALS	Adaptive least squares
ANN	Artificial neural networks
CA	Cluster analysis
CCA	Canonical correlation analysis
CR	Continuum regression
CFA	Correspondence factor analysis
CSA	Cluster significance analysis
2D	Two-dimensional
3D	Three-dimensional
EFMC	European Federation for Medicinal Chemistry
FA	Factor analysis
FB	Fujita-Ban analysis
FW	Free-Wilson analysis
GOLPE	Generating optimal PLS estimations
kNN	*k*-nearest neighbor
LDA	Linear discriminant analysis
LMM	Linear learning machines
MLR	Multiple linear regression
NLM	Non-linear mapping
OLS	Ordinary least squares
PC	Personal computer
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares
QSAR	Quantitative structure-activity relationships
SCD	Single class discrimination
SIMCA	Soft independent modeling of class analogy
SMA	Spectral mapping analysis
SPC	Structure-property correlations

Symbols

$\log 1/C$	Biological activity
E_s	Taft steric parameter
σ	Hammett constant
β	Regression parameter in bilinear equation

1.1 Quantitative Molecular Design

The discovery of biologically active compounds and their development as drugs is a highly complex process which involves many scientific disciplines [1]. Medicinal chemists have for a long time systematically modified a lead compound with the main driving force being synthetic feasibility, experience, intuition and serendipity. Over the last 25 years molecular design strategies have changed considerably [2, 3]. Important contributions to the design of new compounds today come from biostructural research, including protein crystallography, multidimensional bio-NMR and molecular modeling.

Corwin Hansch and co-workers [4 – 6] deserved the success of having propagated the use of physico-chemical properties and statistical methods in structure-activity relationship studies. A general formula for a quantitative structure-activity relationship (QSAR) can be given by the following:

$$\text{activity} = f\ (\text{molecular or fragmental properties}) \qquad (1)$$

The original work of Hansch and co-workers involved linear combinations of suitable descriptors, using multiple linear regression to obtain the now well-known QSAR equations. The Hansch method will be discussed below briefly. For a more detailed discussion, see the first volume of the present series [7]. In order to be able to deal with complex data sets, consisting of more than one biological activity and many (physico-)chemical descriptors, more advanced statistical tools have had to be considered and developed. This is the field of chemometrics, and QSAR, an important branch of chemometrics, is the main focus of this volume.

In drug research today, for some people the QSAR approach is taken to be equivalent to using Hansch-type regression equations, while for others, it includes

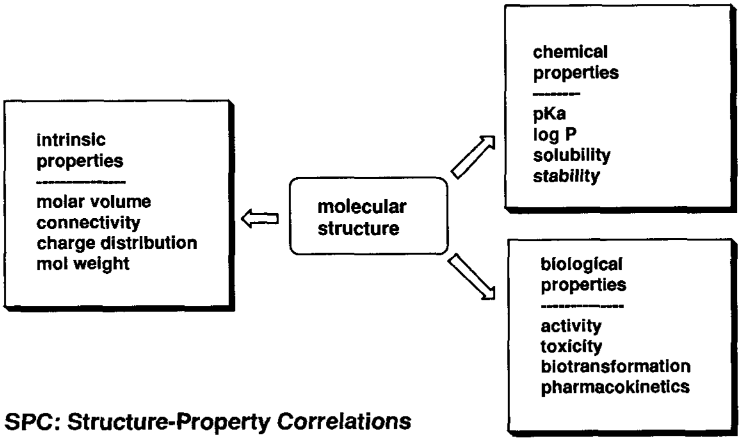


Figure 1. The concept of structure-property correlations (reproduced from Fig.1 of Ref. [2] with permission from the copyright owner).

any statistical mathematical technique which is employed to unravel information obtained from the available biological and chemical data. Therefore, attempts have been made to introduce other terms in order to avoid this confusion [2, 8]. We propose to call all studies which are aimed at broadening the understanding of relationships between intrinsic molecular, chemical and biological properties, as structure-property correlation (SPC) studies [2] (see Fig. 1). QSAR, thus, comes under the generic term of SPC studies. Another source of misunderstanding is the use of terms such as "rational drug design". A drug is a product on the market, which is used in health-care. Such a product is developed from a bioactive molecule, which has been selected and clinically tested for this purpose. Many other biologically active compounds appear not to be suited as a drug due to toxicity or unfavorable side-reactions, or as a result of unfavorable pharmacokinetics. Therefore, we should strictly speaking refer to molecular design. It should also be pointed out that earlier generations of scientists have always conducted research rationally, thus, rendering the word "rational" in the term rational drug design meaningless. The approaches discussed in the present volume should be regarded as computer-assisted molecular design or computer-assisted medicinal chemistry. Since chemistry is an experimental science, these *in computro* methods are only successful under certain conditions. Such conditions require that the biological activity of a series of compounds is, indeed, related to the chemical properties being considered, and that the series is more or less congeneric. That is all the compounds act by the same biological mechanism, e.g. with a similar binding mode at the active site of a given biological target. Medicinal chemists often face the problem of not knowing the 3D structure of their therapeutic target. Both molecular modeling techniques and quantitative statistical methods may then be useful in elucidating structural information at the active site. Molecular modeling provides methods, such as the active analog approach or constrained search [9], to define pharmacophores or the geometry at the active site. The methods discussed in the present volume are either complementary to molecular modelling approaches, or may themselves provide clues about which parts of the molecule are important for activity as well as for inactivity.

1.2 Chemometrics

The term chemometrics was coined in the 1970s and is defined as *the* chemical discipline that uses statistical and mathematical methods for selecting and optimizing analytical and preparative methods, as well as procedures for the analysis and interpretation of data. Chemometrics has found wide application in analytical chemistry [10–15]. Two specific journals are devoted to the development and applications of chemometrics, namely *Chemometrics and Intelligent Laboratory Systems* (1986) and *The Journal of Chemometrics* (1987). A series of books on *Chemometrics* has been started recently [42]. Using essentially the same techniques, medicinal chemists and specialist "drug designers" have further developed the field of quantitative structure-activity relationships (QSAR) [16–20]. Various

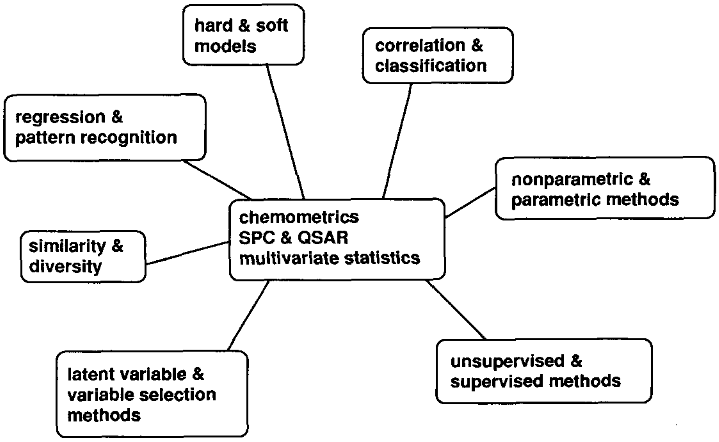

hard & soft models
correlation & classification
regression & pattern recognition
nonparametric & parametric methods
chemometrics SPC & QSAR multivariate statistics
similarity & diversity
latent variable & variable selection methods
unsupervised & supervised methods

Figure 2. Terms used in modeling data.

statistical methods are known under different names, which is certainly confusing for the non-specialist. Terms such as multivariate data analysis, chemometrics, pattern recognition, parametric and non-parametric statistics, regression, latent variable techniques, QSAR and projection methods are often used without definition. In addition different authors may have a different understanding of these terms [21–24].

As already discussed above, a QSAR equation is a correlation between biological and chemical data obtained by Multiple Linear Regression (MLR), sometimes also called Ordinary Least Squares (OLS). MLR is referred to as a variable selection technique [25], while latent variable techniques are techniques, such as Principal Component Regression (PCR) and Partial Least Squares (PLS). MLR is regarded as a "hard" model, whereas SIMCA (Soft Independent Modeling of Class Analogy) and PLS are called "soft" modeling techniques [26].

Two further categories are "supervised" and "unsupervised" methods. Multiple linear regression and backpropagation artificial neural networks are supervised methods, in which a model is fitted to the data, while cluster analysis, principal component analysis and non-linear mapping, for example, are unsupervised, and classification patterns are obtained. One should also distinguish between the quantitative predictions obtained with MLR, and the qualitative predictions obtained with pattern recognition techniques, such as cluster analysis and principal component analysis. Non-parametric statistics, such as Adaptive Least Squares (ALS), do not require a normal distribution of the data.

An important part of each multivariate data analysis is the selection of an appropriate training (modeling or calibration) and test (validation) set. Without a careful selection of the training set, any derived model makes little sense. However, much experience is needed to be able to construct training and test sets which will be of some significance [27].

Regression / Correlation

MLR
FW, FB
LDA, ALS
PLS, GOLPE
PCR
CCA
CR

Pattern Recognition / Classification

PCA, FA, NLM, CFA, SMA
SIMCA
LDA, LLM, KNN, SCD
CA, CSA
ANN

Figure 3. Overview of SPC methods. ALS = Adaptive Least Squares; ANN = Artificial Neural Networks; CA = Cluster Analysis; CCA = Canonical Correlation Analysis; CFA = Correspondence Factor Analysis; CR = Continuum Regression; CSA = Cluster Significance Analysis; FA = Factor Analysis; FB = Fujita-Ban analysis; FW = Free-Wilson analysis; GOLPE = Generating Optimal Linear PLS Estimations; kNN = k-Nearest Neighbor; LDA = Linear Discriminant Analysis; LLM = Linear Learning Machine; MLR = Multiple Linear Regression; NLM = Non-Linear Mapping; PCA = Principal Component Analysis; PCR = Principal Component Regression; PLS = Partial Least Squares of Projected Latent Structures; SCD = Single Class Discrimination; SIMCA = Soft Independent Modeling of Class Analogy (Similarity, Chemistry and Analogy); SMA = Spectral Map Analysis.

1.3 The Hansch Approach

In the early 1960s, Hansch and co-workers systematically investigated ways of expressing the relationships between structural and physico-chemical properties and activities quantitatively. The traditional QSAR paradigm is often formulated as shown in Eq. (1). More recently, due to the confusion surrounding the term QSAR, Hansch has referred to the science of chemical ↔ biological interactions as the "unnamed science" [28]. Since it is evident that a biological effect seldom depends on just one factor, methods have been explored to investigate this multidimensional problem. The first volume of the present series explains the Hansch approach and related techniques in much more detail [7]. It should be emphasized that Hansch analysis is a method, in which the factors which influence biological activity are rationalized, and should not be considered too much as a predictive method, since usually only a limited parameter space is covered.

The first QSAR equations were based on the observation that partition coefficients, as expressed by log P values, are to some extent, correlated to certain biological endpoints. In most cases, this relationship appears not to be linear, but displays an optimum value. The parabolic model of Fujita-Hansch [4] (Eq. (3)) and the bilinear model of Kubinyi [7] (Eq. (4)) describe this empirical observation:

$$\log 1/C = a(\log P) + b \tag{2}$$

$$\log 1/C = a(\log P)^2 + b(\log P) + c \tag{3}$$

$$\log 1/C = a(\log P) - b(\log (\beta P + 1)) + c \tag{4}$$

where C is the molar concentration that produces a certain effect, P is often the 1-octanol/water partition coefficient, and a, b, c and β are regression coefficients. The bilinear model of Kubinyi and the parabolic Hansch model are related and may be derived from the partitioning of simple two- or three-phase solvent systems [29]. The Hansch model is most applicable to complex *in vivo* systems, where a drug has several barriers to cross to reach its target. In less complex systems, e.g. cell cultures, in which only a few membranes must be crossed, the bilinear model best fits the data. Drug transport and distribution is one of the main reasons for the appearance of a lipophilicity descriptor in many SPC studies. The interaction of a ligand to its active site involves different kinds of bonding: H-bonding, ionic forces, van der Waals or hydrophobic, as well as dipole-dipole interactions. These may be parametrized to some extend in a QSAR expression. The so-called Hansch equation (Eq. (5)) takes into account these effects [99]:

$$\log 1/C = a(\log P)^2 + b(\log P) + cE_s + d\sigma + e \tag{5}$$

where E_s is Taft's steric descriptor and σ the well-known Hammett constant, reflecting electronic contributions. Over the years many different molecular and fragmental descriptors have been used in these extrathermodynamic or linear free-energy relationships (LFER) [30]. The traditional method for calculating a quantitative model in a Hansch analysis study is by multiple linear regression (MLR). The most frequently encountered difficulties with multiple linear regression have been discussed fully in Vol. 1 of this series [7]. However, to obtain suitable equations the following are important:

- a ratio of compounds to variables greater than five,
- a minimal intercorrelation among the variables in the final equation.

The quality of a MLR can also be judged by looking at the standard error of the regression coefficients. Some regression programs produce standard deviations, while others give 95% confidence intervals. One should be aware that the latter are out by about a factor of two. Another often misused statistical criterion when comparing two equations, is the correlation coefficient. A statement such as "Equation A ($r = 0.956$) is better than Equation B ($r = 0.918$)" should be treated with caution and a sequential or partial F-test should be performed to justify statements of this kind.

Another pitfall is the use of regression coefficients to discuss the relative contribution of a descriptor to the measured activities. This can only be done after normalizing the equation, i.e. eliminating the constant term [31].

A modern alternative to MLR is partial least squares regression in latent variables (PLS) in combination with cross-validation (see Chaps. 4.4, 5.1 and 5.2)

[32, 100]. Although this method is believed to be very robust, some difficulties should not be overlooked [7, 33, and Chap. 5.2]. In this relatively new statistical method, latent variables or components are extracted from the descriptor variables, which have predictive capability for dependent variables. PLS works for smaller data sets with many descriptors and can treat a set of multivariate biological activities. New faster algorithms have been developed for larger data sets [34], which have, however, been criticized (see Chap. 5.2). PLS is an important component in 3D QSAR or comparative molecular field analysis (CoMFA) [35] (see Vol. 3 [97]). PLS is also widely used to solve the problems of analytical calibration and for optimization in organic synthesis [36]. Interesting alternatives to cross-validation have also been considered [37].

Further alternatives to deriving Hansch-type QSARs are techniques such as principal component regression (PCR) (see Chap. 4.1) and stochastic regression analysis [38].

1.4 Modern Chemometric Approaches in Molecular Design

Biological activities seldom depend on just one or two chemical properties, and subsequently, a complex matrix of data must often be analyzed. Biological data can vary from just mere simple affinity data (IC_{50} values) to complex *in vivo* data, reflecting only the activity or inactivity of a compound. The selection of the appropriate method for handling such data is extremely important if any useful conclusive results are to be obtained.

The present volume first describes molecular concepts and the most important descriptors. More information on chemical descriptors can be found in the series "Methods and Principles in Medicinal Chemistry" (1993) [7]. Every good chemistry experiment, including the synthesis of biologically active compounds, should begin with a good experimental design. The design of a series of compounds is based on synthetic feasibility, chemical intuition, time and availability of chemicals. A number of strategies have been described to make more rational choices in synthesis planning [41–43]. These are presented in Sect. 3.

The remainder of this volume describes methods that analyze biological and chemical data, either separately or the correlations between them (see Fig. 4). Based on the methods already developed, new compounds may be designed, or insight obtained into molecular mechanisms. Therefore, the validation of such methods (Chap. 5.1) and the choice of appropriate methods (Chap. 5.2) are important subjects to discuss.

In drug research today, many disciplines are working closely together. Computer-assisted data handling, including operations such as data retrieval from 2D and 3D chemical databases, pharmacophore generation, molecular modeling, and structure-property correlations (quantitative structure–activity relationships) have be-

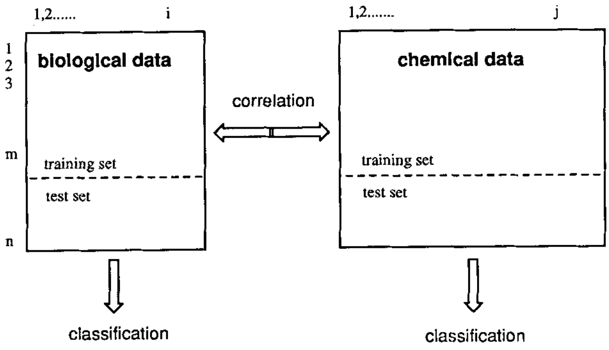


Figure 4. Biological and chemical data tables. Part of the data is used to build a model (for one or more classes) and another part a test set. Biological and chemical data can be used separately or in conjunction to classify compounds. Correlations between biological and chemical data can also be ascertained.

come an integral part of the work of the medicinal chemist [44–46]. The present book gives an overview of some of the current methods and illustrates how modern chemometric techniques can be used in the design of biologically active new chemical entities.

In forthcoming volumes of this series, other computer-assisted medicinal chemistry techniques, such as molecular modeling and structure-based design, will be covered. The increasing importance of 3D data handling and its use in establishing 3D QSAR, is presented in the next volume of this series [97, 35]. With the advent of combinatorial chemistry to improve molcular diversity and the chances of lead discoveries, these methods will become all the more important. Hence, concepts of molecular similarity and dissimilarity [39, 40] are also dealt with in Vol. 3 of the present series "Methods and Principles in Medicinal Chemistry" (1994) [97].

1.5 Software

1.5.1 General Statistical Packages

One of the most difficult tasks for the medicinal chemist is the interpretation of biological test results and how the rest results correlate with the chemical data. The choice of appropriate software tools to achieve this, is a prerequisite for extracting all the available information from the data. Although seemingly trivial, simple 2D scatter plots of either biological or chemical data are still highly informative, as was recently illustrated in an optimization study of antibacterial agents [47]. Similarly, plots of biological data against any of the collated chemical descriptors are most useful, particularly when a color-coding can be employed. Many PC-based programs are in fact suitable for this, and we would like to mention SYSTAT [95]. SPSS/PC^{+}

Table 1. Statistical packages for structure – property correlation studies

Program	Hardware	Proprictor
BMDP	Mainframe	BMDP [56]
BMDP New System	PC	BMDP [56]
EXCEL	PC, Macintosh	Microsoft
GENSTAT	Mainframe, workstation, PC	Numerical Algorithms Group [63]
GRAFTOOL	PC	3-D Visions [65]
JMP	Macintosh	SAS Institute [67, 84]
MACSPIN	Macintosh	D^2-Software [68]
MINITAB		
MULTISTAT	Macintosh	Biosoft [71]
NCSS		
PARVUS	PC	Elsevier [73]
P-STAT		
QUATTRO-PRO	PC	Borland
RS/1	VAX, workstation, PC	BBN [83]
SAS	Mainframe, workstation, PC	SAS Insitute [84]
SIGMAPLOT	PC	Jandel Scientific [85]
SPSS	Mainframe, workstation	SPSS [87]
SPSS/PC$^+$	PC	SPSS [87]
STATA		
STATISTICA	DOS, Windos, Macintosh	StatSoft [89]
STATGRAPHICS	PC	
SYSTAT	Windows, Macintosh	Systat [95]

[87] and STATGRAPHICS [92]. One special feature of some these programs is real-time rotation of 3D plots, using e.g. three independent variables or three components from a principal components analysis. This is available in, e.g. MACSpin [68], JMP [67] and SYSTAT [95]. Table 1 gives a selection of the available statistical data modeling packages. A further selection can be found in the literature [48, 98]. In most SPC and QSAR studies, the first step in looking at data using statistical approaches involves traditional Hansch analysis using multiple linear regression (MLR), or modern partial least squares (PLS) modeling. Multiple regression is available in any statistical package, but unfortunately this is not the case for PLS modeling. Some packages include a programming language which can be used to write macros that can perform the operations required for PLS analysis. The dangers of using incorrectly programmed PLS and cross-validation algorithms are discussed in Chap. 5.2. However, a more specialized and validated software program (see below) is preferred in most cases.

1.5.2 Specialized Software for SPC Studies

The statistical packages discussed previously have been developed for general-purpose statistics and, course, are very useful for most of the analyzes described in this book. However, since many of our chemists are neither trained in statistics, nor in the use of sophisticated statistical packages, it is advizable to have a look

Table 2. Specialized SPC software

Program	Hardware	Proprietor
ADAPT	Vax	Prof. P. C. Jurs [51]
APEX	Workstation	Biosym [52]
ARTHUR	Mainframe	Informetrix [53]
ASP	Workstation	OML [54]
CATALYST	Workstation	BioCad [57]
CERIUS^2	Workstation	MSI [76]
CHEMX	Workstation	CDL [58]
CLOGP	Vax	BioByte [59]
CLUSTAN	Vax	[60]
C-QSAR	Vax	BioByte [61]
GOLPE	Unix	Tripos & MIA [64]
HYPERCHEM	PC, workstation	Autodesk [66]
MOLCONN-X	Vax	Hall Associates Consulting [70]
PCMODELS	Vax, workstation	Daylight CIS [74]
PIROUETTE	PC	Informetrix [75]
POLARIS	Workstation	Molecular Simulations [76]
PROLOGP	PC	Compudrug Chemistry [77]
QSAR	Vax	BioByte [80]
QSAR-PC	PC	Biosoft [81]
RECEPTOR	Workstation	MSI [82]
SIMCA	Vax, PC, workstation	Umetri [86]
SYBYL-CoMFA	Workstation	Tripos [93]
SYBYL-QSAR	Workstation	Tripos [94]
TSAR	Workstation	OML [91]
UNSCRAMBLER	PC	CAMO [96]

at some of the software products, which are more specialized in molecular design. A list of these products, which are currently on the market, is given in Table 2. Reviews of new products appear twice a year in the Newsletter of the EFMC (European Federation for Medicinal Chemistry) [49] and the International QSAR Society [50]. In particular, those products which offer molecular display, statistical and graphical tools, such as TSAR [91], are potentially very useful and would be of considerable interest to any medicinal chemist. Since these products are produced by rather small companies, most of them still give rise to problems as regards to their conceptual basis and implementation, and care must be taken when using them. Various other programs are mentioned in the present volume under specific topics, as well as in Vol. 3 of the present series "Methods and Principles in Medicinal Chemistry" (1994) [97].

References

[1] Spilker, B., *Multinational Drug Companies. Issues in Drug Discovery and Development*, Raven Press, New York, 1989

[2] van de Waterbeemd, H., *Quant. Struct.-Act. Relat.* **11**, 200–204 (1992)

[3] van de Waterbeemd, H., *Drug Des. Disc.* **9**, 277–285 (1993)

[4] Hansch, C., Maloney, P. P., Fujita, T., and Muir, R. M., *Nature* **194**, 178–180 (1962)

[5] Purcell, W. P., Bass, G. E., and Clayton, J. M., *Strategy in Drug Design: A Guide to Biological Activity*, Wiley, New York, 1973

[6] Gould, R. F., ed., *Biological Correlations – The Hansch Approach*, (Adv. Chem. Series 114, American Chemical Society, Washington, 1972

[7] Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches. Methods and Principles in Medicinal Chemistry*, Vol. **1**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1993

[8] Testa, B., and Kier, L. B., *Med. Res. Revs.* **11**, 35–48 (1991)

[9] Dammkoehler, R. A., Karasek, S. F., Shands, E. F. B., and Marshall, G. R., *J. Comp.-Aided Mol. Des.* **3**, 3–21 (1989)

[10] Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., and Kaufman, L., *Chemometrics: A Textbook*. Elsevier, Amsterdam, 1988

[11] Delaney, M. F., *Anal. Chem.* **56**, 261R–277R (1984)

[12] Brown, S. D., Barker, T. Q., Larivee, R. J., Monfre, S. L., and Wilk, H. R., *Anal. Chem.* **60**, 252R–273R (1988)

[13] Brown, S. D., *Anal. Chem.* **62**, 84R–101R (1990)

[14] Berridge, J. C., *Anal. Chim. Acta* **223**, 149–159 (1989)

[15] Kateman, G., *Anal. Chim. Acta* **191**, 125–131 (1986)

[16] Topliss, J. G., ed., *Quantitative Structure-Activity Relationships of Drugs*, Academic Press, New York, 1983

[17] Stuper, A. J., Brügger, W. E., and Jurs, P. C., *Computer-Assisted Studies of Chemical Structure and Biological Function*, John Wiley, New York, 1979

[18] Martin, Y. C., *Quantitative Drug Design. A Critical Introduction*, Marcel Dekker, New York, 1978

[19] Devillers, J., and Karcher, W., ed., *Applied Multivariate Analysis in SAR and Environmental Studies*, Kluwer Academic Publishers, Dordrecht, 1991

[20] Tute, M. S., History and Objectives of Quantitative Drug Design. In: *Quantitative Drug Design* (Comprehensive Medicinal Chemistry, Vol. IV), Hansch, C., Sammes, P. G., and Taylor, J. B., eds., Pergamon Press, Oxford (1990) p. 1–31

[21] Livingstone, D. J., Quantitative Structure-Activity Relationships. In: *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*, Zalewski, R. I., Krygowski, T. M., Shorter, J., eds., Elsevier, Amsterdam (1991) p. 557–627

[22] Livingstone, D. J., *Pattern Recognition Methods in Rational Drug Design*. In: *Molecular Design and Modelling, Concepts and Applications. Methods in Enzymology*, Vol. **203**. Langone, J. J., ed., Academic Press, London (1991) p. 613–638

[23] Wold, S., and Dunn, W. J., *J. Chem. Inf. Comput. Sci.* **23**, 6–13 (1983)

[24] Dunn, W. J., and Wold, S., *Pattern Recognition Techniques in Drug Design*. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Hansch, C., Sammes, P. G., and Taylor, J. B., eds., Pergamon Press, Oxford (1990) p. 691–714

[25] Salt, D. W., Yildiz, N., Livingstone, D. J., and Tinsley, C. J., *Pestic. Sci.* **36**, 161–170 (1992)

[26] Miyashita, Y., Li, Z., and Sasaki, S., *Trends Anal. Chem.* **12**, 50–60 (1993)

[27] World, S., and Dunn, W. J., *J. Chem. Inf. Comput. Sci.* **23**, 6–13 (1983)

[28] Hansch, C., *Acc. Chem. Res.* **26**, 147–153 (1993)

[29] van de Waterbeemd, J. Th. M., and Jansen, A. C. A., *Pharm. Weekbl. Sci. Ed.* **3**, 71–78 (1981)

[30] van de Waterbeemd, H., and Testa, B., *Adv. Drug Res.* **16**, 85–225 (1987)

[31] Mager, H., and Barth, A., *Pharmazie* **34**, 557–559 (1979)

[32] Wold, S., Johansson, E., and Cocchi, M., *PLS – Partial Least Squares Projections To Latent Structures*. In: *3D QSAR in Drug Design. Theory, Methods and Applications*. Kubinyi, H., ed., Escom, Leiden (1993) p. 523–550

[33] Kubinyi, H., and Abraham, U., *Practical problems in PLS Analyses*. In: *3D QSAR in Drug Design. Theory, Methods and Applications*, Kubinyi, H., ed., Escom, Leiden (1993) p. 717–728

[34] Bush, B. L., and Nachbar, R. B., *J. Comput.-Aided Mol. Design* **7**, 587–619 (1993)

[35] Kubinyi, H. ed., *3D QSAR in Drug Design. Theory, Methods and Applications*, Escom, Leiden, 1993

[36] Carlson, R., and Nordall, A., *Topics Curr. Chem.* **166**, 1–64 (1993)

[37] Klopman, G., and Kalos, A. N., *J. Comput. Chem.* **6**, 492–506 (1985)

[38] Devillers, J., Zakaraya, S., Chastrette, M., and Doré, J. C., *Biomed. Environm. Sci.* **2**, 385–393 (1989)

[39] Zalewski, R. I., Krygiwsky, T. M., and Shorterm, J., eds., *Similarity Models in Organic Chemistry, Biochemistry and Related Fields*, Elsevier, Amsterdam, 1991

[40] Johnson, M. A., and Maggiora, G. M., eds., *Concepts and Applications of Molecular Similarity*, Wiley, New York, 1990

[41] Deming, S. N., and Morgan, S. L., *Experimental Design: A Chemometric Approach*, Elsevier, Amsterdam, 1993

[42] Walters, F. H., Parker, L. R., Morgan, S. L., and Deming, S. N., *Sequential Simplex Optimization: A Technique for Improving Quality and Productivity in Research, Development and Manufacturing*, CRF Press Inc., Boca Raton, 1991

[43] Mager, P. P., *Design Statistics in Pharmacochemistry*, Research Studies Press & John Wiley, Chichester, 1991

[44] Wermuth, C. G., Koga, N., König, H., and Metcalf, B. W., *Medicinal Chemistry for the 21st Century*, Blackwell Scientific, Oxford, 1992

[45] Livingstone, D. J., *Data Analysis for Chemists — Application to QSAR and Product Design*, Oxford University Press, Oxford, 1994

[46] Wermuth, C. G., ed., *The Practice of Medicinal Chemistry*, (1994), in preparation

[47] Boyd, D. B., *J. Med. Chem.* **36**, 1443–1449 (1993)

[48] *Sortware for Science*, SciTech International Onc., Bailiwick Court Building, 2231 N. Clybourn Ave. Chicago, IL 60614-3011, USA

[49] *EFMC Newsletter*, Prof. H. Timmerman, Free University of Amsterdam

[50] *International QSAR Society Newsletter*, Prof. J. Block, Oregon State University

[51] *ADAPT*, Prof. P. C. Jurs, PennState University, University Park, PA 16802, USA

[52] *APEX*, Biosym, 9685 Scranton Road, San Diego, CA 92121-2777, USA

[53] *ARTHUR*, Informetrix Inc., 2200 Sixth Avenue, Suite 833, Seattle, Washington 98121, USA

[54] *ASP*, OML, Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, UK

[55] *BIOSOFT*, 49 Bateman Street, Cambridge, CB2 1LR, UK

[56] *BMDP Statistical Software*, University of California Press, Berkeley, CA, USA

[57] *CATALYST*, BioCad Corp., Mountainview, CA, USA

[58] *CHEMX*, CDL (Chemical Design Ltd.), Oxford, UK

[59] *CLOGP*, Daylight CIS, 18500 Von Karman Ave 450, Irvine, CA 92715, USA and BioByte Corp., PO Box 517, Claremont, CA 91711-0517, USA

[60] *CLUSTAN*, Kingsburgh Road 16, Edinburgh EH12 6DZ, Scotland

[61] *C-QSAR*, BioByte Corp.m PO Box 517, Claremont, CA 91711-0517, USA

[62] *EXCEL*, Microsoft

[63] *GENSTAT*, deveoped by the Statistics Department of Rothampsted Experimental Station, and distributed by Numerical Algorithms Groups. Ltd., Mayfield House, 256 Banbury Road, Oxford, OX2 7DE, UK

[64] *GOLPE*, Tripos Associates Inc., St Louis, MO 63144-2913, USA and MIA, Via Gigliarelli 203/207, I-08124 Perugia, Italy

[65] *GRAFTOOL*, 3-D Visions Corp., 2780 Skypark Drive, Torrance, CA 90505, USA

[66] *HYPERCHEM*, Autodesk Inc., 2320 Marinship Way, Sausalito, CA 94965, USA

[67] *JMP*, from SAS Institute Inc.

[68] *MACSPIN*, D^2 Software Inc., P.O. Box 50052, Austin, TX 78763, USA

[69] *MINITAB*

[70] *MOLCONN-X*, Hall Associates Consulting, Quincy, MA 02170, USA

[71] *MULTISTAT*, Biosoft, 49 Bateman Street, Cambridge, CB2 1LR, UK

[72] *NCSS*

[73] *PARVUS*, Elsevier Scientific Software, Amsterdam

[74] *PCMODELS*, Daylight CIS, 18500 Von Karman Ave 450, Irvine, CA 92715, USA

[75] *PIROUETTE*, Informetrix Inc., 2200 Sixth Avenue, Suite 833, Seattle, Washington 98121, USA

[76] *POLARIS*, Molecular Simulations Inc., 200 Fifth Avenue, Waltham, MA 02154, USA

[77] *PROLOGP*, Compudrug Chemistry Ltd., PO Box 405, H-1395 Budapest, Hungary

[78] *PSTAT*

[79] *QUATTRO-PRO*, Borland

[80] *QSAR*, BioByte Corp., PO Box 517, Claremont, CA 91711-0517, USA

[81] *QSAR-PC*, Biosoft, 49 Bateman Street, Cambridge, CB2 1LR, UK

[82] *RECEPTOR*, MSI (Molecular Simulations Inc.), 200 Fifth Avenue, Waltham, MA 02154, USA

[83] *RS/1*, BBN Software Products Corp., 10 Fawcett Street, Cambridge, MA 02238, USA

[84] *SAS*, SAS Institute Inc., Box 8000, Cary, NC 27511, USA

[85] *SIGMAPLOT*, Jandel Scientific, 2591 Kerner Blvd., San Rafael, CA 94901, USA

[86] *SIMCA*, Umetri AB, PO Box 1456, S-90124 Umea, Sweden

[87] *SPSS*, SPSS Inc., 444 North Michingan Avenue, Chicago, IL, USA

[88] *STATA*

[89] *STATISTICA*, Statsoft, 2325 E. 13 St., Tulsa, OK 74104, USA

[90] *STATlab*, Slp Statistiques, 51/59 Rue Ledru-Rollin, F-94853 Ivry, France

[91] *TSAR*, OML, Oxford Molecular Ltd., The Magdalen Centre, Oxford Science Park, Sandford-on-Thames, Oxford OX4 4GA, UK

[92] *STATGRAPHICS*, Statgraphics Inc., 2115 East Jefferson Street, Rockville, MD 20852, USA

[93] *SYBYL-COMFA*, Tripos Associates Inc., St Louis, MO 63144-2913, USA

[94] *SYBYL-QSAR*, Tripos Associates Inc., St Louis, MO 63144-2913, USA

[95] *SYSTAT*, Systat Inc., 1800 Sherman Avenue, Evanston, IL 60201, USA

[96] *UNSCRAMBLER*, Camo AS, Jarleveien 4, N-7041 Trondheim, Norway

[97] van de Waterbeemd, H., ed., *Advanced Computer-Assisted Techniques in Drug Discovery. Methods and Principles in Medicinal Chemistry*, Vol. **3**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1995

[98] *Scientific Software Catalog for the PC and Macintosh*, Europa Scientific Software Corporation es^{2}c, 14 Clinton Drive, Hollis, NH 03049, USA

[99] Topliss, J. G., *Perspect. Drug Disc. Des.* **1**, 253–268 (1993)

[100] Cramer, R. D., *Perspect. Drug Disc. Des.* **1**, 269–278 (1993)

Chemometric Methods in Molecular Design
edited by Han van de Waterbeemd
© VCH Verlagsgesellschaft mbH, 1995

2 Molecular Concepts

2.1 Representations of Molecules

Peter C. Jurs, Steven L. Dixon, Leanne M. Egolf

Abbreviations

BA	Biological activity
CPSA	Charged partial surface area
HMO	Hückel molecular orbital
MO	Molecular orbital
NATOMS	Number of atoms
NOCC	Number of occupied orbitals
QSAR	Quantitative structure-activity relationship
RP-HPLC	Reversed-phase high performance liquid chromatography

Symbols

B_1	Substituent length parameter measured along attachment bond axis
B_5	Substituent length parameter measured perpendicular to attachment bond
$C, [C]$	Molar concentration
E_{HOMO}	Energy of the highest occupied molecular orbital
E_{LUMO}	Energy of the lowest unoccupied molecular orbital
E_{s}	Tafts steric substituent parameter
F	Dewar and Grisdale field substituent constant
$\mathscr{F}$	Swain and Lupton field substituent constant
f_i^{ELEC}	Electrophilic frontier orbital density for atom, *i*
f_i^{NUCL}	Nucleophilic frontier orbital density for atom, *i*
K	Equilibrium constant
k	Rate constant; capacity factor
L	Substituent length parameter measured along attachment bond axis
M	Dewar and Grisdale mesomeric substituent constant
n	Number of observations in a statistical correlation
P	Octanol/water partition coefficient
R	Original Randić molecular connectivity index
$\mathscr{R}$	Swain and Lupton resonance substituent constant
r	Statistical correlation coefficient; van der Waals radius
s	Standard deviation of regression
S_i^{ELEC}	Electrophilic superdelocalizability for atom, *i*

S_i^{NUCL}	Nucleophilic superdelocalizability for atom, *i*
${}^n\kappa$	Kappa shape index for paths of length, *n*
${}^n\kappa_\alpha$	Kappa shape index corrected for atom type
ν	Charton steric constant
ν_{eff}	Charton energy-corrected steric constant
π	Hydrophobic substituent constant
ϱ	Hammett reaction constant
${}^n\chi$	Path-*n* molecular connectivity index
${}^n\chi^{\mathrm{v}}$	Valence-corrected molecular connectivity index
σ	Hammett electronic substituent constant
σ^0	Normalized electronic substituent constant
σ^-	Exalted substituent constant for electron-withdrawing groups
σ^+	Exalted substituent constant for electron-releasing groups
σ^*	Taft inductive substituent constant
σ_{I}	Taft and Lewis fundamental inductive substituent constant
σ_{m}	Hammett constant for *meta* substituents
σ_{p}	Hammett constant for *para* substituents
σ_{R}	Taft and Lewis fundamental resonance substituent constant

2.1.1 Introduction

The objective of a QSAR study is to develop a relationship between the structures of a set of compounds and the biological activity (BA) of interest [1]. Such a relationship can be codified as follows:

$$\mathrm{BA} = f(\text{molecular structure}) = f(\text{descriptors}) \tag{1}$$

The nature of the descriptors used, and the extent to which they encode the structural features of the molecules that are related to the biological activity, is a crucial part of any QSAR study. The descriptors may be physico-chemical parameters (hydrophobic, steric or electronic), structural descriptors (frequency of occurrence of a substructure), topological (connectivity index), electronic (from a molecular orbital calculation), geometric (from a molecular surface area calculation), or they may be one of the hundreds of other descriptors, which have been proposed by researchers in this area.

The first studies in QSAR used an approach derived from physical organic chemistry and variations of the Hammett equation. This was soon followed by factorizing the interactions into three contributions-electronic, steric, and hydrophobic interations. The QSAR relationship, thus, became,

$$\log(1/C) = f(\sigma, E_s, \pi) \tag{2}$$

where C is the molar concentration of a compound producing a standard response, σ denotes electronic interations, E_s is the Taft steric substituent constant or a variation thereof and π is the hydrophobic substituent parameter. In each case, these substituent parameters are defined only for compounds in a homologous series.

The availability of substantial computational power has led more recently to the development of many sophisticated computed descriptors. Many of them follow the same reasoning as regards to the partitioning of molecular features into electronic, steric, and hydrophobic interactions. Descriptors introduced more recently, however, use other interpretations of molecular structure. For example, molecular connectivity indices and other topological indices depict molecules as graphs and use a graph theoretical approach to descriptor development. On the other hand, descriptors derived from the solvent accessible surface area depict the molecule as a collection of overlapping spherical atoms.

In this chapter, we will be discussing a variety of molecular representations that have been developed. First, we will introduce the basic concepts and point to the literature where necessary for further details. It should be emphasized that many authors have developed descriptors which are suited for particular problems, such as in QSAR studies, but which are also suitable in other types of investigations (e.g. structure-property studies, studies of chromatographic retention time as a function of structure). Reports of these investigations are scattered throughout the literature, and no attempt shall be made here to gather information about every descriptor reported to date.

2.1.2 Substituent Constants

QSAR grew out of physical organic chemistry studies on how differential reaction rates of chemical reactions depend on the differences in molecular structure. Characterizations of these differences in structure, which are due to the substitution of functional groups on to a fixed core structure, led to the development of *substituent constants*. It was not until with the appropriate substituent constants, encoding the electronic, hydrophobic, and steric aspects of a series of compounds, that QSARs could be developed for understanding structure-activity relationships.

2.1.2.1 Electronic Substituent Constants

Electronic substituent constants were as a direct result of the empirical observation made from certain chemical systems that substituents have the same relative effects on the rates of reaction equilibria, regardless of which reaction was being studied. The most significant breakthrough in this area occurred in 1937, when Hammett [2] proposed the, now famous, Hammett equation for the rate constants and equilibrium constants of reactions of *meta*- and *para*-substituted benzoic acid derivatives:

$$\log k = \log k_0 + \varrho\sigma \tag{3}$$

$$\log K = \log K_0 + \varrho\sigma \tag{4}$$

The constants, k_0 and K_0, refer to the unsubstituted compound, while k and K refer to a *meta*- or *para*-substituted version. The substituent constant, σ, reflects the inherent polar effect a given substituent has, on the rate or equilibrium of a reaction, relative to hydrogen. This effect is, in principle, independent of the reaction. The reaction constant, ϱ, depends on the nature and experimental conditions of the reaction under consideration and measures the sensitivity of the process to polar effects exerted by substituents. The reference reaction chosen for determining σ values was the aqueous dissociation of benzoic acids at 25 °C, where ϱ was defined to be unity.

A remarkably wide range of data involving benzene derivatives [3] has been successfully correlated according to the Hammett equation, using a single set of substituent constants, σ_m and σ_p, for *meta-* and *para*-substitution, respectively. There are many instances, however, when Eqs. (3) and (4) break down. These deficiencies have led to the development of several alternative substituent constant scales.

The first and most obvious limitation of the Hammett equation is that it does not hold, in general, for *ortho* substituents. This so-called "*ortho* effect" was identified by Ingold [4] as being steric in nature. Based on Ingold's hypothesis, Taft [5] proposed a quantitative measure for separating the inductive influence of a substituent from its steric effect. The substituent constant σ^* was based on the rates of acid- and base-catalyzed hydrolysis of esters of the form $X-CH_2-COOR$:

$$\sigma^* = (1/2.48)\,[\log(k/k_0)_{\mathrm{BASE}} - \log(k/k_0)_{\mathrm{ACID}}] \tag{5}$$

where X = H for k_0. Taft argued that σ^* should measure only the inductive influence of a substituent for two reasons: 1) the steric and resonance effects should essentially be the same in acidic and alkaline hydrolysis, and 2) the inductive component in $\log(k/k_0)_{\mathrm{ACID}}$ should be much smaller than in $\log(k/k_0)_{\mathrm{BASE}}$, because the ϱ values for the acidic hydrolysis of esters are much smaller than for alkaline hydrolysis. The factor 2.48 corresponds to the alkaline hydrolysis ϱ value and, thus, puts σ^* on the same scale as the Hammett σ_m and σ_p values.

Discrepancies in the Hammett equation were also noted in the ionization of phenols and anilines, when a strongly electron-withdrawing group such as $-NO_2$ was present in the *para* position. The most widely accepted explanation for this, is that the substituent receives electron density via "cross-conjugation" [6] or "through resonance", [7] and that this phenomenon is more important in the base, e.g. *p*-nitrophenoxide ion, than in the corresponding acid, as illustrated in Fig. 1. The increased stability of the base would then account for the unusually high acidities of *p*-nitrophenol and *p*-nitroanilinium ion. Studies of such systems have provided a set of exalted constants σ^- which may be used in favor of σ_p when cross-conjugation with an electron-withdrawing substituent occurs [8].

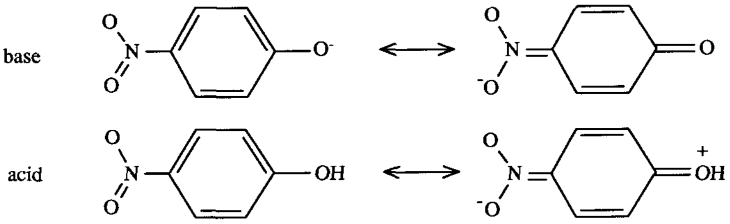


Figure 1. The resonance structures of *p*-nitrophenoxide ion and *p*-nitrophenol.

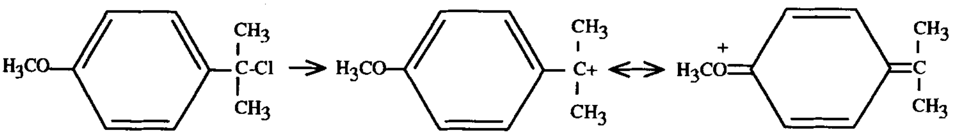


Figure 2. Illustration of the stabilization of an intermediate carbocation by the methoxy group.

In analogy to the σ^- constants for electron-withdrawing groups, Brown and coworkers [9] developed a set of substituent constants σ^+ for groups that release electron density via resonance. The reaction selected for defining σ^+ was the S_{N^1} solvolysis of *t*-cumyl chlorides in 90% aqueous acetone at 25 °C. Electron-releasing substituents such as $-OCH_3$ speed up the reaction by stabilizing the intermediate carbocation, as illustrated in Fig. 2. As with the σ^- constants, σ^+ may be used in place of σ_p for electron-releasing groups.

While the σ^- and σ^+ scales enabled the correlation of a broader set of data using the Hammett equation, the use of these exalted constants did not address the fundamental issue of how to account for resonance effects in general. Taft [10] showed that part of the problem stemmed from the fact that cross-conjugation was an important factor, even in the reference benzoic acid system. Although it was previously believed that cross-conjugation was of equal significance in benzoic acids and the corresponding benzoate ions, there was increasing evidence [11] that the effect was more important in the acid when groups such as $-OCH_3$ were present, as illustrated in Fig. 3. The implication was that the Hammett σ_p constants were biased from the beginning for certain electron-releasing substituents. To test this hypothesis, Taft [10] studied the ionization of *meta*/*para*-substituted phenylacetic and 3-phenylpropionic acids and the akaline hydrolysis of *meta*/*para*-substituted ethyl phenylacetates and benzylacetates. In all cases, the reaction center was insulated from the ring by one or two methylene units, which, as Taft argued, should minimize any cross-conjugation effects. The *meta*-substituted systems were used to determine the appropriate reaction constants ϱ, and these, in turn, were used to find the normalized substituent constants, σ^0, for *para* substituents. The σ^0 values indicated that groups such as $-N(CH_3)_2$, $-NH_2$, and $-OCH_3$ had much less of an influence in these systems than in the corresponding *para*-substituted benzoic acids. This helped to confirm that cross-conjugation was a significant component of the Hammett constants, σ_p for these substituents.

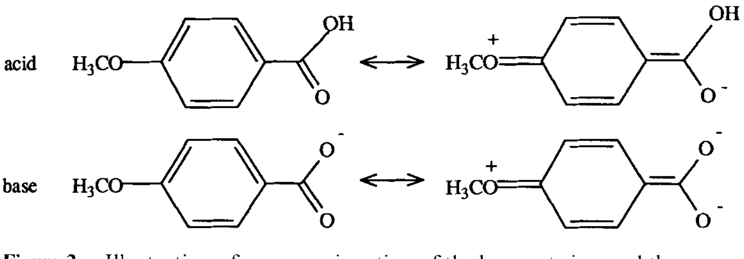


Figure 3. Illustration of cross-conjugation of the benzoate ion and the corresponding benzoic acid.

The proliferation of substituent constant scales increased the widespread desire to break down the effect of a substituent into the fundamental components of induction and resonance. Taft and Lewis [12, 13] proposed the following relationships for the original Hammett constants:

$$\sigma_p = \sigma_I + \sigma_R \tag{6}$$

$$\sigma_m = \sigma_I + \alpha\sigma_R \tag{7}$$

The inductive contribution σ_I was a scaled version of Taft's σ^* parameter [5], and α was a transmission coefficient for the resonance parameter, σ_R. Taft and Lewis suggested that the model could be used to determine whether deviations from the basic equations; Eqs. (6) and (7), were due to a breakdown in the applicability of the inductive scale, the resonance scale, or both.

Dewar and Grisdale [14] proposed that a molecule's structure be incorporated into a field and mesomeric representation of the substituent constant:

$$\sigma = F/r_{ij} + Mq_{ij} \tag{8}$$

The electrostatic field term was dependent on the inverse distance, $1/r_{ij}$, between the point of attachment, i, of the substituent and the point of attachment j, of the reaction center. The mesomeric term utilized the formal charge, q_{ij}, which arose at point j due to the attachment of a $-CH_2^{\ominus}$ group at point i. Dewar and coworkers [15] later modified the scheme to allow for a mesomeric field effect which had originally been ignored.

Swain and Lupton [16] showed that many of the previous substituent scales could be accurately represented according to the following equation:

$$\sigma = f\mathscr{F} + r\mathscr{R} \tag{9}$$

In deriving the field and substituent constant, $\mathscr{F}$, they assumed that the σ' scale for the dissociation of 4-substituted bicyclo[2.2.2]octane-1-carboxylic acids [17] could be written as a linear combination of the Hammett σ_m and σ_p constants, and that σ' contained no resonance component, i.e.,

$$\sigma' = a\alpha_m + b\sigma_p = \mathscr{F} \tag{10}$$

The constants, a and b, were determined by a least-squared fitting procedure for 14 different substituents. This alowed $\mathscr{F}$ to be calculated for any substituent, for which σ_m and σ_p were known. In order to determine the resonance constant, $\mathscr{R}$, Swain and Lupton assumed that σ_p for the substituent $-{}^{\oplus}N(CH_3)_3$ had no resonance component and solved Eq. (9) for f. By setting r equal to unity for the σ_p series, the $\mathscr{R}$ values for a large number of substituents were computed. With $\mathscr{F}$ and $\mathscr{R}$ defined for each substituent, the various σ scales were regressed against these constants to obtain appropriate values for f and r. From f and r, the percent field and resonance contributions were determined for each σ scale.

2.1.2.2 The Hydrophobic Substituent Constant, π

When working with a set of derivatives, the hydrophobicity of the compounds in the series can be represented on a relative scale with the hydrophobic substituent constant, π [18, 19]. The value for the substituent X is then defined as follows:

$$\pi_X = \log P_{RX} - \log P_{RH} \tag{11}$$

where P_{RX} is the partition coefficient of the derivative, and P_{RH} is the partition coefficient of the parent compound. The variable π_X expresses the variation in lipophilicity, which results when the substituent X replaces H in RH. For example, the value for the chloro substituent, π_{Cl}, is the difference between the partition coefficient for chlorobenzene and that of benzene. When π has a positive value, the substituent causes the derivative to favor the lipid phase, and when π has a negative value the derivative is more hydrophilic than the parent compound.

This equation can be reversed to allow the calculation of $\log P$ values for derivatives, given the $\log P$ for the parent compound and π values for the substituents of interest. Thus, $\log P$ for chlorotoluene is calculated as follows:

$$\log P_{\text{chlorotoluene}} = \log P_{\text{benzene}} + \pi_{Cl} + \pi_{Me} = 2.13 + 0.71 + 0.56 = 3.40 \tag{12}$$

The measured value of $\log P$ for chlorotoluene is 3.33, and thus, the agreement is good. When multiple substituents are present, simple additivity can fail due to interactions which must be taken into account in such circumstances. Compendia of π values are available in the literature [e.g. 20, 21].

2.1.2.3 Partition Coefficient – Log *P*

The relative affinity of a drug molecule for an aqueous or lipid medium is an important correlate of drug activity due to absorption, transport, and partitioning phenomena. The most widely used molecular structure descriptor to encode this property is the logarithm of the partition coefficient, *P*, between 1-octanol and water:

$$P = \frac{[C]_{\text{1-octanol}}}{[C]_{\text{aqueous}}} \tag{13}$$

where in this model $[C]_{\text{1-octanol}}$ is the concentration of a solute in the lipid phase, 1-octanol, and $[C]_{\text{aqueous}}$ is the concentration of the solute in the aqueous phase [21]. Compounds for which $P > 1$ are lipophilic or hydrophobic, and compounds for which $P < 1$ are hydrophilic.

Log *P* has been shown to be highly correlated with a diversity of biological activities, including drug activity, toxicity, pesticidal activity, genotoxic activity, and others. It is evident that $\log P$, as an operational definition of lipophilicity or hydrophobic bonding, plays a significant role in the interactions between drugs and their receptors. The lipophilic character of drugs is also an important factor in drug metabolism. In addition, the absorption, distribution, and excretion of many classes of drugs have been shown to be dependent on $\log P$.

The organic phase used most frequently for determining log P is 1-octanol. Studies have shown that it is a good compromise as regards to solvent properties, and most measurements and compilations of values have now been made with this lipid phase being used. Other lipid phases are also used, but less often.

Log P values have been measured by two main methods: the "shake-flask" method and liquid chromatographic methods. The shake-flask method involves the distribution of a compound between an aqueous phase and organic phase, and once equilibrium has been attained, measuring the concentrations of the compound in the two phases. The values obtained depend on a number of experimental factors, including the pH and ionic strength of the aqueous phase, the nature of the buffer used, the purity of the organic phase, purity and stability of the drug compound, solute concentration, temperature, stirring, the analytical method used to determine the equilibrium concentrations, as well as other factors.

Reversed-phase high performance liquid chromatography (RP-HPLC) is now the method of choice for measuring log P. This method has been reviewed [22] and yields log k, which is a capacity factor, calculated as follows:

$$k = (t_R - t_O)/t_O \tag{14}$$

where t_R and t_O are the retention times of the drug compound and a non-retained compound, respectively. A lipophilic stationary phase is used, such as an inert support coated with 1-octanol or alkylsilylated silicas. The mobile phase consists of a buffered aqueous phase and an organic modifier such as methanol, acetonitrile, or acetone. Extensive studies on these systems have general many papers discussing the applicability of the method. The RP-HPLC method has many advantages over the shake-flask method, including greater accuracy and precision, a wider range of applicability, decreased dependence on impurities, speed, and only small amounts of the drug compound are required.

Measured log P values have been evaluated and gathered into a database now containing more than 40000 log P values which have been measured in more than 300 solvent systems. The database contains more than 18000 log P values measured with the octanol/water system. A subset of 8162 selected values is called the Starlist. The entire database will be published soon [23] and is part of the Pomona College MedChem project.

There have been a number of methods, including substituent additivity, developed for the calculation of log P from molecular structure, fragments, atomic contributions and/or surface area, molecular properties, and solvatochromic parameters [24]. The first general method for the calculation of log P was proposed by Rekker and Nys [25] in 1973. In this method, after summing fragment constants for the molecule in question, any necessary correction factors for intramolecular interactions between the fragments, such as electronic, steric, or hydrogen-bonding effects, were added.

This fragment addition method led to the method which now is the mostly widely used and which was developed by Leo and Hansch [21, 23, 24]. Here, the log P of a compound is computed by summing over the contributions for the fragments and then applying a number of correction factors as needed.

$$\log P = \sum_i a_i f_i + \sum_j b_j F_j \tag{15}$$

where f_i are fragment constants and F_j are correction factors. The solute structure is broken down into a series of fragments which are separated by isolating carbon atoms according to a set of rules. An isolated carbon is a carbon that is not doubly or triply bonded to a heteroatom. The groups of atoms that remain are polar fragments. The contribution for each fragment, f_i, multiplied by the occurrence of that fragment, is added accumulatively. This sum is then corrected for a number of factors according to solvent theory. Each correction factor has an associated value, F_j, and this is multiplied by the number of instances of the correction in the structure, b_j. Correction factors include those due to molecular flexibility, branching, polar fragment interaction factors, *ortho* effects, and aromatic interactions [21, 24].

This fragment additivity method has been implemented in a commercially available software package named CLOGP, which currently exists in version three as CLOGP-3. The only input needed to the program is a structural representation of the compound. A description of the current status of the software, its limitations, and planned enhancements has appeared recently in the literature [24].

2.1.2.4 Steric Substituent Constants

In a homologous series of compounds, the different biological activity for the compounds is often related to the size of the substituents. Bulky substituents can interfere with the intermolecular reactions, which lead to drug activity. The quantitative encoding of the steric aspect of drug structure has been accomplished by a series of steric substituent constants.

Taft's Steric Parameter, E_s

The first steric parameter, E_s, was developed by Taft [5] and describes the intramolecular steric effects on the rate of reaction. Acid-catalyzed ester hydrolysis was used to derive the following relationship:

$$E_s = \log (k_R/k_{Me})_A \tag{16}$$

where k_R and k_{Me} are the acid-catalyzed rate constants of hydrolysis for the compounds RCOOR′ and MECOOR′. This equation assumes that there is no inductive or resonance contribution and that E_s is dependent of the medium in which the rate constants are measured. By definition, $E_s = 0$ for the methyl group. Tables of E_s values have been published [e.g., 26].

Bulkier substituents usually generate negative values for E_s. The E_s parameter is correlated with the van der Waals radii of the substituent's atoms and is also related to electronic contributions.

Although these E_s parameters were derived for physical organic chemistry, they have also been extensively applied to biological activity problems with success and form one of the traditional substituent constant parameters for QSARs.

Charton's Steric Constant, v

Efforts to bypass the uncertainties and limitations which accompany the Taft method led to new research in the area. Charton, for example, observed that the E_s parameter closely paralleled group radii [26, 27]. Hence, he developed a new set of steric measures, termed ν values, which have the following form,

$$\nu_X = r_{vX} - r_{vH} = r_{vX} - 1.20 \qquad (17)$$

where r is the van der Waals radius of the symmetrical substituent (e.g. F, Cl, $-CH_3$, $-CBr_3$, $-t\text{-}C_4H_9$) and 1.20 is the radius (in Angstroms) of hydrogen, the standard [28]. In order to incorporate the effects which were as a result of the conformation which, in turn, were dependent on energetic factors, a modified parameter, the effective steric value, v_{eff}, was later introduced [20, 29]. This parameter takes into account both the $\log (k_R)_A$, as above, as well as the original ν value corresponding to the minimum substituent radius. This results in a scale of steric measures which essentially is energy-corrected and is, thus, independent of the medium. Charton has published ν_{eff} values for more than 300 substituents [30]. Hansch and Leo [20] illustrated the wide diversity of substituent types, which can be characterized as a whole, while also demonstrating the relationship between Taft's E_s constant and Charton's ν_{eff} by the following correlation.

$$E_s = -2.062\nu_{\mathrm{eff}} - 0.194$$
$$n = 104 \qquad r = 0.978 \qquad s = 0.250 \qquad (18)$$

The fact that there is no obvious structural explanation as to why several substituents deviate significantly from the majority of substituents in the data set, indicates that much is still unknown about the representation of the steric nature of molecules.

STERIMOL Parameters

In an atempt to go beyond the Taft parameters, which were designed for simple homogeneous organic reactions, Verloop and coworkers [31, 32] designed a multiparametric method for characterizing the steric features of substituents in more complex biological systems. With their computer program STERIMOL [32], covalent and van der Waals radii, along with standard bond angles and lengths, are used to build chemically feasible three-dimensional models of molecular substituents. From these models, the spatial requirements of any type of end group can be effectively represented by distance-based measures.

The substituent to be described is represented by the van der Waals radii of the atoms, forming the group, by standard bond lengths and angles and by reasonable conformations derived from molecular mechanics. In the original approach [31], five directions were used to represent the shape of the substituent. In a later variation [33], just three are used, L, B_1 and B_5. The length parameter, L, is defined as the length of the substituent measured along the axis of the bond that joins the substituent to the parent molecule. B_1 is the smallest distance from the axis of the attachment bond, measured perpendicularly to the edge of the substituent. B_5 is the maximum width of the substituent and has no directional relationship with B_1. The ratios L/B_1 and B_5/B_1 are useful measures of the relative deviations of the substituent's shape from a sphere. The B_1 parameter has been shown to be highly correlated with Taft's E_s parameter as well as Charton's ν_{eff} parameter.

A table with more than 100 values for L, B_1, and B_5 has been published [26]. The successful application of these parameters, alone or in combination with other physico-chemical descriptors has been illustrated by examples in a review by Fujita and Iwamura [34] as well as in a number of additional studies, where the interactions between various ligands and biomolecules were explored [26].

2.1.3 Whole Molecule Representations

Developments in computer methods for structural representation and manipulation of chemical structures have led to the generation of a host of methods for representing entire molecular structures. Many of the whole molecule descriptors are extensions of the substituent constant approach, but many of them are also completely new approaches to the problem of representing whole molecules.

Descriptors, based on the connection table for a molecule, are topological in nature, and their values are independent of a three-dimensional conformation. These descriptors can be counts of the substructures present in the molecules being encoded, or they can be calculated topological indices that attempt to encode the size, shape, or branching in the compound by manipulation of graph-theoretical aspects of the structures.

In contrast to the topological descriptors, descriptors which are derived from a three-dimensional conformation of the molecule are dependent on the exact conformation chosen and, therefore, on the molecular modeling program employed. There are now many commercial molecular modeling programs available, and many of them have the capability of producing descriptors from the molecular models they develop. Since conformational analysis often requires the calculation of atomic charges, these programs can also produce electronic descriptors.

2.1.3.1 Topological Descriptions

The basic information about the structure of an organic compound is contained in the corresponding connection table, which is a compact representation of types of atoms and bonds, and of the connections forming the molecule. Since the connection table is the usual storage medium for structures in chemical database systems, they are easily accessible and have been used to develop descriptors.

Substructure-Based Descriptors

With the advent of chemical structure handling computer systems a convenient and fast substructure searching was facilitated. A byproduct of this capability is that compounds in QSAR studies may be represented by integers, which in turn, are derived from substructure counts. Encoding organic structures as numbers, which reflects their constituent substructures, is appealing in its simplicity and conforms to the organic chemistry point of view of chemical structures and the importance of component parts of the structure.

One application of substructure descriptors is as indicator variables. Indicator variables have been used in QSAR studies for a long time, since it is easy to generate just one such variable. Two sets of compounds which differ from each other only by a substructure existing in one set but not the other can be studied as an entire set when using an indicator variable. This yields a model which simultaneously utilizes all other independent variables and then combines the models via the indicator variable. Strictly speaking, such an approach should only be used when the two sets of compounds are identical in every respect, except for the substructure being coded with the indicator variable. Such a strategy has been used in many studies to build a model for a larger set of compounds than would otherwise have been possible.

Another use of substructure-based counts is found in certain approaches in QSAR studies, involving a systematic examination of substructures and how they are related to the biological activity. Such an approach has been successfully implemented in the CASE program by Klopman and coworkers [35, 36]. In this work, the substructural units were built up from bonded pairs of atoms to larger units which were called biophores and biophobes, depending on whether their presence correlates with the presence or absence of biological activity. The software produced a list of those substructural units that correlate most highly with activity and inactivity among the training set compounds, and then a mathematical model was built using these substructural units as the independent variables. Thus, the molecules in the training set were represented by a list of the substructural units present.

Topological Indices

A long-standing goal in chemistry is to represent chemical structures in numerical form as succintly, but as completely as possible. When molecular structures are represented as graphs, [37] then this quest can be equated to seeking ways in which graphs can be represented as numbers. *Topological indices* have been developed by chemists in pursuit of this goal [26, 38]. A topological index is a numerical quantity that is mathematically derived in a direct and unambiguous manner from the structural graph of a molecule. Since isomorphic graphs possess identical values for any given topological index, these indices are referred to as graph invariants. Topological indices ordinarily encode both molecular size and shape at the same time. More than 50 topological indices have been presented in the literature since their first development. In this section we will discuss several widely used topological indices and show how they have been applied to chemical problems.

The first index based on a graph approach to molecular structure was developed by Wiener [39] in 1947. The path number was defined as the number of bonds between all pairs of atoms in an acyclic molecule. Using the path number and an additional index, Wiener was able to fit alkane boiling points fairly well. The Wiener number is inversely proportional to the compactness of a molecule. In 1971 Hosoya [40] connected the Wiener number with graph theory, pointing out that the Wiener number is the half-sum of all the distance matrix entries for a molecule.

In 1975, Randić [41] proposed a topological index that has now evolved into the most widely used of topological indices in chemical studies. This branching or connectivity index was originally defined as,

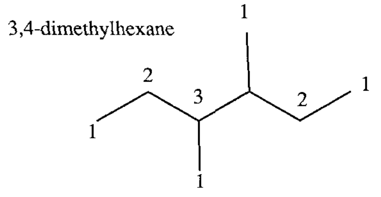
$$R = \sum_{\text{all bonds}} \frac{1}{(mn)^{1/2}} \qquad (19)$$

where the summation includes one term for each edge in the hydrogen-suppressed structural graph. Thus, when the graph is representing an organic molecule, there is one term in the summation for each bond in the structure. The variables, m and n, are the degrees of the adjacent modes joined by each edge. In terms of chemical structure, this is the number of bonds attached to each atom participating in the bond.

Fig. 4 shows the sequence of steps for the calculation of the value for this simple topological index for the example molecule, 3,4-dimethylhexane. At the top of the figure, the molecular structure is shown as a graph with the degree of each node labeled. The value ${}^1\chi = 3.717$ reflects both the size and the degree of branching of the structure. It is related to the size of the molecule, because when extra atoms and bonds are added, more terms are added to the summation, and the value increases. ${}^1\chi$ is also related to the degree of branching of the molecule, because when more branching occurs, the denominators for those terms become larger and the terms themselves become smaller, thus, decreasing the overall value for the index.

The normal valency of a carbon atom is 4, so the valencies of the nodes in the hydrogen-suppressed structural graphs of simple alkanes cannot exceed 4. Therefore, there are 10 possible sets of edges: $1-1$, $1-2$, $1-3$, $1-4$, $2-2$, $2-3$, $2-4$, $3-3$, $3-4$, $4-4$. The $1-1$ edge type occurs only in ethane, and the edges of type $1-4$ and $2-2$ each yield the same product. Thus, this branching index is based on the decomposition of a compound into eight different carbon-carbon bond types. Since the number of different bond types is limited, it follows that the branching index value can be the same for different molecules. For example, 3-methylheptane and 4-methylheptane have identical values for this branching index.

The simple branching index discussed above involves a summation over all paths of length 1 being treated in the graph. This viewpoint has been extended to include the definition of additional indices corresponding to paths of length 2, 3, or longer, and to other subgraphs such as clusters and path-clusters. This entire class of



$${}^1\chi = \frac{1}{\sqrt{1{\times}2}} + \frac{1}{\sqrt{2{\times}3}} + \frac{1}{\sqrt{1{\times}3}} + \frac{1}{\sqrt{3{\times}3}} + \frac{1}{\sqrt{1{\times}3}} + \frac{1}{\sqrt{2{\times}3}} + \frac{1}{\sqrt{2{\times}1}}$$

$$= .707 + .408 + .577 + .333 + .577 + .408 + .707$$

$$= 3.717$$

Figure 4. The calculation of ${}^1\chi^v$ for the example molecule 3,4-dimethylhexane.

Table 1. Valence delta values for carbon, nitrogen, and oxygen in different bonding environments for use in calculating valence-corrected molecular connectivity indices

$-CH_3$	1	$-NH_2$	3	$-OH$	5
$-CH_2-$	2	$>NH$	4	$-O-$	6
$>CH-$	3	$=NH$	4	$=O$	6
$=CH-$	3	$>N-$	5		
$=C<$	4	$=N-$	5		
$>C<$	4	$\equiv N$	5		

topological indices are commonly called *molecular connectivity indices* [42]. The original Randić branching index is referred to as the *path-1 molecular connectivity* ${}^1\chi$. The higher order indices are calculated by equations analogous to the simple equation for path-1 molecular connectivity.

The following equation generates the path-2 molecular connectivity from the degrees of the three edges involved in the definitions of paths of length two:

$$ {}^2\chi = \sum_{\substack{\text{length 2}\\ \text{paths}}} \frac{1}{(mnp)^{1/2}} \qquad (20)$$

where m, n, and p are the degrees of the atoms of each path of length two. For 3,4-dimethylhexane there are six terms in the summation for the six paths of length two in the molecule. The denominator contains the following terms, $(1 \cdot 2 \cdot 3)^{1/2}$, $(2 \cdot 3 \cdot 3)^{1/2}$, $(2 \cdot 3 \cdot 1)^{1/2}$, $(3 \cdot 3 \cdot 1)^{1/2}$, $(1 \cdot 3 \cdot 2)^{1/2}$, and $(3 \cdot 2 \cdot 1)^{1/2}$, and the overall value for ${}^2\chi$ for 3,4-dimethylhexane is 2.201.

The simplest molecular connectivity indices described in this context do not allow for the differentiation of atom types. In order to generalize the molecular connectivity indices and make them more useful for the characterization of organic molecules containing heteroatoms, the following enhancement has been developed. In the denominator of the equation, *delta* values were used in place of the degree of the node. The *delta* values are defined as,

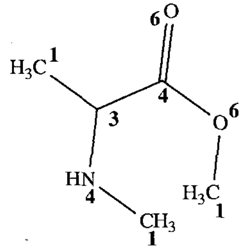
$$ \delta^{\mathrm{v}} = Z^{\mathrm{v}} - h \qquad (21)$$

where Z^{v} is the number of valence electrons for the atom, and h is the number of attached hydrogens. Thus, a carbonyl oxygen has a $\delta^{\mathrm{v}} = 6$, and a nitrogen atom as a secondary amine has a value of $\delta^{\mathrm{v}} = 4$. Table 1 provides a complete list of the valence delta values for carbon, nitrogen, and oxygen atoms in various bonding environments.

Molecular connectivity indices calculated with these delta values are referred to as *valence molecular connectivity indices* and have the superscript v. Fig. 5 shows the calculation of the valence path-1 molecular connectivity index ${}^1\chi^{\mathrm{v}}$ for 2-(methylamino)propionic acid methyl ester (or *N*-methyl alanine methyl ester).

The valence molecular connectivity index has been correlated with many physico-chemical properties of organic compounds. The index is easy to compute and is thus, more accessible than values derived from complicated experimental measurements. An example to demonstrate this is given by the correlation between

2-(methylamino)propionic acid methyl ester



$$ {}^1\chi^v = \frac{1}{\sqrt{1\times3}} + \frac{1}{\sqrt{3\times4}} + \frac{1}{\sqrt{4\times1}} + \frac{1}{\sqrt{3\times4}} + \frac{1}{\sqrt{4\times6}} + \frac{1}{\sqrt{4\times6}} + \frac{1}{\sqrt{6\times1}} $$

$$ = 2.47 $$

Figure 5. The calculation of ${}^1\chi^v$ for the example molecule 2-(methylamino) propionic acid methyl ester.

log P and ${}^1\chi^v$ for 138 simple organic compounds, including 24 esters, 9 carboxylic acids, 49 alcohols, 28 amines, 16 ketones, and 12 ethers [43]. The log P for a compound is the logarithm of the partition coefficient of the compound between water and 1-octanol. Log P of organic compounds has been shown to be related to biological activity and environmental transport rates in hundreds of studies, and is thus, of great interest. The correlation between measured log P values and calculated ${}^1\chi^v$ values is illustrated graphically in Fig. 6. The equation for the best fit:

$$ \log P = 0.95\,{}^1\chi^v - 1.48 \qquad (22) $$

$$ n = 138 \qquad r = 0.986 \qquad s = 0.152 $$

Thus, molecular connectivity indices can be used to encode information about molecular structures that are also represented by experimentally measured quantities. The molecular connectivity indices are the most widely used of topological indices for quantitative structure-activity relationship and quantitative structure-property relationship studies. Kier and Hall [42] include a list of 158 references for examples, in which molecular connectivity played a prominent role. A recent paper reports the availability of software for calculating molecular connectivity indices using a microcomputer [44].

Kappa Indices

A series of graph theoretical indices have been develped by Kier, which relate to the molecular shape of a molecule [45–47]. The method is based on graph theory and is not dependent on molecular geometry. In its simplest form, the shape calculations weight all non-hydrogen atoms and bonds equally. Other forms use atom and bond type information. In all cases, hydrogen atoms are not treated explicitly.

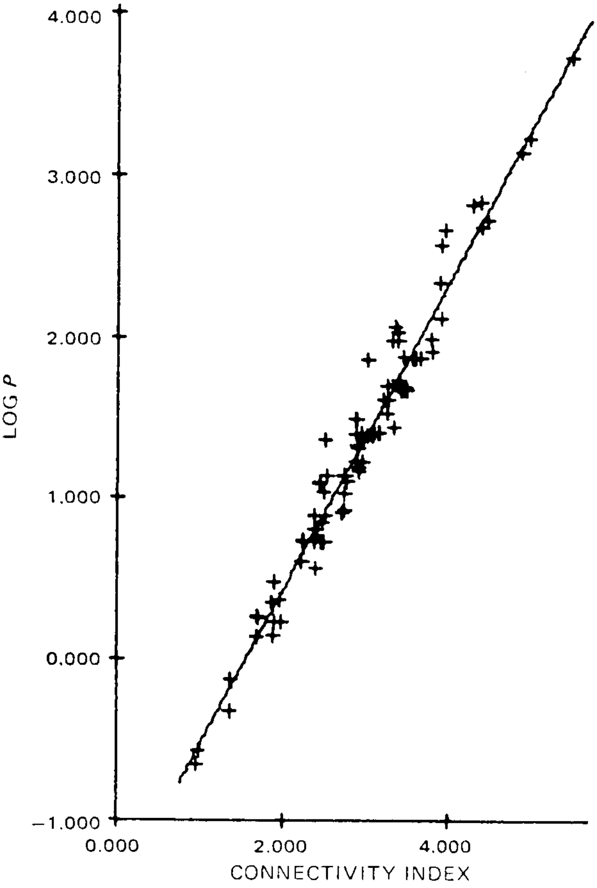


Figure 6. Plot of the logarithm of the octanol/water partition coefficient versus the path-one molecular connectivity $^1\chi^v$ for a set of 138 simple organic compounds. (Reproduced with permission of the American Pharmaceutical Association.)

Kappa indexes are calculated relative to the least branched (linear) and most branched (star) compounds with the same number of atoms as the molecule beinginvestigated. The equation for $^2\varkappa$ illustrates this,

$$^2\varkappa = 2(^2P_{\max})\,(^2P_{\min})/(^2P_i)^2 \tag{23}$$

where $^2\varkappa$ is the shape index based on paths of length 2, $^2P_{\max}$ is the maximum number of 2 bond fragments possible with the number of atoms in a molecule, *i*, $^2P_{\min}$ is the minimum number of 2 bond fragments possible with the number of atoms in a molecule, *i*, and ^2P_i is the number of 2 bond fragments in a molecule, *i*.

The equations for $^1\varkappa$ and $^3\varkappa$ are similar:

$$^1\varkappa = 2(^1P_{\max})\,(^1P_{\min})/(^1P_i)^2 \tag{24}$$

$$^3\varkappa = 2(^3P_{\max})\,(^3P_{\min})/(^3P_i)^2 \tag{25}$$

The atom type may be accounted for by using a corrective term, α, that is derived from the covalent radius of an atom relative to the radius of an sp^3 carbon. Using atom type corrections the equation for ${}^2\chi_\alpha$ is as follows:

$$ {}^2\chi_\alpha = (A + \alpha - 1)(A + \alpha - 2)^2/({}^2P_i + \alpha)^2 \tag{26} $$

where A is the number of atoms in a molecule, i, and $(A - 1)(A - 2)^2 = 2({}^2P_{\max})({}^2P_{\min})$. For a straight chain graph: ${}^2\chi_\alpha = {}^2\chi + \alpha$.

2.1.3.2 Electronic Whole Molecule Descriptors

A large variety of electronic whole molecule descriptors have been used to encode electronic features in QSAR and QSPR investigations. These descriptors are distinguished from electronic substituent constants in that a single value is asigned for a given compound. These values range from experimental to semi-empirical and to quantum mechanical values, and may encode either general features of the entire molecule or local features of a specific site in the molecule. Some of the more commonly used descriptors are covered here; more extensive compilations can be found in the literature [e.g. 48, 49].

A number of electronic descriptors may encode the effects or strengths of intermolecular interactions. The more commonly recognized intermolecular forces arise from the following interactions: ion-ion, ion-dipole, dipole-dipole, dipole-induced dipole, dispersion, and hydrogen bonding. Certain electronic descriptors are clearly associated with one or more of these types of interactions.

Ionic interactions have been encoded in drug potency studies through the use of ionization constants [50]. As a descriptor, the ionization constant provides information about the extent to which a drug molecule ionizes, which is known to strongly influence the absorption and distribution of the drug [51].

Electric dipole moments obviously encode the strength of polar-type interactions and Lien et al. [52] have reviewed their use as descriptors in QSAR studies. While extensive compilations of experimental dipole moments are available [53, 54], many accurate empirical [55–57] and quantum mechanical [58–61] techniques exist for estimating them.

Molecular polarizability and molar refractivity are closely related properties that are a measure of a molecule's susceptibility to becoming polarized. These descriptors are often useful in situations, where dipole-induced dipole and dispersion interactions play an important role. They are readily calculated [62] from the refractive index and the molar volume; however, applications in QSAR and QSPR usually employ empirical estimates, based on atomic, bond, or group contributions. A recent paper by Miller [63] includes a review of techniques that have been used to estimate molecular polarizabilities. Methods for estimating the molar refractivity may be found in the literature [e.g. 64, 65].

Hydrogen bonding has long been recognized as an important factor in the physical properties and biological activities of compounds. Fujita et al. [66] have reviewed the use of hydrogen bonding parameters in QSAR studies. Most applications have involved the use of an indicator variable for the presence of a hydrogen bond donor or acceptor group. Kamlet and Taft [67, 68] developed more quantitative scales based on solvatochromic shifts for carefully selected solutes with known hydrogen bonding characteristics.

While descriptors related to intermolecular interactions are useful for predicting bulk physical properties and certain types of biological activities, they provide little direct information about the reactivities of compounds. This type of information, however, is available through molecular orbital (MO) calculations. The Hückel molecular orbital (HMO) method [69] has provided a number of so-called reactivity indices [70], although any MO technique could have been used to calculate these descriptors.

Reactivity indices are usually categorized as either electrophilic or nucleophilic, depending on whether the reaction of interest involves electrophilic or nucleophilic attack. Perhaps the simplest of such descriptors are E_{HOMO} and E_{LUMO}, the energies of the highest occupied and lowest unoccupied MOs, respectively. The HOMO energy is roughly related to the ionization potential of a molecule, while the LUMO energy is related to the electron affinity. The magnitudes of these quantities are measures of the overall susceptibility of the molecule to losing a pair of electrons to an electrophile or accepting a pair of electrons from a nucleophile.

Site-specific reactivity indices are obtaied by considering electronic information at specific locations in the molecule. The electrophilic and nucleophilic superdelocalizabilities of atom i are energy-weighted atomic electron densities, which, for the HMO method, are given by [70]:

$$S_i^{\mathrm{ELEC}} = \sum_{j=1}^{NOCC} \frac{2c_{i,j}^2}{|e_j|} \tag{27}$$

$$S_i^{\mathrm{NUCL}} = \sum_{j=NOCC+1}^{NATOMS} \frac{2c_{i,j}^2}{|e_j|} \tag{28}$$

Here, $c_{i,j}$ is the LCAO-MO coefficient for atomic orbital i in MO j, and e_j is the energy of MO j. The sums in Eqs. (27) and (28) represent the occupied and unoccupied MO's, respectively, and the factor of 2 assumes a double occupation of each MO. The electrophilic superdelocalizability is a rough measure of the availability of electrons in atom, i; nucleophilic superdelocalizability is a measure of the availability of "room" on atom, i, for additional electron density. While these indices are atomic in nature, they may be classed as whole molecule descriptors if atom, i, has a fixed position in a series of congeners, or if the maximum superdelocalizability among all the atoms has been chosen.

If one considers only the electron densities in the highest occupied and lowest unoccupied MOs, the so-called electrophilic and nucleophilic frontier orbital densities are given by:

$$f_i^{\mathrm{ELEC}} = 2c_{i,NOCC}^2 \tag{29}$$

$$f_i^{\mathrm{NUCL}} = 2c_{i,NOCC+1}^2 \tag{30}$$

These descriptors assume that the HOMO and LUMO are far more important than the other MOs in determining the position and likelihood of electrophilic or nucleophilic attack. Again, when used in the manner discussed previously, these atom-specific indices become whole molecule descriptors.

2.1.3.3 Geometric Descriptors

Biological activity is often related to the shape and size of the active compounds as well as the degree of complementarity of the compound and a receptor. With the given methods for generating three-dimensional molecular models of compounds, these models can be used to develop geometric descriptors. Many molecular modeling routines have the capability of calculating geometric descriptors from the resulting conformations. An extensive study of molecular conformation, and a detailed investigation of interactions between drug molecules and receptors (which often employ interactive computer graphics), goes beyond the scope of this chapter, although this is an extremely active area of research in QSARs.

Molecular Volume

One of the most commonly calculated decriptors for biological activity investigations is the molecular volume. An early volume approximation method, introduced by Bondi [71], hinges on group contribution techniques. By treating van der Waals radii as adjustable parameters, Bondi derived group contribution values for individual atoms and functional groups. Thus, when presented with a new molecule, whose volume was as yet unknown, the researcher merely has to add up the pre-established increments in order to calculate the Bondi van der Waals volume.

Probably the most widely used volume estimation technique in recent years the volume estimation technique developed by Pearlman [72]. This algorithm utilizes numerical integration, in which a molecule is viewed as a set of overlapping atomic spheres. The integration technique divides each sphere either into uniform lunes or longitudinal sections. For a given atom, the volume of each lune, that is occluded by intersection with neighboring spheres, is subtracted from the total volume of that atom. The total molecular volume is then simply the sum of the atomic contributions.

Molecular Surface Area

Surface area has a prominent effect on the interactions which occur between a drug molecule and its surroundings. When the surface area is introduced as a descriptor in chemometric analyzes, it has been found to contribute statistically significant information in correlations developed for water solubility, octanol-water partition coefficients, activity coefficients and boiling points [73–79].

While a number of surface area approximation techniques [80, 81] have been proposed, the methods, which currently gain the most attention, are those of Lee and Richards [82], Hermann [83], and Pearlman, [75, 84] who developed a significantly more efficient algorithm based on Hermann's original work [83]. In these three algorithms, atomic surface areas are determined by cutting a molecule's individual spheres into flat slices, in analogy to the algorithm of Lee and Richards [82], or to the algorithm of lunes, as described by Hermann and Pearlman [83]. The overlap between spheres is calculated and the non-occluded areas are summed over to yield the surface area that is associated with a molecule.

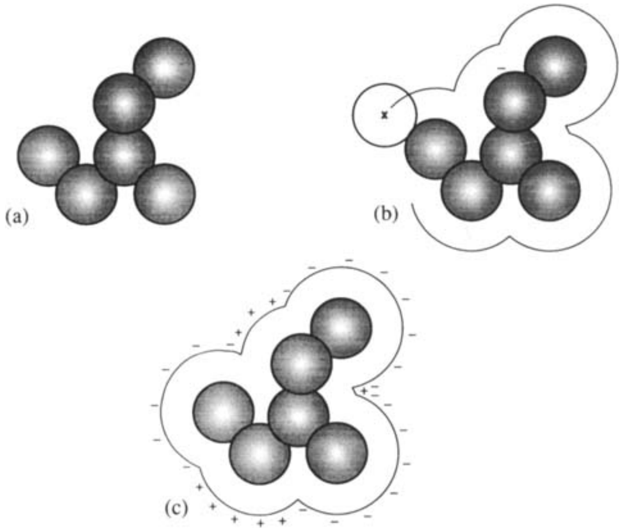


Figure 7. Diagram of the molecular representation used to derive the charged partial surface area (CPSA) descriptors.

Often a more pertinent and useful structural parameter in molecular design studies has been proven to be the solvent-accessible surface area, which is simply a mathematical extension of the surface area just described. Since many properties and activities (e.g. drug transport, docking) are a consequence of the type and strength of the solute-solvent interactions, this parameter was designed to reflect the amount of a molecule's exposed surface, which is actually capable of coming into direct physical contact with a neighboring solvent molecule. This accessible, or contact area is determined by adding the solvent radius (1.5 Å for water) to the original van der Waals radii as previously defined. Conceptually, this new area is viewed as being the surface, traced out by the center of a solvent sphere, as that sphere is rolled over the entire van der Waals area of the molecule of interest (see Fig. 7a and 7b). The utility of this information, quantified through this parameter, is clearly illustrated in the following section.

Charged Partial Surface Area

Properties influenced by interactions, which are polar in nature, have traditionally been difficult to model. Since the strength of these interactions is thought to be a function of the size, shape and charge distribution throughout a molecule, attempts to better understand the added structural complexities of polar molecules spurred on researchers to develop new groups of descriptors, which could capitalize on both the combination of surface area and charge information.

Advances in this area began with Grigoras' work with electrostatic molecular surface interaction terms [85]. Two structural features are first quantified: the solvent-accessible surface area of individual atoms and the molecular energy assigned to the exposed areas. The surface area associated with each atom – in the molecular environment – is determined using Pearlman's SAREA (Surface AREA) program

[84]. By modifying this program, the researcher can then use the net atomic charges calculated via the EHT molecular orbital method [86] to estimate the surface distribution of the molecular energy. Finally, charge scaling factors are incorporated to correct for discrepancies in estimating the site and strength of the polar interactions expected within the molecule.

Four descriptors are derived from this charge and surface area information. One is simply the total molecular surface area. The other three combine charges, surface areas and correction factors to yield a negatively charged surface area term, a positively charged surface area term and a hydrogen bonding term (where applicable). The structure-property relationship is ultimately developed by regressing these four terms, Grigoras' successful predictions of both critical temperatures and critical volumes [85] illustrate the advantages of these methods.

A continuation of this research soon followed with the development of Stanton and Jurs' charged partial surface area (CPSA) descriptors [87]. These parameters present various combinations of solvent-accessible surface area information from the SAVOL (Surface Area and VOLume) program of Jurs et al. [88] based on algorithms by Pearlman [72] as well as partial atomic charge information from Dixon and Jurs' [89] expanded version of Abraham and Smith's CHARGE algorithm (see Fig. 7c) [90]. Both σ and π charges are included in this iterative algorithm.

Twenty-five CPSA descriptors were proposed. Key structural information which is represented includes the summed accessible surface areas of the positively charged atoms, the charge associated with the exposed areas, the total positive charge in the molecule, the summed positive surface area relative to the total molecular surface area, and the charge of the most positive atom relative to the total positive charge. The corresponding information can also be obtained for the negative charges and negatively charged surface area. Finally, the differences between the positive- and negative-specific descriptors are also calculated to reveal net charge and surface area information.

The CPSA descriptors have found immediate use in both structure-property and structure-activity studies. These descriptors, when used in combination with other physico-chemical features, have been instrumental in developing strong correlations for numerous chemical and engineering properties including surface tensions [91], chromatographic retention indices [92, 93], boiling points [94–96], critical temperatures [97] and auto-ignition temperatures [98]. Although these parameters have not been studied as extensively in conjunction with biological processes, investigations of Henry's Law constants [99], odor thresholds [92] and odor intensities [100], show that the CPSA descriptors have considerable potential in this context as well.

References

[1] Ramsden, C. A., ed., *Quantitative Drug Design*, Pergamon Press, Oxford, 1990

[2] Hammett, L. P., *J. Am. Chem. Soc.* **50**, 96–103 (1937)

[3] Jaffé, H. H., *Chem. Rev.* **54**, 191–261 (1953)

[4] Ingold, C. K., *J. Chem. Soc.* **1930**, 1032–1039

[5] Taft, R. W., Separation of Polar, Steric, and Resonance Effects in reactivity. *In: Steric Effects in organic Chemistry*, Newman, M. S., ed., Wiley, New York (1956) p. 556–675

[6] Hansch, C., and Leo, A., *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York, 1979

[7] Bowden, K., Electronic Effects in Drugs. In: *Comprehensive Medicinal Chemistry, Vol. 4, Quantitative Drug Design*. Ramsden, C.A., ed., Pergamon Press, New York, 1990, p. 205–239

[8] Ehrenson, S., Brownlee, R. T. C., and Taft, R. W., *Prog. Phys. Org. Chem.* **10**, 1–80 (1973)

[9] Brown, H. C., and Okamoto, Y., *J. Am. Chem. Soc.* **80**, 4979–4987 (1958).

[10] Taft, R. W., *J. Phys. Chem.* **64**, 1805–1815 (1960)

[11] van Bekkum, H., Verkade, P. E., and Wepster, B. M., *Recl. Trav. Chim. Pays-Bas* **78**, 815–850 (1959)

[12] Taft, R. W., and Lewis, I. C., *J. Am. Chem. Soc.* **80**, 2436–2443 (1958)

[13] Taft, R. W., and Lewis, I. C., *J. Am. Chem. Soc.* **81**, 5343–5352 (1959)

[14] Dewar, M. J. S., and Grisdale, P. J., *J. Am. Chem. Soc.* **84**, 3548–3553 (1962)

[15] Dewar, M. J. S., Golden, R., and Harris, J. M., *J. Am. Chem. Soc.* **93**, 4187–4195 (1971)

[16] Swain, C. G., and Lupton, E. C., *J. Am. Chem. Soc.* **90**, 4328–4337 (1968)

[17] Roberts, J. D., and Moreland, W. T., *J. Am. Chem. Soc.* **75**, 2167–2173 (1953)

[18] van de Waterbeemd, H., and Testa, B., The Parameterization of Lipophilicity and other Structural Properties in Drug Design. In: *Advances in Drug Research*, Vol. **16**, Testa, B., ed., Academic Press, New York (1987) p. 85–225

[19] Hansch, C., and Fujita, T., *J. Am. Chem. Soc.* **86**, 1616–1626 (1964)

[20] Hansch, C., and Leo, A. J., *Substituent Constants for Correlation Analysis in Chemistry and Biology*, John Wiley & Sons, New York, 1979

[21] Leo, A. J., Methods of Calculating Partition Coefficients. In: *Comprehensive Medicinal Chemistry. Vol. 4., Quantitative Drug Design*, Hansch, C., Sammes, P. G., and Taylor, J. B., eds., Pergamon Press New York (1990) p. 295–320

[22] Braumann, T., *J. Chromat.* **373**, 191–225 (1986)

[23] Hansch, C., and Leo, A. J., *Correlation Analysis in Chemistry and Biology*, American Chemical Society, Washington, DC, 1994

[24] Leo, A. J., *Chem. Rev.* **93**, 1283–1306 (1993)

[25] Nys, G. G., and Rekker, R. F., *Chim. Ther.* **8**, 521–529 (1973)

[26] Silipo, C., Vittoria, A., Three-Dimensional Structure of Drugs. In: *Comprehensive Medicinal Chemistry, Vol.* **4**, *Quantitative Drug Design*. Hansch, C., Sammes, P. G., and Taylor, J. B., eds., Pergamon Press, New York (1990) p. 154–204

[27] Charton, M., *J. Am. Chem. Soc.* **41**, 1976, p. 2217–2220

[28] Exner, O., *Correlation Analysis of Chemical Data*, Plenum Press, New York, 1988

[29] Charton, M., *J. Org. Chem.* **41**, 2217–2220 (1976)

[30] Charton, M. T., *Curr. Chem.* **114**, 59–91 (1983)

[31] Verloop, A., Hoogenstaaten, W., and Tipker, J., Development and Application of New Steric Substituent Parameters in Drug Design. In: *Drug Design*, Vol. **VII**, Ariëns, E. J., ed., Academic Press, New York (1976) p. 165–207

[32] Verloop, A., *The STERIMOL Approach to Drug Design*, Marcel Dekker, New York, 1987

[33] Verloop, A., and Tinker, J., Physical Basis of Sterimol and Related Steric Constants. In: *Pharmaco Chemistry Library, QSAR in Drug Design and Toxicology*, Vol. **10**, Hadzi, D., and Jerman-Blazic, B., eds., Elsevier, Amsterdam (1987) p. 97–102

[34] Fujita, T., and Iwamura, H., Application of Various Steric Constants to Quantitative Analysis of Structure-Activity Relationships. In: *Steric Effects in Drug Design*, Charton, M., and Motoc, I., eds., Springer, Berlin (1983) p. 119–157

[35] Klopman, G., Balthasar, D. M., and Rosenkranz, H. S., *Environ. Tox. Chem.* **12**, 231–240 (1993)

[36] Rosenkranz, H. S., and Klopman, G., *Mutation Research* **228**, 105–124 (1990)

[37] Trinajstić, N., *Chemical Graph Theory*, 2nd ed., CRC Press, Boca Raton, FL, 1992

[38] Kier, L. B., Molecular Connectivity as a Descriptor of Structure for SAR Analyses. In: *Physical Chemical Properties of Drugs*, Yalkowsky, S. H., Sinkua, A. A., and Valvani, S. C. eds., Marcel Dekker, New York (1980) p. 277–320

[39] Wiener, H., *J. Am. Chem. Soc.* **69**, 17–22 (1947)

[40] Hosoya, H., *Bull. Chem. Soc. Japan* **44**, 2332–2337 (1971)

[41] Randić, M., *J. Am. Chem. Soc.*, **97**, 6609–6615 (1975)

[42] Kier, L. B., and Hall, L. H., *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1986

[43] Murray, W. J., Hall, L. H., and Kier, L. B., *J. Pharm. Sci.*, **64**, 1978–1982 (1975)

[44] Sabljić, A., and Horvatić, D., *J. Chem. Inf. Comput. Sci.* **33**, 292–295 (1993)

[45] Kier, L. B., *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.*, **4**, 109–116 (1985)

[46] Kier, L. B., *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.*, **5**, 1–7 (1986)

[47] Kier, L. B., *Quant. Struct.-Act. Relat. Pharmacol., Chem. Biol.*, **5**, 7–12 (1986)

[48] van de Waterbeemd, H., and Testa, B., *Adv. Drug. Res.* **16**, 85–225 (1987)

[49] Purcell, W. P., Bass, G. E., and Clayton, J. M., *Strategy of Drug Design*, Wiley, New York, 1973

[50] Martin, Y. C., The Quantitative Relationships Between pK_a, Ionization and Drug Potency: Utility of Model-Based Equations. In: *Physical Chemical Properties of Drugs*, Yalkowsky, S. H., Sinkula, A. A., and Valvani, S. C., eds., Marcel Dekker, New York (1980) p. 49–110

[51] Seydel, J. K., and Schaper, K. J., *Pharmacol. Ther.* **15**, 131–182 (1982)

[52] Lien, E. J., Guo, Z.-R., Li, R.-L., and Su, C.-T., *J. Pharm. Sci.* **71**, 641–655 (1982)

[53] Nelson, R. D., Lide, D. R., and Maryott, A. A., *Nat. Stand. Ref. Data Set, Natl. Bur. Stand.* No. 10 (1967)

[54] McClelland, A. L., *Table of Experimental Dipole Moments*, Freeman, San Francisco, 1963

[55] Smith, R. P., Ree, T., Magee, J. L., and Eyring, H., *J. Am. Chem. Soc.* **73**, 2263–2268 (1951)

[56] Abraham, R. J., and Smith, P. E., *J. Comp. Chem.* **9**, 288–297 (1988)

[57] Dixon, S. L., and Jurs, P. C., *J. Comp. Chem.* **13**, 492–504 (1992)

[58] Pople, J. A., and Segal, G. A., *J. Chem. Phys.* **44**, 3289–3296 (1966)

[59] Dewar, M. J. S., and Thiel, W., *J. Am. Chem. Soc.* **99**, 4899–4907 (1977)

[60] Dewar, M. J. S., Zoebisch, E. G., Healy, E. F., and Stewart, J. J. P., *J. Am. Chem. Soc.* **107**, 3902–3909 (1985)

[61] Stewart, J. J. P., *J. Comp. Chem.* **10**, 209–220 (1989)

[62] Atkins, P. W., *Physical Chemistry*, Freeman, New York, 1982

[63] Miller, K. J., *J. Am. Chem. Soc.* **112**, 8533–8542 (1990)

[64] Vogel, A. I., Cresswell, W. T., Jeffery, G. H., and Leicester, J., *J. Chem. Soc.* 514–549 (1952)

[65] Hansch, C., Leo, A., Unger, S. H., Kim, K. H., Nikaitani, D., and Lien, E., *J. Med. Chem.* **16**, 1207–1222 (1973)

[66] Fujita, T., Nishioka, T., and Nakajima, M., *J. Med. Chem.* **20**, 1071–1081 (1977)

[67] Kamlet, M. J., and Taft, R. W., *J. Am. Chem. Soc.* **98**, 377–383 (1976)

[68] Kamlet, M. J., and Taft, R. W., *J. Am. Chem. Soc.* **98**, 2886–2894 (1976)

[69] Streitweiser, A., *Molecular Orbital Theory for Organic Chemists*, Wiley, New York, 1961

[70] Kier, L. B., *Molecular Orbital Theory in Drug Research*, Academic Press, New York, 1971

[71] Bondi, A., *J. Phys. Chem.* **68**, 441–451 (1964)

[72] Pearlman, R. S., Molecular Surface Areas and Volumes and Their Use in Structure-Activity Relationships. In: *Physical Chemical Properties of Drugs*, Vol. **10**. Yalkowsky, S. H., Sikula, A. A., and Valvani, S. C., eds., Marcel Dekker, New York, 1980, p. 321–345

[73] Pearlman, R. S., *Quantum Chem. Prog. Exchange Bull.* **1**, 15–16 (1981)

[74] Pearlman, R. S., Molecular Surface Area and Volume: Their Calculations and Use in Predicting Solubilities and Free Energies of Dissolution. In: *Partition Coefficient Determination and Estimation*, Dunn, W. J., Block, J. H., and Pearlman, R. S., eds., Pergamon Press, New York (1986) p. 3–10

[75] Pearlman, R. S., Yalkowsky, S. H., and Banerje, S. J., *J. Phys. Chem. Ref. Data* **13**, 555–562 (1984)

[76] Camilleri, P., Watts, A., and Boraston, J. A., *J. Chem. Soc. Perkin Trans.* II 1699–1707 (1988)

[77] Amidon, G. L., Yalkowsky, S. H., Anik, S. T., and Valvani, S. C., *J. Phys. Chem.* **21**, 2239–2246 (1975)

[78] Koehler, M. G., Grigoras, S., and Dunn III, W. J., *Quant. Struct.-Act. Relat.* **7**, 150–159 (1988)

[79] Dunn, W. J., Koehler, M. G., and Grigoras, S., *J. Med. Chem.* **30,** 1121–1126 (1987)

[80] Bondi, A., *J. Phys. Chem.* **68**, 441–451 (1964)

[81] Harris, M. J., Higuchi, T., and Rything, J. H., *J. Phys. Chem.* **77**, 2694–2703 (1973)

[82] Lee, B., and Richards, F. M., *J. Mol. Biol.* **55**, 379–400 (1971)

[83] Hermann, R. B., *J. Phys. Chem.* **76**, 2754–2759 (1972)

[84] Pearlman, R., *SAREA* Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, Program #432

[85] Grigoras, S., *J. Comp. Chem.* **11**, 493–510 (1990)

[86] Howell, J., Rossi, A., Wallace, D., Haraki, K., and Hofman, R., *FORTICON 8* Quantum Chemistry Program Exchange, University of Indiana, Bloomington, IN, Program #469

[87] Stanton, D. T., and Jurs, P. C., *Anal. Chem.* **62** 2323–2329 (1990)

[88] Stuper, A. J., Brugger, W. E., and Jurs, P. C., *Computer-Assisted Studies of Chemical Structure and Biological Function*, Wiley, New York, 1979

[89] Dixon, S. L., and Jurs, P. C., *J. Comp. Chem.* **13**, 492–504 (1992)

[90] Abraham, R. J., and Smith, P. E., *J. Comp. Chem.* **9**, 288–297 (1988)

[91] Stanton, D. T., and Jurs, P. C., *J. Chem. Inf. Comput. Sci.* **32**, 109–115 (1992)

[92] Anker, L. S., Jurs, P. C., and Edwards, P. A., *Anal. Chem.* **62**, 2676–2684 (1990)

[93] Stanton, D. T., and Jurs, P. C., *Anal. Chem.* **61**, 1328–1332 (1989)

[94] Stanton, D. T., Jurs, P. C., and Hicks, M. G., *J. Chem. Inf. Comput. Sci.* **31**, 301–310 (1991)

[95] Stanton, D. T., Egolf, L. M., Hicks, M. G., and Jurs, P. C., *J. Chem. Inf. Comput. Sci.* **32**, 306–316 (1992)

[96] Egolf, L. M., and Jurs, P. C., *J. Chem. Inf. Comp. Sci.* **33**, 616–625 (1993)

[97] Egolf, L. M., and Jurs, P. C., in preparation

[98] Egolf, L. M., and Jurs, P. C., *Ind. Eng. Chem. Res.* **31**, 1798–1807 (1992)

[99] Russell, C. J., Dixon, S. L., and Jurs, P. C., *Anal. Chem.* **64**, 1350–1355 (1992)

[100] Egolf, L. M., and Jurs, P. C., *J. Chem. Inf. Comp. Sci.*, in press

2.2 Atom-Level Descriptors for QSAR Analyzes

Lemont B. Kier

Abbreviations and Symbols

δ	Count of bonded atoms other than hydrogen
d_i^v	Count of valence electrons other than those bonding to hydrogen
E-state	Electrotopological state
I	Intrinsic value
log *MAC*	log of the minimum anesthetic concentration
MAO	Monoamine oxidase
N	Principle quantum number
pI_{50}	Inhibitory potency
S_i	Electropological state of atom *i*
X_{KH}	Kier/Hall electronegativity
X_{m}	Mulliken-Jaffe valence state electronegativity

2.2.1 Introduction

The non-empirical molecular descriptors such as molecular connectivity [1, 2] and the *kappa* shape indices [3] have served us well in the creation of models, relating structure to biological activity (see Chapt. 2.1). These models define path fragments of importance to the encoding of salient molecular features governing a measured activity. Numerous examples have revealed the value of this paradigm [4–6]. In spite of these successes, we are aware of the generally held view that atom-level parts of molecules are the critical ingredients in meaningful drug-receptor or drug-enzyme interactions. It is the atom or the group, which engages a complementary receptor feature to initiate a chain of events leading to an effect.

Somehow, the significance of this concept has escaped many of the early developers of structure-activity quantitation, as they have laid heavy emphasis on physical properties to model the contribution of the whole molecule to a biological activity. The developers of topological indices have also neglected this reality and have concentrated on the entire molecule in their quest to encode the structure.

The early interest in molecular orbital indices to quantitate atom contribution to activity [7] was a recognition of the importance of an atom or group in this process. Calculations of atom charges using several levels of rigor have been considered. From these indices, molecular electrostatic potential maps have been calculated to encode atom-level information. The concept of molecular fragmentation to encode local information such as lipophilicity is another such area of study.

At the level of rigor in topological indices, there is clearly a need for atom level indices, reflecting both the electronic environment near an atom (group) and nearby topological state. This need fueled the development of electrotopological state indices by Hall and Kier [10–13] over the last few years. The train of thought leading to this development began in 1987 when Kier, [8] seeking a way to identify topologically equivalent atoms for use in the Shannon information theory equation, proposed an atom index in which all atom pairs, $\delta_i^v\delta_j^v$, were identified within a molecule (see Chapter 3 for definition of *S* values). A numerical index was derived by taking the geometric mean of each product. For each atom, *i*, summation of these $(\delta_i^v\delta_j^v)^{1/2}$ terms over all pairs in the molecule gave rise to a relative topological state for that atom in that molecule. Hall [12], reported that this index gave unique values for a large number of test cases. This index could, thus identify topologically equivalent atoms in any molecule with the aid of a computer. The Shannon equation can thus be calculated entirely by computation, without the need for external atom equivalence recognition. In a later, more detailed study Hall and Kier [9] sought to improve the uniqueness of this atom index. Several algorithms were examined, which were all based on the geometric mean of the valence *delta* values of the atoms in each path from an atom, *i*. The uniqueness was, thus, greatly improved in this study. Subsequent studies have produced an electropological state index, encoding both electronic and topological information in a unified attribute index for atoms in molecules [10–13]. This work has been reviewed recently [14]. We shall develop the concept and form of the electrotopological state index in this article followed by some recent examples of applications.

2.2.2 An Atom-Level Description of Structure

2.2.2.1 The Field

The attributes of atoms or groups in a molecule that engage a receptor or an enzyme active site must certainly be electronic and topological in character. There is indeed little else apart from these features that would attract our interest. We have developed the view that an atom in a molecule is part of a field of information relating to electronic influences and topological environment [10–14]. This field is an environment that can cause two methyl groups in a molecule to be very different or identical. This field produces changes in the state of an atom or group, when changes in the molecule are introduced. If we can quantify the influence of this field on any atom, then we have the opportunity to relate this influence to the biological performance of the molecule. More specifically, we have an opportunity to identify those atoms within the molecule, which are exhibiting field-induced changes, and which correlate with a biological response. The goal of our research to date has been the quantification of the principle ingredients in this field, i.e. the electronic and the topological influences on atoms.

The quantification of the influence of this field on an atom is dependent upon three components. The first is the attribute associated with each atom, which we call the intrinsic state of that atom. This is the quantitation of the composition, hybrid state, topology and hydride state of the atoms (groups) in isolation. The second component is the quantification of the field effect which is the influence of one atom on another within the molecule. Finally, we must include the information concerning the separation or distance between any two atoms in a convenient metric. We will briefly review each of these components leading to an index defining the state of an atom in a molecule.

2.2.2.2 The Intrinsic State of an Atom

An intrinsic state of an atom in a molecule encodes the basic information associated with that atom regardless of its environment. Because we ultimately want to account for the electronic and topological influences of one atom on another within a molecule, it is apparent that these two attributes must be encoded into the intrinsic state. We further require that the molecule is represented as a chemical graph, in which the hydrogen atoms have been suppressed, and is the familiar skeleton representation of a molecule.

The electronic influence is most conveniently summarized into a value, which reflects the electronegativity of an atom or group. Since we are considering skeleton or chemical graph representations of molecules, we might turn to our previous work on molecular connectivity for guidance in quantitating this attribute. In 1981 Kier and Hall [15] found a close relationship between the two molecular connectivity *delta* values and the Mulliken-Jaffe valence state electronegativity, X_M [17]. This relationship is approximately:

$$X_M = 2(\delta^v - \delta) + 7 \qquad (1)$$

for second row atoms, where δ^v is the number of valence electrons on an atom in a chemical graph (excluding those bonding to hydrogen) and δ is the number of sigma electrons from that atom (excluding those bonding to hydrogen). The equation explains 98% of the variation in the X_M value.

An interpretation of this relationship is that $\delta^v - \delta$ is simply the number of pi and lone pair electrons on an atom in a molecule. Kier and Hall [15] reconciled this relationship by invoking the reduced shielding of the core induced by a pi or lone pair of electrons relative to a sigma bonding electron on that same atom. Another form of the Kier/Hall electronegativity, X_{KH}, which is useful in student lectures is:

$$X_{KH} = \frac{\text{Periodic Table column No.} - \text{Number of sigma bonds}}{(\text{Periodic Table row No.})^2} \qquad (2)$$

Using the *delta* values, we can define an intrinsic atom state as a function of electronegativity $(\delta^v - \delta)$ and of topology. The topology is certainly encoded in the simple *delta* value as an index reflecting the number of adjacent atoms. An initial statement of the intrinsic state can be expressed by:

$$I = (\delta^v - \delta)/\delta \qquad (3)$$

Calculations of the various atoms and their hydride groups, using this expression, reveal that $\delta^v = \delta$ for alkane hydride groups, thus, these would have redundant values. This is simply dealt with by modifying the expression with a constant:

$$I = (\delta^v - \delta + 1)/\delta \tag{4}$$

By adding 1 to this expression we can simplify it further to:

$$I = (\delta^v + 1)/\delta \tag{5}$$

To account for the diminished electronegativity of atoms in higher quantum levels, an addition to the intrinsic state must be considered. Several possibilities exist, but the one chosen was the modification of the δ^v value to reflect the principal quantum number, N, relative to the value of 2 (the N value for C, N, O, F). The general expression for δ^v in the general equation (5) is $(2/N)^2\ \delta^v$ and Eq. (5) becomes:

$$I = [(2/N)^2\ \delta^v) + 1]/\delta \tag{6}$$

Calculated values using Eq. (6) are shown in Table 1. An inspection of this table reveals that the electronic and topological information is reflected in the I values. As an atom (group) becomes more electron-rich in terms of valence electrons, the value of I increases. As the atom becomes more "buried" in the molecule (as opposed to having mantle status) the value of I decreases. This is an acceptable definition of the intrinsic state of an atom encoding both electronic and topological attributes.

Table 1. Intrinsic state values

Atom (skeletal hydride group)	I $[(\delta^v + 1)/\delta]$
$>C<$	1.250
$>CH-$	1.333
$-CH_2-$	1.500
$>C=$	1.667
$-CH_3$, $=CH-$, $>N-$	2.000
$\equiv C-$, $-NH-$	2.500
$=CH_2$, $=N-$	3.000
$-O-$	3.500
$\equiv CH$, $-NH_2$	4.000
$=NH$	5.000
$=N$, $-OH$	6.000
$=O$	7.000
$-F$	8.000
$-Cl$	4.111
$-Br$	2.750
$-I$	2.120
$=S$	3.667
$-SH$	3.222
$-S-$	1.833

2.2.2.3 The Field Effect on Each Atom

The second ingredient in an atom-level index must describe the field effect on each atom. Stated another way, this contribution must encode information about the interaction and relative perturbation that each atom contributes to the electronic and topological attributes of every other atom. This perturbation, ΔI, can take many forms, but the one we have chosen is based on the intrinsic states themselves, which are sources of perturbation. The simplest form would be the difference between any two I values relative to atom i, summed over the entire molecule.

Thus, we have

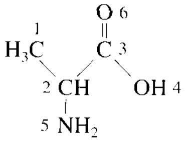
$$\Delta I = \Sigma(I_i - I_j) \tag{7}$$

as a source of the perturbation.

This expression is not complete, in that the distance between i and j is not stated, but is nevertheless highly relevant. This third component is reflected in the number of atoms, separating and including i and j in the chemical graph. The graph distance, r, is included and is expressed to the second power, however, the program MOLCONN-X [16] permits the power to be varied. The final expression of the

Table 2. Electrotopological state calculations for alanine

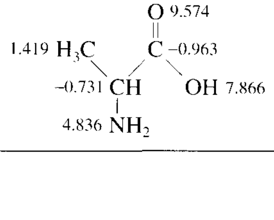
Atom Numbering:



Intrinsic Values: $I(1) = 2.000$ $I(4) = 6.000$
$I(2) = 1.333$ $I(5) = 4.000$
$I(3) = 1.667$ $I(6) = 7.000$

	$(I_j - I_j)/r^2$ Matrix						$\Delta I =$
i			j				
	1	2	3	4	5	6	row sum
1	0.0	0.1667	0.0370	−0.2500	−0.2222	−0.3125	−0.5810
2	−0.1667	0.0	−0.0833	−0.5185	−0.6667	−0.6296	−2.0648
3	−0.0370	0.0833	0.0	−1.0833	−0.2593	−1.3333	−2.6296
4	0.2500	0.5185	1.0833	0.0	0.1250	−0.1111	1.8657
5	0.2222	0.6667	0.2593	−0.1250	0.0	−0.1875	0.8356
6	0.3125	0.6296	1.3333	0.1111	0.1875	0.0	2.5741
							0.0000

$S_i = I_i + \Delta I_i$



field perturbation of the intrinsic state is:

$$\Delta I = \Sigma(I_i - I_j)/r^2 \tag{8}$$

The field effect, ΔI, modifies the intrinsic state, producing a configuration or state value which we call the electrotopological state, S_i:

$$S_i = I + \Delta I \tag{9}$$

This, in its abbreviated form, is called the E-state value of an atom. An example of this calculations is shown in Table 2.

2.2.3 Strategies for Use of E-State Indices

The E-state indices reflect the electronic and topological state of atoms and groups in the molecule. These states change as a function of the distance and of the state of other atoms along the chemical graph path, which separates two atoms. In a series of molecules, with a core component remaining constant and some other part varying in structure, it is possible to quantify in relative terms the extent of the through-graph influence on atoms and the focus of this influence. In QSAR analyzes this is a powerful tool for gaining insight into which structural changes in a chemical series are influencing the rest of the molecule. This information, plus the quantitation of this influence, produces a potential for identifying atoms or groups, which are responsible for a measured biological activity. The ultimate aim is that drug design may proceed on a rational basis with such an insight.

The method described here is more effective, if the molecules in a series have more than one substituent. This makes it possible to avoid extensive intercorrelation of influences experienced by single substituent series. It is possible to group nearby structural features or fragments and to identify them as possible salient features, which could be influenced by the substitution patterns in the rest of the molecule.

2.2.4 Examples of E-state QSAR

Over the past five years a number of applications of E-state analysis have been reported in the literature. A brief survey of these studies is presented here to demonstrate the utility and breadth of application with this atom-level index.

2.2.4.1 MAO Inhibition with Hydrazides

Hall, Mohoney and Kier [11] have reported a study on a series of aryloxyacetohydrazides as potential monoamine oxidase (MAO) inhibitors [18]. The E-state indices correlating with the activity were assigned to the respective atoms of the parent structure given in Fig. 1. A comparison between the E-state indices and the molecular

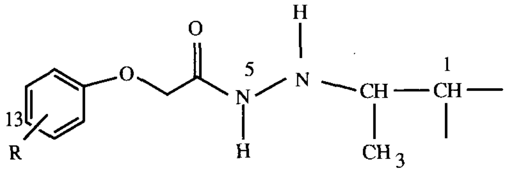


Figure 1. The aryloxyacetohydrazide parent structure for MAO inhibitors.

orbital parameters for these same atoms using the AM1 Hamiltonian, showed that the E-state values were significantly better in modeling a structure-activity relationship. The atoms implicated by the E-state analysis are the same as those implicated using the less successful molecular orbital model. The equation, relating the E-state indices to the inhibitory potency, pI_{50} is:

$$\begin{gathered} pI_{50} = 1.69S_1 - 9.14S_5 + 0.15S_{13} + 32.15 \\ PRESS\, r^2 = 0.848\,, \quad s = 0.23\,, \quad n = 24 \end{gathered} \tag{10}$$

2.2.4.2 Adenosine A_1 Inhibitors

Joshi and Kier [19] using E-state indices, analyzed a series of xanthines reported by Jacobson et al. [20] to be inhibitors of adenosine A_1. Analysis of the ring atoms and substituents using the E-state indices revealed a good correlation with affinity, $\log K_1$, as modeled by the equation:

$$\begin{gathered} \log K_1 = -1.17S_7 - 0.97S_{10} - 0.22S_{12} + 1.71 \\ r^2 = 0.88\,, \quad s = 0.33\,, \quad n = 28 \end{gathered} \tag{11}$$

The atoms implicated in this QSAR analysis are shown in Fig. 2.

2.2.4.3 Anesthetic Concentration of Haloalkanes

In a study of the anesthetic effect of several haloalkanes reported by Larsen [21], Tsantili-Kakoulidou, Kier and Joshi reported an E-state analysis of several derivatives [22]. The series of molecules analyzed were $CXYZ\text{-}CF_3$ analogues where XYZ were halogen or hydrogen atoms. A good correlation was found between the log of the minimum anesthetic concentration, log *MAC* and the sum of the E-state

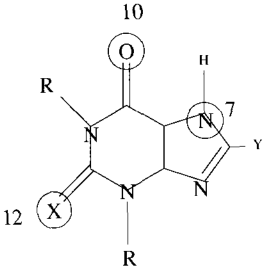


Figure 2. The parent xanthine structure for adenosine A_1 inhibitors.

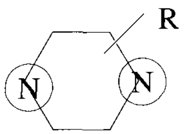


Figure 3. The parent pyrazine structure for odor threshold analysis.

values for the -CF$_3$ group. The summation of two E-state indices to produce a single value was initiated in our earlier work [11]. Van de Waterbeemd has considered this idea in a recent article [23]. This approach is important to the development of group E-state indices.

$$\begin{gathered}\log MAC = 33.19 - 1.01S_{\mathrm{CF_3}} \\ r^2 = 0.892\,, \quad s = 0.23\,, \quad n = 11\,.\end{gathered} \tag{12}$$

2.2.4.4 Odor Sensitivity of Pyrazines

Tsantili-Kakoulidou and Kier [24] have analyzed a series of alkyl substituted pyrazines using E-state indices for the ring atoms. A close correlation was found between the sum of nitrogen atom E-state values, S_{N}, (see Fig. 3) and the odor threshold concentration, log *ppb*.

$$\begin{gathered}\log ppb = 94.87S_{\mathrm{N}} - 13.17S_{\mathrm{N}}^2 - 165.39 \\ r^2 = 0.979\,, \quad s = 0.25\,, \quad n = 13\end{gathered} \tag{13}$$

2.2.5 Conclusions

The electrotopological state (E-state) method is a new and powerful approach to encoding information about an atom in its molecular environment. The relative perturbation of each atom, as molecular structure is varied in a series, is quantitated in a manner suitable for QSAR analysis. Atoms (or groups) implicated in a biological event may be identified in this analysis, if the data set has been constructed with this in mind. The information generated lends itself ideally to the design of new molecules, since structural influences are easily identified, quantitated and translated into structural changes. The studies utilizing the E-state paradigm are increasing with extended applications becoming more prominent. Several new innovations are being developed and will be described in later reports. These will enhance the ability of E-state analysis to promote theoretical approaches in molecular design.

References

[1] Kier, L. B., and Hall, L. H., *Molecular Connectivity in Chemistry and Drug Research*, Academic Press, New York, 1976

[2] Kier, L. B., and Hall, L. H., *Molecular Connectivity in Structure-Activity Analysis*, John Wiley, London, 1986

[3] Kier, L. B., *Med. Res. Revs.* **7**, 417–440 (1987)

[4] Hall, L. H. and Kier, L. B., *Eur. J. Med. Chem.* **6**, 399–405 (1981)

[5] Hall, L. H., and Kier, L. B., *J. Molec. Struct.* **134**, 309–315 (1986)

[6] Hall, L. H., and Kier, L. B., *Environ. Tox. and Chem.* **8**, 431–437 (1989)

[7] Kier, L. B., *Molecular Orbital Theory in Drug Research*, Academic Press, New York, 1971

[8] Kier, L. B., *Quant. Struct. – Act. Relat.* **6**, 8–12 (1987)

[9] Hall, L. H., and Kier, L. B., *Quant. Struct. – Act. Relat.* **9**, 115–131 (1990)

[10] Kier, L. B., and Hall, L. H., *Pharm. Res.* **7**, 801–807 (1990)

[11] Hall, L. H., Mohoney, B., and Kier, L. B., *Quant. Struct. – Act. Relat.* **10**, 43–51 (1991)

[12] Hall, L. H., Mohoney, B., and Kier, L. B., *J. Comp. Inf. Comp. Sci.* **31**, 76–82 (1991)

[13] Kier, L. B., Hall, L. H., and Frazer, J. W., *J. Math. Chem.* **7**, 229–241 (1991)

[14] Kier, L. B., and Hall, L. H., *Adv. Drug Res.* **22**, 1–38 (1992)

[15] Kier, L. B., and Hall, L. H., *J. Pharm. Sci.* **70**, 583–589 (1981)

[16] Program MOLCONN-X is available from Hall Associates Consulting, 2 Davis St., Quincy, MA 02170 USA

[17] Hinze, J., and Jaffe, H. H., *J. Am. Chem. Soc.* **84**, 540–552 (1962)

[18] Fulcrand, P., Berge, G., Noel, A. M., Chevallet, P., Castel, J. and Orzaleski, H., *Eur. J. Med. Chem.* **13**, 177–182 (1978)

[19] Joshi, N., and Kier, L. B., *Med. Chem. Res.* **1**, 409–416 (1992)

[20] Jacobson, K. A., Kiriasis, L. Barone, S., Bradbury, B. J., Udai, K., Campagne, J. M., Secunda, S., Daly, J. W., Neumeyer, J. L., and Pfleiderer, W., *J. Med. Chem.* **32**, 1873 (1989)

[21] Larsen, E. R., *Fluorine Chem. Rev.* **3**, 1–6 (1959)

[22] Tsantili-Kakoulidou, A., Kier, L. B., and Joshi, N., *J. Chem. Phys.* **89**, 1729–1733 (1992)

[23] van de Waterbeemd, H., Carrupt, P.-A., Testa, B., and Kier, L. B., *Multivariate Data Modeling of New Steric, Topological and CoMFA-Derived Substituent Parameters.* In: *Trends in QSAR and Molecular Modelling 92.* Wermuth, C. G. (ed.), Escom, Leiden (1993) p. 69–75

[24] Tsantili-Kakoulidou, A., and Kier, L. B., *Pharm. Res.* **9**, 1321–1323 (1992)

Chemometric Methods in Molecular Design
edited by Han van de Waterbeemd
© VCH Verlagsgesellschaft mbH, 1995

3 Experimental Design in Synthesis Planning and Structure-Property Correlations

3.1 Experimental Design

Volkhard Austel

Abbreviations and Symbols

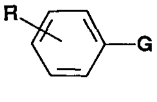
HIP	Hypersurface iterative projection
MR	Molar refractivity (parameter of size or polarity)
π	Hansch-Fujita value (parameter of lipophilicity)
σ	Hammett constant (parameter of electronic properties)

3.1.1 The Importance of Experimental Design in Medicinal Chemistry

The number of compounds that is synthesized and tested for every new chemical entity introduced onto the market is rising steadily. At present, estimates range from 10000 to 20000. As the resources for chemical synthesis and biological testing are limited, there is an urgent need for preventing, or at least slowing down, further increases in the number of compounds being synthesized.

In theory, the most promising way in which this might be achieved, is to investigate the causes of diseases and the possibilities of intervention at the molecular level and by using this insight for designing test compounds. In most cases of interest, the available information is, however, not detailed enough for deriving structure-activity relationships that would allow sufficiently potent compounds to be designed more directly. More informative structure-activity relationships are, therefore, required, which at present can only be derived empirically, i.e. with sets of test compounds.

Depending on the composition of such sets the average structure-activity information per compound can vary greatly. Consider for example the set shown in Fig. 1 whose elements are characterized by a common pharmacophore, consisting of a basic skeleton G, to which a variously substituted phenyl ring is attached. The structure-activity information per compound, obtained from this set, is com-



R = -H, 4-OCH_3, 4-OC_2H_5, 4-CH_3, 4-F, 4-SCH_3

Figure 1. Example of an uneconomical set of test compounds.

paratively low, since only two of these compounds (e.g., R = H and $-OC_2H_5$) adequately represent the whole set. In addition, potentially important factors such as steric interactions between the phenyl ring and G, e.g. with an *ortho* substitutent, or the influence of hydrophilic or strongly electron withdrawing groups are not addressed at all.

It is clearly uneconomical to work with such data sets, and the problems encountered with these can, however, be avoided by carefully designing the test sets. To this ends methods for experimental design have been developed in the past decades.

This chapter will be confined to those methods that in my opinion might be particularly useful for the bench chemist and that could be (but need not be) applied qualitatively without computational input.

3.1.2 Strategies in Experimental Design

The experimental design methods that have been proposed in the literature can be devided into two categories:
1) Methods which are aimed at a direct and, therefore, (supposedly) quick optimization of lead compounds, and
2) methods that provide a strong basis for deriving reliable structure-activity relationships.

Methods which belong to the first category are only suitable in the final stages of an optimization procedure, for which reliable fundamental structure-activity relationships are already available and which require a certain modification. Typical examples would be the optimization of substitution patterns of aromatic rings or of aliphatic chains. In addition, these methods frequently give rise to biased results, since they clearly do not put structure-activity relationships to the test. Such a procedure is frequently chosen in medicinal chemistry in order to reduce the experimental input. However, this always entails the risk of overlooking interesting routes for a lead optimization or even for the discovery of new leads. In the latter stages of an optimization, this risk is comparatively minor, but it may, however, become more significant if the structural features that are essential for a particular activity have not yet been fully elucidated.

The methods which belong to the second category in principle do not need any prior structure-activity information and are, therefore, applicable at any stage of the search for new drugs. These methods, in particular, should be used for deriving qualitative and quantitative structure-activity relationships. In practice structure-activity relationships are relevant for determining those structural requirements which give rise to sufficient potency. This also includes the determination of bulk areas which may become an important feature in the optimization of pharmacokinetic and metabolic properties or for achieving selectivity.

The difference between the two categories becomes evident if one considers optimization on a more abstract level: In order to interact with its target, e.g. a receptor or an enzyme, a drug molecule must be able to present an appropriate pattern of physico-chemical properties in the correct spatial arrangement. This pattern is usually described in an indirect manner with the aid of molecular descriptors (see Chap. 2). Every one of these descriptors can be associated with one of the components of an n-dimensional *parameter space*. The compounds of a test set are represented by points in this space. If the descriptors have been correctly chosen, the active compounds will be concentrated in only one or, at least in a few localized areas in the corresponding parameter space.

Methods that are aimed at a direct optimization usually cover only a limited area of parameter space surrounding a previously identified active compound. Therefore, active compounds that are located in other areas cannot be detected.

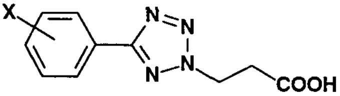
Such disadvantages can be significantly reduced by applying methods belonging to the second category, which allow large areas of the parameter space to be investigated in a systematic manner. However, this method is usually at the cost of greater synthetic efforts. In practice, one can reduce the experimental effort by using the different densities of data points (test compounds) in different areas of the parameter space, i.e. a higher density surrounding active compounds than in the other parts of the parameter space.

3.1.3 Selected Methods for Experimental Design

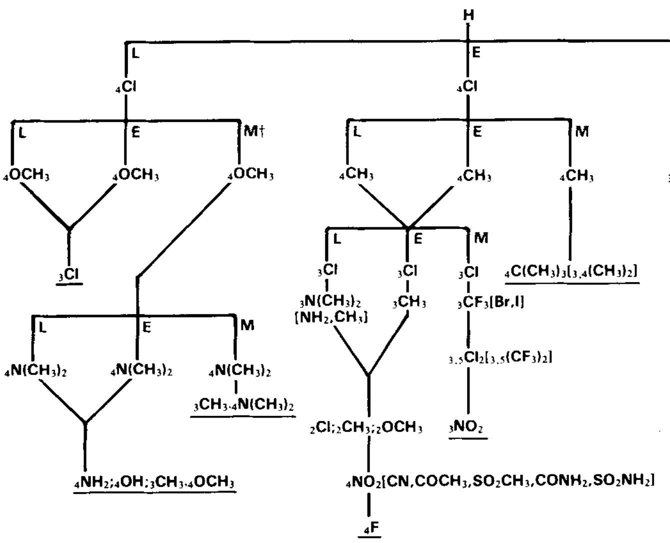
3.1.3.1 Methods for the Direct Optimization of Lead Compounds

Of the many procedures for direct lead optimization, three are particularly suited for the purposes of the bench chemist. One of these methods uses operational schemes, known as *Topliss trees* [1], which have been designed for substituents on aromatic rings (Fig. 2) and for modifying aliphatic chains (Fig 3).

In the case of the aromatic substitution, one first compares the unsubstituted compound with the 4-chloro derivative. If the latter is more active, one continues by preparing the 3,4-dichloro derivative. Should this lead to a further increase in activity then the 3-CF_3, 4-Cl and the 3-CF_3, 4-NO_2 derivatives can be considered as candidates for maximal activity. The other branches of the tree are followed analogously. As an example, with phenyl tetrazoles of type (**1**), the unsubstituted compound showed a higher anti-inflammatory activity than the 4-chloro derivative. Consequently, the chlorine was replaced by 4-methoxy, which reduced activity even further. Under these circumstances the scheme suggests the 3-chloro derivative, which in the present example, was indeed the most active compound (along with the 5-bromo derivative).



(**1**)

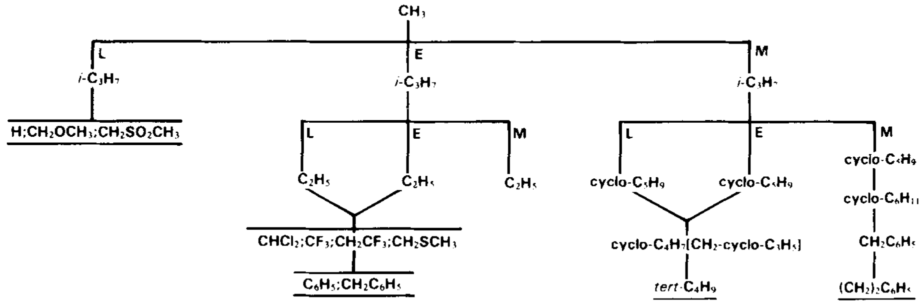


M = More active, E = equiactive, L = less active. Descending lines indicate sequence. Square brackets indicate alternate
†Compared to 4-H compound.

Figure 2. Operational scheme for the optimization of aromatic substitution patterns (reprinted with permission from Ref. [1] Copyright 1972, American Chemical Society).

The scheme for modifying aliphatic chains can also be applied analogously.

The second procedure for optimizing the substitution on aromatic rings, also suggested by Topliss [2], begins with a set of five compounds which consists of the unsubstituted compound and the 4-chloro, 4-methyl, 4-methoxy and 3,4-dichloro derivatives. The relative activities of these derivatives are considered indicative of a particular quantitative dependence on electronic properties (represented by Hammet σ constants) and lipophilicity (Hansch-Fujita π values). For example, the order 3,4-Cl_2 > 4-Cl or 4-CH_3 > 4-OCH_3 > H is assumed to signify that the biological response is dependent on the term $(2\pi - \sigma)$. On the basis of this relationship,

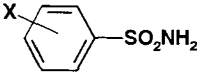


M = More active, E = equiactive, L = less active. Descending lines indicate sequence. Square brackets indicate alternates.

Figure 3. Operational scheme for modifications of aliphatic chains (reprinted with permission from Ref. [1], Copyright 1972, American Chemical Society).

additional substitution patterns have been proposed that might possibly improve activity. In the present example 4-*i*-Pr, 4-*t*-Bu, 3,4-di-Me, 4-O-*n*-Bu, 4-O-Bz, and 4-N(Et)$_2$ were proposed as additional candidates.

An illustration using literature data comes from the inhibition of carbonic anhydrase by compounds of the general structure (**2**). The ranking of the primary compounds was found to be 3,4-Cl$_2$ > 4-Cl > 4-CH$_3$ > 4-OCH$_3$, H suggesting a ($\pi + \sigma$) relationship. The corresponding additional substitution patterns comprise 3-CF$_3$, 4-NO$_2$ which was found to be the most active of the reported compounds.



(2)

Other rankings and the corresponding relationships as well as additional substitution patterns are also given by Topliss [2].

A third method, which uses the *sequential simplex* technique, was introduced into medicinal chemistry by Darvas [3] and was developed further by Gilliom et al. [4]. This method begins with a lead compound and as many analogs of this compound as there are parameters to be considered. In an *n*-dimensional parameter space, the point corresponding to the least active compound of this set, is reflected through the center of gravity of the remaining points. A new analog is designed, so that its corresponding point in parameter space is located as closely as possible to the point of the reflection. The least active compound of the original set is discarded and the operation is repeated with the remaining set. This procedure may be stopped prematurely, if the new analog is less active than the other members of the new set. The modification suggested by Gilliom et al. [4] can circumvent such problems.

A method that combines direct optimization, with indications as to which parts of a parameter space have not yet been investigated, was proposed by Boyd [5] under the name "hypersurface iterative projection" (HIP). This method which uses multidimensional scatter plots requires, however, computerization, even though it allows the medicinal chemist to select new substituents from a graphical representation of the data.

3.1.3.2 Methods for the Systematic Investigation of a Parameter Space

The first stage of the *second Topliss method* (see above) already contains elements of a systematic investigation of a parameter space. Thus, the first five compounds are selected with the intention of covering a significant part of a σ/π parameter space. However, the selection is not quite optimal, since the hydrophilic parts of that space are not represented.

A better representation could be achieved by applying one of the various methods that have been developed for series design. With most of these methods one selects a number of structural moieties from a larger predefined set and attaches them to a basic skeleton. In most cases, the structural moieties refer to substituents on an aromatic ring. It is, however, also possible to spefically design structural moieties according to predefined physico-chemical or conformational properties. In the present context, only manual procedure for the design of a series will be described (for computer-aided versions and other computerized methods see Chap. 3.2).

If only one feature (parameter), e.g., lipophilicity (π), is to be investigated, one simply selects a small number of compounds that cover a sufficiently wide range of parameter values, e.g., Ph, Me, H, OH, with the corresponding π-values (taken from Ref. [6]) being 1.96, 0.56, 0.00, -0.67.

A selection with respect to two parameters can be done visually according to Craig [7] by setting up a two-dimensional plot *(Craig plot)* in which the structural moieties (e.g., substituents) appear as points. Moieties are then selected so that the corresponding points are evenly distributed over the plot.

If more than two parameters need to be considered, one can resort to sets of substituents that have been reported in the literature, as being more or less evenly distributed over a larger area of a multi-dimensional parameter space.

Thus, Wootton [8] designed 10 sets, each with 10 members taking into account the lipophilic, resonance, inductive and space filling properties of 35 substituents for aromatic systems. An earlier publication by Franke et al. [9] presented similar sets, based on 90 substituents. Schaper [10] has reported selections that optimized the sets with respect to the extension of the parameter space area being investigated and to the mutual independence of the parameters and aditionally took into account *synthetic accessibility.*

Van de Waterbeemd et al. [11] have analyzed the mutual similarity of 59 substituents and divided them into 5 groups accordingly. Again, selection of one substituent out of each group should result in a representative set to start with. The groups were formed with consideration of the two most important *principal components* which had resulted from a corresponding analysis of 74 descriptors. These comprised various lipophilicity parameters, parameters describing electronic properties, steric properties, connectivity indices, indicators for hydrogen bonding and other indicator variables. Closer examination of the two *principal components* revealed that they largely represented two properties, i.e. that of bulk and polarity. The performance of *principal component analysis* in series design (see Chap. 4) was compared with that of *cluster analysis* (see below). The former method was given preference, because it led to groupings that could be interpreted in terms of the physico-chemical properties.

Cativiela et al. [12] have selected various sets of 10 heterocyclic systems each that represent altogether 18 systems, and in such a way that maximum structure-activity information was obtained. The selection was based on a *Free-Wilson* representation in conjunction with *D-optimal design.*

Hansch et al. [6, 13] have used *cluster analysis* in order group substituents. The members in each group are considered to be similar, so that any can be chosen to represent the whole group. Different groups represent different parts of the corresponding parameter space. A set of test compounds is assembled by selecting one substituent from every group (cluster). Substituents to be introduced into the aromatic system of a basic skeleton are clustered, taking into account lipophilic, resonance, inductive, steric, and H-bonding properties. With substituents for aliphatic components, the same criteria were applied, except for resonance effects.

	A	B	AB	C	AC	BC	ABC	D ...
1	-	-	+	-	+	+	-	-
2	+	-	-	-	-	+	+	-
3	-	+	-	-	+	-	+	-
4	+	+	+	-	-	-	-	-
5	-	-	+	+	-	-	+	-
6	+	-	-	+	+	-	-	-
7	-	+	-	+	-	+	-	-
8	+	+	+	+	+	+	+	-
9	-	-	+	-	+	+	-	+
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
2n+1	0	0	0	0	0	0	0	0 ...

Figure 4. General form of a 2^n-factorial scheme used for series design. The rows refer to individual compounds, whereas the columns represent structural features or physico-chemical properties. Plus signs denote the presence of one of two alternative structural features, or a high level of a particular physico-chemical property. Minus signs refer to the presence of the other feature, or a low level of the physico-chemical property. The last row, containing only zeros, refers to a compound that represents intermediate levels of the physico-chemical properties. The part separated by solid lines (upper left corner) constitutes a 2^2-factorial scheme.

Hansch et al. [6, 13] divided the basic sets of substituents into 5, 10, and 20 clusters, thereby allowing test sets with 5, 10, and 20 compounds to be assembled. The smaller test sets would normally suffice for an initial preliminary investigation for particular types of chemical structures. If this yields promising results, one usually proceeds on to the larger sets in order to derive reliable structure-activity relationships.

The use of predefined test sets, as in the examples given above, is confined to structural moieties (substituents), which are members of the basic set, and to the particular selection criteria employed, irrespective of how relevant they are for the biological property of interest. A manual series design method that is not subject to such limitations is *factorial design.*

In its simplest form, a 2^n-factorial scheme, expresses structural features and physico-chemical properties in binary terms, e.g. as the presence of one of two alternative features vs the presence of the other, or as a high level of a property vs a low level. These classifications are then assigned plus and minus signs in the factorial schemes. A general scheme is shown in Fig. 4.

If the molecular properties are expressed in physico-chemical terms, it is necessary to specify, at least approximately, an upper and a lower limit of that property. The level of a property is considered "high" if it is closer to the upper limit, and "low" otherwide. It is generally advisable, however, to also represent the intermediate levels of physico-chemical properties (denoted by 0), in which case one introduces an additional row into the factorial scheme (the last row in Fig. 4).

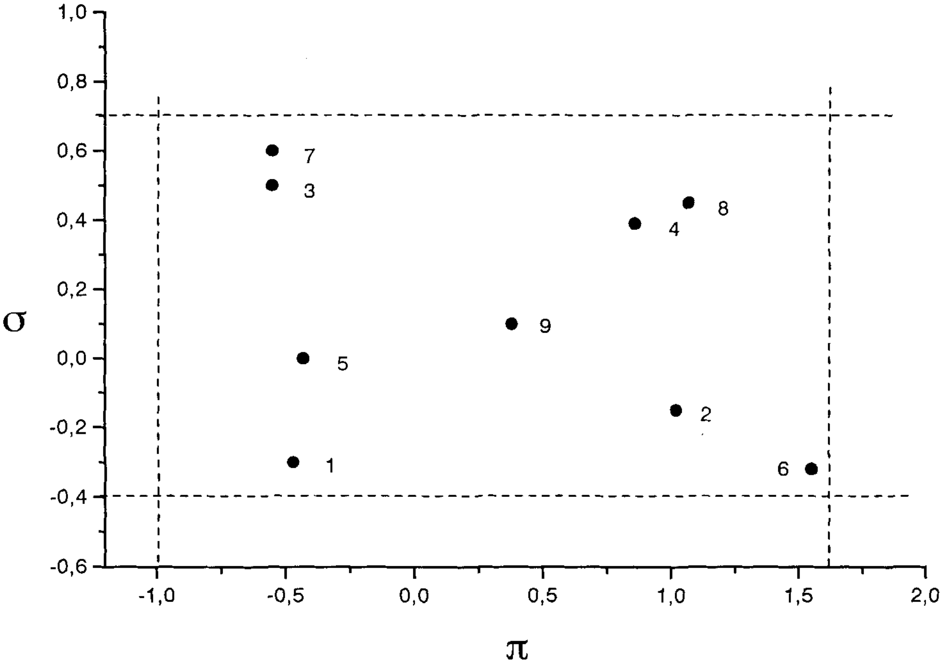


Figure 5. π/σ Craig plot of the set of substituents listed in Table 1. The dashed lines refer to the limits chosen for π and σ, respectively. An additional *MR* axis perpendicular to the plane of the diagram would separate the pairs 3/7, 4/8, 1/5 and 2/6 further (numbering of the substituents as in Table 1).

An illustration of this procedure is given in Table 1 which refers to the selection of 9 substituents for aromatic rings, based on the list of well-characterized substituents for aromatic systems taken from the literature [6], in conjunction with the preset limits to the applied parameters, which are listed in Table 1.

A π/σ Craig diagram (Fig. 5) demonstrates that the 9 compounds are evenly distributed over the chosen area (within the dashed lines). Note that the points lying close to each other (3/7, 4/8, 1/5, 2/6) differ in their *MR* values.

In practice, it is usually more economical to derive approximate structure-activity relationships from small test series, and only then to design new compounds (see the second Topliss procedure). Thus, in the previous example, one may wish to start with less than 9 compounds and factorial schemes allow for such a reduction by confounding one or more of the parameters with cross terms of the scheme *(fractional factorial schemes)*. In the present example, MR could be represented by the AB column in Fig. 4. The scheme now defines 4 compounds (the 4 combinations of levels within the solid lines) and a fifth may be added to include intermediate levels. The resulting set consists of substituents **5**, **2**, **3**, **8**, and **9** of Table 1. For more detailed discussions see Austel [14].

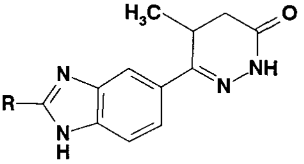
One particularly interesting application of factorial schemes is the design of structural moieties based on a qualitative description of their properties. The

Table 1. Selection of substituents for an aromatic ring. The top part refers to the physico-chemical properties that are considered, and the parameters by which they are represented, including the lower and upper limits of the properties covered. The bottom part lists the selected substituents, together with the combination of levels, which is represented by a particular substituent. The actual values of the parameters are also given (taken from Ref. [6])

property	parameter	low limit	high limit
lipophilicity	π	-1.0	$+1.6$
electronic properties	σ	-0.4	$+0.7$
size	MR	8.0	23.0

No	levels	subst.	π	σ	MR
1	$- - -$	m-$NHCH_3$	-0.47	-0.3	10.33
2	$+ - -$	p-C_2H_5	1.02	-0.15	10.30
3	$- + -$	p-$COCH_3$	-0.55	0.50	11,18
4	$+ + -$	m-Br	0.86	0.39	8.88
5	$- - +$	p-$NHCOC_2H_5$	-0.43	0.00	19.58
6	$+ - +$	p-OC_4H_9	1.55	-0.32	21.66
7	$- + +$	m-$SO_2C_3H_7$	-0.55	0.60	22,79
8	$+ + +$	p-$COOC_3H_7$	1.07	0.45	22,17
9	0 0 0	m-OC_2H_5	0.38	0.10	12.47

following example from the field of cardiotonic pyridazinylbenzimidazoles (**3**) may serve to illustrate the procedure. The question was concerned with which properties of the substituent R were compatible with high potency.



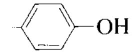
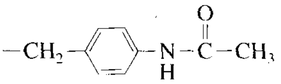
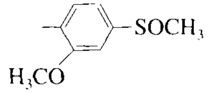
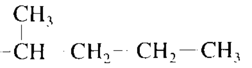
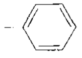
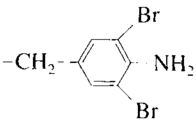
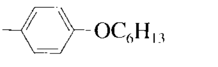
(**3**)

Special features of interest were:

1) nature of the substituent [aliphatic (−) or aromatic (+)] (the term "aromatic" refers to an aromatic ring that is directly joined to the benzimidazole)
2) size [small (−) or large (+)]
3) lipophilicity [hydrophilic (−) or lipophilic (+)].

The latter two features need further specificaiton: the smallest common aromatic substituent is a phenyl group and, therefore, it would be reasonable to consider groups with less than, say 9 non-hydrogen (second row) atoms as "small". The nature of the phenyl ring would also determine the borderline between "hydrophilic" and "lipophilic" groups, in such a way that hydrophilic substituents on the phenyl ring, such as a hydroxyl could render it "hydrophilic". Therefore, groups with an estimated π value of less than about 1.6 may be considered as "hydrophilic".

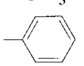
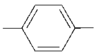
Table 2. Substituents for structure (3) designed directly with a 2^3-factorial scheme, considering the nature [aliphatic (−)/aromatic (+)], the size [<9 non-hydrogen atoms (−)/(+) otherwise], and lipophilicity [π < ca. 1.6 (−)/(+) otherwise]. The left column refers to the combination of levels represented by that particular substituent, the right column gives the increase (% of initial value) in contractility of the heart after i.v. administration of 0.1 mg/kg to anesthetized cats

No	properties			substituent	biological response
	aliph./ arom.	small/ large	hyd./ lip.		
1	−	−	−	$-CH_2-OCH_3$	86
2	+	−	−		75
3	−	+	−		inactive
4	+	+	−		38
5	−	−	+		83
6	+	−	+		ca. 100
7	−	+	+		98
8	+	+	+		inactive

With these specifications it was now possible to design a test series using the first 8 rows and the columns A, B, and C of Fig. 4 (2^3-factorial scheme). According to this scheme, the first substituent (− − −) is a small hydrophilic aliphatic group (e.g., $-CH_2OCH_3$), whereas the sixth substituent is a small lipophilic aromatic group (e.g., $-C_6H_5$). The complete set is listed in Table 2.

Examination of the biological data suggest that a wide range of lipophilicity and size is compatible with a reasonable potency, and that it does not matter whether the substituent has a directly connected aromatic ring or not. However, there appears to be limitations on the length of the substituents. Unfortunately, both inactive compounds contain aromatic rings, albeit in one case not directly adjoining the benzimidazole. In order to clarify whether the presence of the aromatic ring or the length, or both, are responsible for the low potency, one can design a second set of substituents using the following features:

Table 3. Substituents for structure (**3**) designed directly with a 2^2-factorial scheme, considering the features aliphatic (−)/aromatic (+) and short (−)/long (+) (at least 9 bonds from the benzimidazole). Biological data refer to increase (%) in contractility of the heart after i.v. administration of 0.1 mg/kg to anesthetized cats

No	**properties**		**substituent**	**biological response**
	aliph./arom.	**short/long**		
1	−	−	$-CH_3$	85
2	+	−		ca. 100
3	−	+	$-C_{11}H_{23}$	inactive
4	+	+	 OC_6H_{13}	inactive

1) aliphatic (−)/aromatic (+)
2) short (−)/long (+) (long, e.g. with a length of at least 9 bonds beginning at the 2-position of the benzimidazole)

The design is based on 2 features so that four rows (1−4) and two columns (A and B) of Fig. 4 (2^2-factorial scheme) are sufficient. The test set and the corresponding biological data shown in Table 3 support the view that the alkyl chain length is, generally, a limiting factor for the cardiotonic potency of compounds of the general structure (**3**).

All examples of series design which have been discussed so far, do not take into account one very important aspect of biological activity, i.e. the conformation of test compounds. It would be very desirable to plan test series so that this property is represented in its entirety, and yet, without redundance. *Factorial design* can serve this very purpose.

Basically, the conformation is described in terms of the distances between characteristic points within the molecule. These marker points must span the whole structure and may be (but need not be) placed at the center of groups that are relevant for biological activity. Again for every distance, one has to define the range to be covered. Distances that are closer to the upper limit of the range are denoted by (+) and in the converse by (−). These distances may, in the simplest application, be obtained from conventional mechanical molecular models.

In the corresponding factorial schemes, the columns refer to the individual distances, whereas the rows refer to individual conformations. A flexible compound will, therefore, often cover several rows. The test series must be designed so that the following conditions are fulfilled:

1) all the possible conformations are represented by at least two different molecules,
2) for every conformation there are at least two molecules which cannot adopt this conformation,
3) there is no complete correlation between the occurrance of two (or more) conformations.

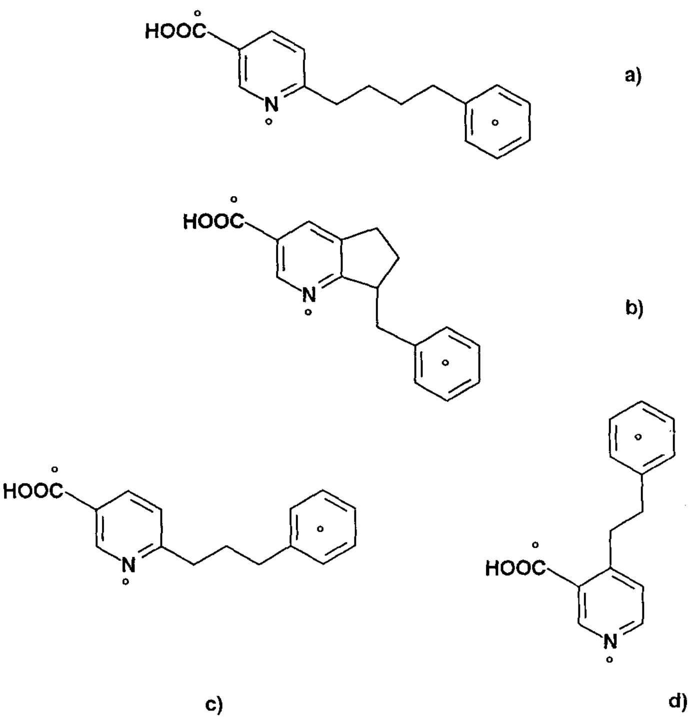


Figure 6. Set of nicotinic acid derivatives, which gives a complete account of the conformational properties of this type of structure within the limits outlined in Table 4. The open circles indicate the position of the marker points.

A simple theoretical example is given in Fig. 6 which shows phenylalkyl nicotinic acids. Three marker points are placed at strategic positions, i.e. on the central carbon atom of the carboxylic group, the ring nitrogen, and the center of the phenyl group. Two of the resulting three distances, i.e. those between the phenyl group and the carboxyl group (ranging from 0.3 to 1.6 nm) and the ring nitrogen (ranging from 0.3 to 1.0 nm) respectively are considered. Therefore, the conformational properties can be treated with a 2^2-factorial scheme which defines four conformations. Table 4 indicates the conformations which can be adopted by the compounds

Table 4. Conformations occupied by the compounds of Fig. 6. Every first column of the signs refers to the distance between the center of the phenyl ring and the carbon atom of the carboxyl group [(−): 0.3 to 0.95 nm, (+): 0.95 to 1.6 nm], the second column represents the distance between the center of the phenyl group and the ring nitrogen [(−): 0.3 to 0.65 nm, (+): 0.65 to 1.0 nm]

compound	a)	b)	c)	d)
conformations	− −	− −		− −
	+ −	+ −	+ −	
	− +			− +
	+ +		+ +	

shown in Fig. 6. All the compounds are more or less flexible and can, therefore, each represent more than one of the conformations defined by the scheme. Altogether the conditions for a complete test series, as outlined above, are fulfilled. A detailed description of a computerized version of the method, which also includes a practical example from the field of negative chronotropic drugs, is currently in press [15]. Finally, 2^n-factorial schemes can also be used for the design of test series to be analyzed, according to the Free-Wilson method [16].

3.1.3.3 Choice of Molecular Descriptors

A fundamental problem in experimental design, apart from selecting a suitable method, is the choice of molecular descriptors. The two factors, which in this context must be taken into consideration are:

1) the type of descriptors to be used, e.g. continuous physico-chemical parameters such as π, σ or quantum mechanical parameters, classifying parameters (e.g. large, medium, small), or indicator variables (e.g. presence or absence of a particular structural moiety)
2) the relevance of the descriptors for the biological process being investigated.

As to which type of descriptor is the most suitable, depends entirely on the problem being addressed. During the search for a lead compound or a preliminary optimization procedure, indicator variables and perhaps classifying parameters may suffice. The latter should be particularly suitable in the more advanced stages of the optimization. Here the aim is to find compounds that surpass a predefined level of potency, which is based on the requirements for therapeutic use. If this level has been reached, factors other than potency, such as selectivity, pharmacokinetics, and metabolism, become the main objective in the optimization. To this end, again classifying descriptors are most suitable. Continuous desriptors are needed if detailed quantitative structure-activity relationships are to be derived, e.g. in order to elucidate binding modes or biomolecular mechanisms. In this context computerized design (see Chapt. 4) may become the method of choice.

In order to obtain well-defined structure-activity relationships, one should ideally use only those parameters that are relevant for the biological activity under consideration. However, such prior knowledge is frequently not available, particularly in most cases of drug design. This problem may be overcome in an iterative manner: one first chooses a set of descriptors, preferably indicator variables or classifying parameters, which one considers potentially relevant. From the resulting test series, preliminary structure-activity relationships are derived, which in turn, will give an indication as to which descriptors contribute most strongly to the variance in biological response. In addition, if inconclusive structure-activity relationships emerge, one might wonder whether important descriptors have been missed. A simple illustrative example is provided by the set in Table 2. The descriptors, on which this series was based, did not lead to conclusive structure-activity relationships. Therefore, these descriptors are, at least within the range covered by this set, not relevant. More significant results could, however, be obtained by introducing a new descriptor, i.e. length.

3.1.4 Summary and Conclusion

Careful design of test series is a prerequisite for an economical use of resources in drug design. A broad spectrum of appropriate methods have been reported in the literature. Each of these methods is applicable to particular types of problems. This issue has been addressed in the present review in a general way. As to which one of these methods is the most suitable in a particular case must, however, be decided upon by the medicinal chemist, based on his own judgement or experience.

References

[1] Topliss, J. G., *J. Med. Chem.* **15**, 1006–1011 (1972)
[2] Topliss, J. G., *J. Med. Chem.* **20**, 463–468 (1977)
[3] Darvas, F., *J. Med. Chem.* **17**, 799–804 (1974)
[4] Gilliom, R. D., Purcell, W. P., and Bosin, T. R., *Eur. J. Med. Chem.* **12**, 187–192 (1977)
[5] Boyd, D. B., *J. Med. Chem.* **36**, 1443–1449 (1993)
[6] Hansch, C., and Leo, A., *Substituent Constants For Correlation Analysis in Chemistry and Biology*, John Wiley & Sons, New York, 1979
[7] Craig, P. N., *J. Med. Chem.* **14**, 680–684 (1971)
[8] Wootton, R., *J. Med. Chem.* **26**, 275–277 (1983)
[9] Streich, W. J., Dove, S., and Franke, R., *J. Med. Chem.* **23**, 1452–1456 (1980)
[10] Schaper, K.-J., *Quant. Struct.-Act. Relat.* **2**, 111–120 (1983)
[11] Van de Waterbeemd, H., El Tayar, N., Carrupt, P. A., and Testa, B., *J. Comput. Aided Mol. Des.* **3**, 111–132 (1989)
[12] Cativiela, C., Garcia, J. I., Elguero, J., Mathieu, D., and Phan Tan Luu, R., *Quant. Struct.-Act. Relat.* **6**, 173–178 (1987)
[13] Hansch C., and Unger, S. H., *J. Med. Chem.* **16**, 1217–1222 (1973)
[14] Austel, V., *Eur. J. Med. Chem.* **17**, 9–16 (1982)
[15] Müller, P., Austel, V., Reiffen, M., Wagner, K., Prox, A., Luger, P., and Witschel, W., *Proc. Natl. Acad. Sci. USA*, in press
[16] Austel, V., Manual Design of Test Series for Free-Wilson Analysis. In: *QSAR and Strategies in the Design of Bioactive Compounds*, Seydel, J. K., ed., Verlag Chemie, Weinheim (1985) p. 247–250

3.2 Applications of Statistical Experimental Design and PLS Modeling in QSAR

Michael Sjöström and Lennart Eriksson

Abbreviations

COST	Change-one-substituent-at-a-time
PCA	Principal component analysis
PLS	Partial least squares in latent variables
PPs	Principal properties
QSAR	Quantitative structure-activity relationships

3.2.1 Introduction

The basis of statistical experimental design in QSAR has been reviewed in the preceding Chapter (3.1). A recent review is also given by Pleiss and Unger [1]. In this chapter we will discuss this topic further and, in particular, outline and exemplify a strategy for QSAR development, in which statistical experimental design plays an important role. An often overlooked problem in QSAR is the selection of the compounds with which to calibrate the model, i.e. how to design the so-called training set. This is unfortunate and may result in an unbalanced test series, which in turn will give rise to QSARs of poor quality. The training set compounds must be representative of the class of compounds, from which they originate, that is, they must be chosen in such a way that they efficiently cover the physico-chemical domain of that class. One approach is to use statistical experimental designs, which are optimal schemes informationally for the selection of efficient training sets [2].

The goal of quantitative structure-activity modeling is to derive a mathematical model, having as good predictive capabilities as possible, of the biological effects of new compounds. However, first the model must be calibrated, using, for example, easily accessible physico-chemical descriptors and measured biological responses for the training set compounds. In order to adequately capture the often complex nature of many biological systems, it is necessary to use a series of several relevant physico-chemical descriptors. This view, the multivariate analogy approach to QSAR modeling, which was introduced by Wold and Dunn [3] and Hellberg [4], assumes that the factors governing the events in a biological system are represented by a multitude of physico-chemical descriptors. In other words, within a series of compounds, it is assumed that a small change in chemical structure will be accompanied by an analogous small change in biological activity, and that the multivariate physico-chemical description will reveal these analogies.

Analogy models can be regarded as linearizations of "real" complicated relationships between chemical properties and biological responses. Wold and Dunn [3] have shown that such analogy models typically have local validity only, that is, they can only encompass compounds having fairly similar structures, and which show commonality in chemical or biological mechanisms. Thus, a QSAR study should be based on a series of chemically and biologically similar compounds. It must be noted, however, that the compounds must be dissimilar enough to cause some systematic change in the biological activity.

Besides being multivariate, QSAR data are often crude, imprecise and strongly collinear. This implies that traditional regression techniques, like multiple linear regression, that assume the physico-chemical descriptors to be exact, 100% relevant, and independent of each other, will not always work well. Thus, in situations where many strongly collinear physico-chemical descriptors and/or biological responses operate together, data analytical methods, other than the classical multiple linear regression technique, must be used. Partial least squares projections to latent structures (PLS) is a projection method, which is particularly well suited for handling these problems. For a more thorough discussion of the data analytical method selection problem, *see Chap 5.2.* PLS is presented in detail in Chap. 4.4 and is well suited for data sets where the number of descriptor variables exceeds the number of compounds. PLS can also tolerate a moderate number of missing observations.

3.2.2 A Strategy for QSAR Development in Drug Design

In the preceding paragraph, some general, but important remarks on modern QSAR analysis were given. These considerations have been incorporated into a strategy for QSAR development, which is described in the next few sections. This strategy consists of six steps, which are closely linked to each other, and are based on the two principal methods of statistical experimental design and multivariate data analysis [5]. Briefly these steps are: (1) Formulation of classes of similar compounds, (2) structural description and definition of design variables, (3) selection of the training set of compounds, (4) biological testing, (5) QSAR development, and (6) validation and predictions for non-tested compounds. In the next paragraph these steps are discussed in more detail. In the examples in paragraph 3.2.3, the emphasis is placed mainly on the statistical experimental design (Step 3), and the QSAR modeling (Step 5).

3.2.2.1 Formulation of Classes of Similar Compounds (Step 1)

Since the mechanism of biological action usually differs between different types of classes of compounds, one can not construct QSARs, which are based on compounds that are too diverse structurally. Thus, the first step of the strategy consists of formulating classes of similar compounds. The ideal situation corresponds to classes, where, within each class, all the compounds are structurally similar and function according to the same mode of action. In reality, this is difficult to achieve and some deviations can be expected.

The formation of classes of similar compounds consists of dividing the series of compounds of interest into categories on the basis of their chemical structure. This may, for instance, be achieved according to their general backbone, their substituents, or perhaps according to crucial properties such as hydrophobicity or chemical reactivity, and knowledge of the biological mechanism. The subsequent multivariate data analysis (Step 2) may give information about deviations from the class similarity, provided that the majority of the compounds are, indeed, chemically and biologically similar. If the data analysis reveals that the investigated compounds do not form a homogeneous class, new classes should be formed. This means that this step is sometimes an iterative procedure.

3.2.2.2 Structural Description and Definition of Design Variables (Step 2)

Once a class of similar compounds has been compiled, the next question is how to appropriately describe the structural variation. Obviously, the demands on the structural description depend not only on the considered compounds, but also on the nature of the biological system under investigation. In general, the more complicated the system under observation is, then the more unlikely it is that a single descriptor variable will contain sufficient information about a given biological phenomenon. Thus, the structural description is multivariate, but to what extent, varies from case to case. The structural and physico-chemical descriptors can be categorized into two groups, viz. (1) global types and (2) substituent types [6]. Global variables, such as log P, are based on the whole molecule, whereas substituent descriptors correspond to a certain part or moiety of a molecule. Depending on the application, the two categories of descriptors can be used independently, or in conjunction with each other. Regardless of which type of variable is chosen, it is usually difficult to predict in advance which descriptor variables will be useful. If no prior knowledge or information exists about the importance of certain factors, it is usually recommended that at least the hydrophobic, steric and electronic properties of the compounds are described.

Prior to the selection of a series of compounds for synthesis according to a statistical experimental design scheme, it is necessary to decide on a set of independent design variables, which might have an influence on the biological effect. In technical optimization applications, variables, such as time, temperature, pressure, pH, etc., usually can be varied independently of each other. In the optimization of molecules, however, where substitution patterns or the whole molecular structure is changed, it is usually not possible to discern design variables that can be changed independently of each other. For example, if size and lipophilicity of the varied substituents are used as design variables, they are rarely independent of each other. If the size of the side-chain is varied, the lipophilicity is also altered. Furthermore, changes in molecular structures are discrete in nature, which means that it is not possible to find combinations of substituents that exactly match a statistical design.

By analyzing the multivariate physico-chemical data table with principal component analysis (PCA), the original number of descriptors are contracted into a few and information-rich principal components. (For a thorough presentation of PCA see Chap. 4.1). The term principal components shall be used interchangeably with design variables in this context, because they can be used as variables in statistical experimental designs, or "principal properties (PPs)". This is simply because they can be assumed to reflect the most important features of the compounds that are hidden in the total variation of all descriptors. The use of PPs in connection with statistical experimental design is discussed further in Chap. 3.4. The concept of using PPs as variables in statistical experimental designs has also been useful, for example, in the selection of solvents, catalysts, etc., in organic synthesis [7].

3.2.2.3 Selection of the Training Set of Compounds (Step 3)

The purpose of the third step is to select a training set of compounds for biological investigations. Unfortunately, this step is often ignored in QSAR research. It is of crucial importance for any QSAR model, irrespective of its origin and future use, that the set of chemicals used to calibrate the model exhibits a well-balanced distribution and contains representative compounds. This can only be attained by a systematic selection of the training set of compounds, where the major structural features are varied systematically and simultaneously. Here, statistical experimental designs are invaluable. This stems from the fact that they generate the training set by introducing systematic variation in all the variables or PPs simultaneously, and not just in one design variable at a time. There are different categories of statistical experimental designs, which are of great practical importance, such as factorial designs (FD), fractional factorial designs (FFD), and D-optimal designs [8 – 10]. The resulting models are easy to interpret, and, with regard to the FDs and FFDs, they are easy to construct and modify.

In an FD or FFD (see Chap. 4.1), each PP (Design variable or factor) is usually given two fixed levels. With more than three design variables (or PPs), FDs usually require too many experiments (in this case compounds). In such situations, FFDs are more attractive, because the number of compounds needed for biological testing is drastically decreased with little loss of information. FDs and FFDs only allow linear and interaction terms to be estimated. However, if FDs and FFDs are complemented with interior centerpoints, they also permit a rough estimation of the quadratic terms, which reflect curvature. These designs may also be complemented to form central composite designs, which allow a more rigorous quantification of curved phenomena. The principles behind the use of PPs for constructing FDs and FFDs are discussed and exemplified in more detail in Chap. 4.1. In the selection of training sets for QSAR applications, so-called D-optimal designs are also of interest. Such designs are particularly attractive in situations where constraints exist in the physico-chemical domain of possible compounds. With a D-optimal design, a subset of a given larger set of compounds which fulfill these restrictions are selected, so that they span the physico-chemical space as well as possible. For a discussion of D-optimal designs and a review of subset selection algorithms, see Carlson [7] and Baroni et al. [8].

3.2.2.4 Biological Testing (Step 4)

One of the core concepts underlying this QSAR strategy, is that the biological testing should be minimized as far as possible. Thus, the basic idea is to merely subject the representative training set compounds to extensive testing, in order to obtain a broad and stable picture of their biological properties. This implies that a large number of biological measurements should be undertaken, so that the response matrix contains biological variables that span as many aspects of the biological profiles of the investigated compounds as possible. The more biological tests that are performed for each compound, the better is the stability of the resulting QSAR model, and this will likely also lead to an improved predictive capability. Besides economic considerations, the testing of a few representative compounds also saves time, and adheres to the principles of animal welfare.

Another general remark about the biological testing, is that such measurements are commonly recorded as dose-response curves, showing the relationships between the administered doses and the responses that they elicit. Typically, the information content of such curves is summarized in a single value, such as LD_{50}, EC_{50}, etc. This need not be a problem if the curves are congruent and exhibit the same general features. Then in such a case, a single value will adequately reflect the existing information. However, if the curves are incongruent, i.e., they are influenced by more than one factor, summarizing a dose-response curve with only a single value may lead to a loss of valuable information. It is, therefore, recommended that the whole (multivariate) dose-response curves are used, whenever possible, in QSAR analyzes.

3.2.2.5 QSAR Development (Step 5)

In the fifth step of the strategy, the main objective is to calculate the best mathematical expression linking together the physico-chemical descriptors and biological responses. During this procedure, information, regarding the essential features of the chemical and biological data structure, is obtained. There may, for instance, be a need to transform some of the descriptor variables, or delete compounds, exhibiting deviating chemical and/or biological properties. The QSAR analysis also provides information on whether a descriptor variable is relevant for a certain application.

In practice, there are two ways, in which the physico-chemical variation of the studied compounds may be represented. One way is to use the PPs for the QSAR development as well. If these are few, and provided that they are information-rich, the calculated QSAR is easy to interpret. The problem might arise, however, that the PPs are not sufficiently adequate for QSAR development. Although the PPs are derived by a maximum variance projection in PCA, some of the residual variance might be essential to QSAR development. In the situation, where the PPs are found insufficient for QSAR analysis, it is recommended that one returns to the use of the original physico-chemical descriptors. These might lead to an improvement of the model.

In QSAR development, we recommend the use of multivariate partial least squares projections to latent structures (PLS). This is because PLS is a projection-based method and estimates the correlation structure among the collinear descriptors in terms of a limited number of latent variables. This means that PLS can analyze any number of variables, regardless of the number of compounds in the training set, which is beneficial, since the projections to latent structures become more stable as more informative variables are included. The statistical significance of the QSAR is then assessed by means of cross-validation [11]. A measure of the predictive capability of a model, based on cross-validation, is Q^2 (sometimes denoted as R^2_{CV}, describing the amount of variance in y that can be predicted). This can be compared with the R^2 value, expressing the variance modeled in y. Both these statistical measures vary between 0 and 1, where 1 signifies a perfect model, and 0 a model, which has no relevance. The percent variance predicted or explained is expressed as $100{*}Q^2$ or $100{*}R^2$. A large discrepancy between R^2 and Q^2 might indicate an overfitting of the QSAR model. Cross-validation is sometimes referred to as an "internal" procedure to ascertain the predictive capability of a QSAR. Cross-validation is discussed in more detail in Chaps. 5.1 and 5.2 and briefly in the next paragraph (3.2.2.6).

3.2.2.6 Validation and Predictions for Non-Tested Compounds (Step 6)

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

It is desirable that the compounds in the validation set adequately span the physico-chemical domain and the biological activity range of interest. Conveniently, the validation set may be selected according to a statistical experimental design in order to result in a series of representative compounds. In fact, the validation set can be selected already at the third stage of the strategy, simultaneously with selection of the training set. It also seems relevant to stress that the cross-validation procedure (internal validation) and the verification of the validation set (external validation), are not mutually exclusive. On the contrary, these methods should be regarded as being complementary and, which can be used to obtain an estimate of the precision, with which the biological activity can be predicted with the QSAR model.

3.2.3 Examples of Design and PLS Modeling

In this section eight examples are given with the aim of illustrating the concepts of experimental design and PLS analysis in the development of QSARs. The examples concern a diverse series of peptides or peptoids (Secs. 3.2.3.1 to 3.2.3.4), halogenated alkanes (Sec. 3.2.3.5), dibenzofuranes (Sec. 3.2.3.6), aromatics (Sec. 3.2.3.7) and corrosive carboxylic acids (Sec. 3.2.3.8). In the examples given in Secs. 3.2.3.1 to 3.2.3.4, the statistically designed training sets were constructed in retrospect from earlier and more or less systematically varied series of compounds, but for which QSARs were developed. In the examples given in Secs. 3.2.3.5 to 3.2.3.8, on the other hand, a class of interesting compounds was defined prior to the biological testing. Then, a limited set of compounds was selected by statistical experimental design. Models were then developed, and the validity of the models was tested by separate validation sets.

3.2.3.1 Bradykinin Potentiating Pentapeptides

A series of 30 pentapeptides with bradykinin potentiating activity, which had variable amino acid sequences, was reported by Ufkes et al. [12, 13]. The biological activity was expressed as a relative activity index, *RAI*, relative to one of the peptides. A QSAR was developed using PLS based on a numerical description of each of the varied positions on the peptides in terms of three PPs, denoted as z_1, z_2 and z_3 (see below), for the 20 coded amino acids. Thus, each pentapeptide was described by 15 variables.

The three z scales for the amino acids were calculated by PCA from an autoscaled and mean centered multiproperty matrix of 29 physico-chemical variables [14]. The scales or PPs of the amino acids are listed in Table 1, and plotted against each other in Fig. 1. These PPs have been shown to be relevant in the development of numerous peptide QSARs [6, 15 – 17]. Moreover, they have been extended also to comprize a large number of non-coded amino acids [18].

The PLS analysis resulted in a model with two significant components, where $R^2 = 0.82$, i.e. the model described 82% of the variance in the data, and $Q^2 = 0.70$, i.e. 70% of the variance in the biological activity was predicted by cross-validation. A plot of the observed values against the calculated values is shown in Fig. 2a. In

Table 1. PPs or descriptor scales, z_1, z_2 and z_3, for the coded amino acids (AA)

AA	z_1	z_2	z_3	AA	z_1	z_2	z_3
Ala(A)	0.07	−1.73	0.09	His(H)	2.41	1.74	1.11
Val(V)	−2.69	−2.53	−1.29	Gly(G)	2.23	−5.36	0.30
Leu(L)	−4.19	−1.03	−0.98	Ser(S)	1.96	−1.63	0.57
Ile(I)	−4.44	−1.68	−1.03	Thr(T)	0.92	−2.09	−1.40
Pro(P)	−1.22	0.88	2.23	Cys(C)	0.71	−0.97	4.13
Phe(F)	−4.92	1.30	0.45	Tyr(Y)	−1.39	2.32	0.01
Trp(W)	−4.75	3.65	0.85	Asn(N)	3.22	1.45	0.84
Met(M)	−2.49	−0.27	−0.41	Gln(Q)	2.18	0.53	−1.14
Lys(K)	2.84	1.41	−3.14	Asp(D)	3.64	1.13	2.36
Arg(R)	2.88	2.52	−3.44	Glu(E)	3.08	0.39	−0.07

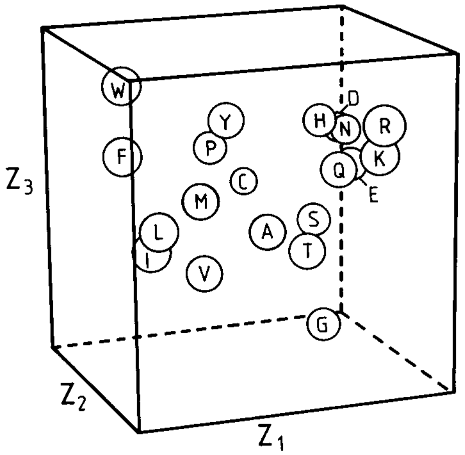


Figure 1. Scatter plot of the three PPs (z_1, z_2 and z_3) for the 20 coded amino acids.

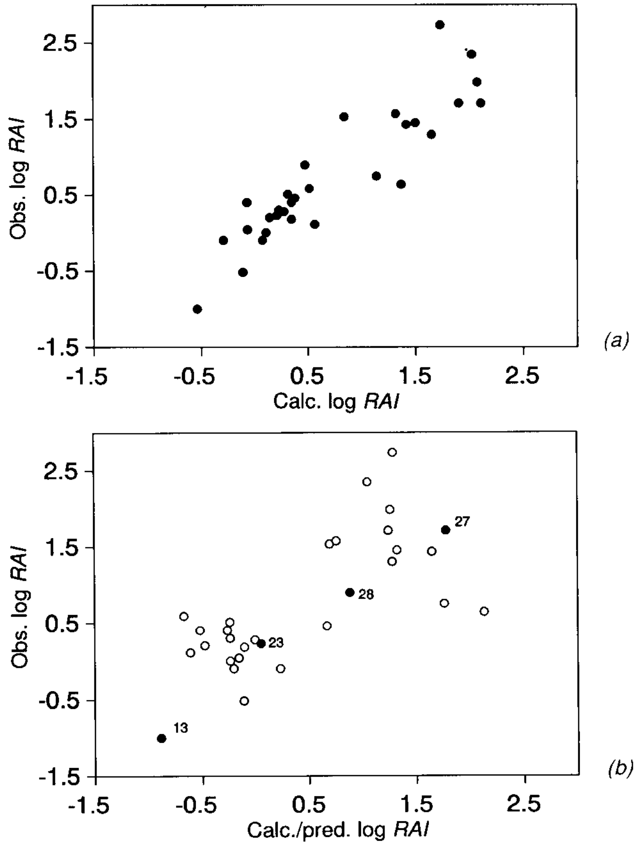


Figure 2. a) A plot of the observed activities against the calculated activities for a PLS model, based on 30 pentapeptides. b) The observed activities plotted against the predicted activities (open circles) for a PLS model, based on a designed training set (filled circles) for z_1 at position 3 and 4 (see Table 1).

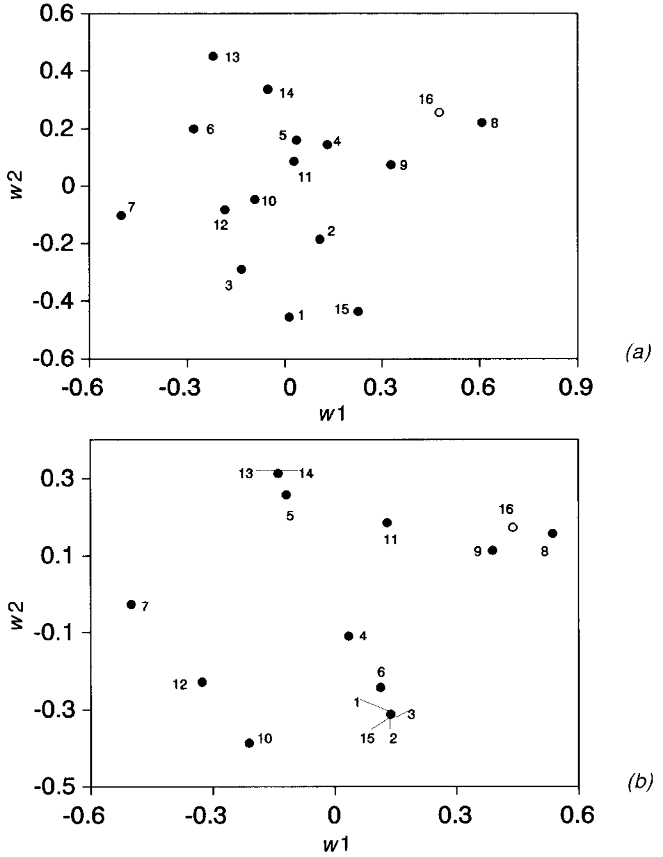


Figure 3. a) PLS weights for the model based on all 30 pentapeptides, and b), based on the designed training set. The numbering refers to the amino acid position and the scale used, i.e. 1–3 (z_1-z_3) at position 1, 4–6 (z_1-z_3) at position 2, 7–9 (z_1-z_3) at position 3, etc. Variable 16 is the weight of the biological activity.

Fig. 3a, the PLS weights are plotted against each other. The plot in Fig. 3a shows that the weights for the variables 7, 8 and 9, corresponding to position 3, have the largest absolute values for the first model dimension. Thus, position 3 is the most influential for regulating the *RAI*.

The present 30 peptides were not synthesized according to a statistical design. In retrospect, however, we have investigated to determine whether the apparent success of the QSAR is due to an intrinsic design among the 30 pentapeptides. Indeed, an approximate full factorial design in just z_1 in the most varied positions 3 and 4, was present among the 30 pentapeptides (see Table 2). A PLS model based on this set of four peptides, as a training set, results in a model yielding satisfactory predictions, as shown in Fig. 2b. This shows that a design with few compounds can be valuable for screening purposes even if not all of the design variables are informative. We also noted that the PLS weights are quite similar in the model, based on the

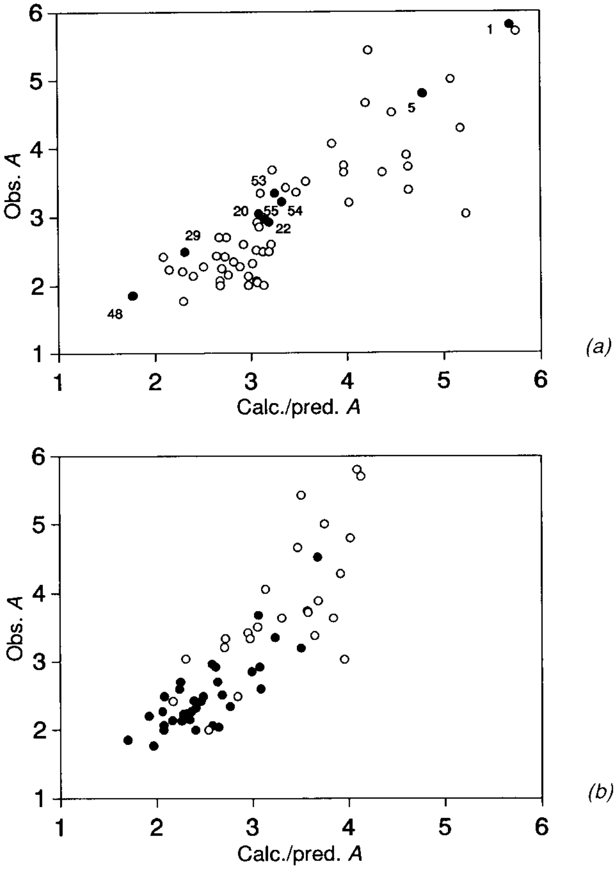


Figure 4. a) The observed activities plotted against the calculated activities (open circles) for the QSAR model, based on a designed training set with nine dipeptides (filled circles) in z_1 and z_2 at positions 1 and 2 (see Table 3). b) The observed activities plotted against the predicted activities (open circles) for a COST designed training set with 34 dipeptides (filled circles) and a PLS model, based on linear terms only. c) The plot shown is similar to the one in b), except that it is a PLS model, based on linear, quadratic and cross terms. d) The observed activities plotted against the calculated activities for a PLS model, based on all 58 dipeptides.

design, as well as for the model, based on all 30 peptides (compare Figs. 3a and 3b). For example, the variables, 7, 8 and 9, have high absolute values of their weights for the first dimension in Fig. 3a and exhibit a similar size and sign as the corresponding weights plotted in Fig. 3b.

Table 2. The 2^2 FD for the selected training set of the pentapeptides. *PP*1, i.e. z_1, is used as design variable for both positions 3 and 4 (amino acids in bold letters)

FD		no.[a]	Pentapeptide	Setting in *PP*1	
$z_1(3)$	$z_1(4)$			$z_1(3)$	$z_1(4)$
−	−	**27**	VEWVK	−4.75	−2.69
+	−	**23**	VAAWK	0.07	−4.75
−	+	**28**	PGFSP	−4.92	1.96
+	+	**13**	VGGGK	2.23	2.23

[a] The numbers correspond to those plotted in Fig. 2b and to the data given by Hellberg et al. [15].

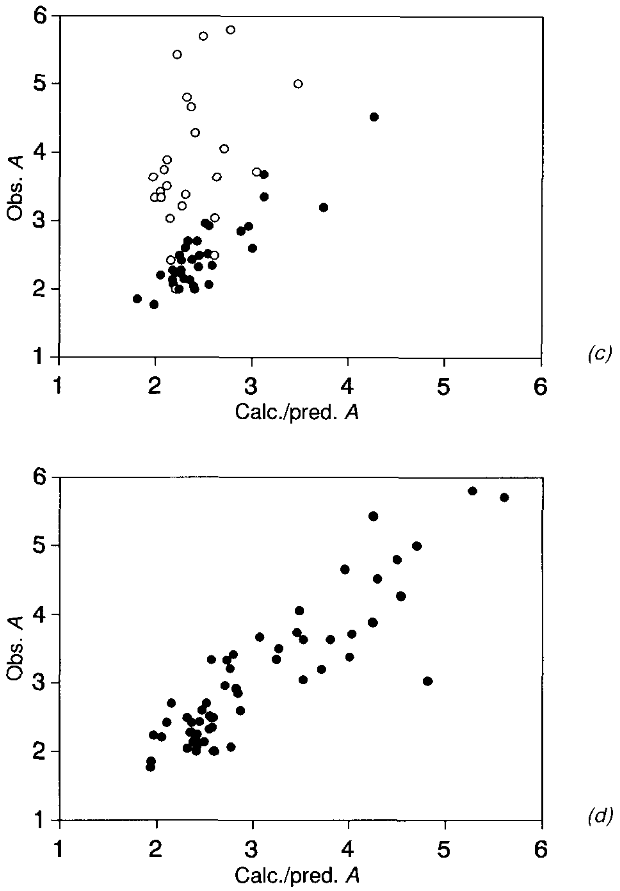


Figure 4. Continued.

3.2.3.2 Dipeptides (Inhibiting the Angiotensin Converting Enzyme)

In this section, we will show that the predictive capability of a QSAR model is strongly dependent on the strategy used for selecting the compounds in the training set. Thus, we have, in retrospect, compiled two training sets, one based on a statistical design, and one on a change-one-separate-feature-at-a-time (COST) design from a series of 58 dipeptides which inhibit the angiotensin converting enzyme. The activity, A, is expressed as $6 + \log(1/I_{50})$, where I_{50} is the concentration (in μM), which inhibits 50% of the angiotensin converting enzyme. The biological data was compiled by Cheung et al. [19]. The results from the statistical design and the COST approaches are compared to the result of a reference QSAR model with all 58 dipeptides included (see page 14).

Table 3. The 2^{4-1} FFD for z_1 and z_2 for a peptide varied at two positions (1 and 2). The design is complemented with a center point. Dipeptides (DP) corresponding approximately to the settings of the angiotensin data are given

FFD				no.[a]	DP	Settings in PPs			
$z_1(1)$	$z_2(1)$	$z_1(2)$	$z_2(2)$			$z_1(1)$	$z_2(1)$	$z_1(2)$	$z_2(2)$
−	−	−	+	**1**	VW	−2.7	−2.5	−4.8	3.7
+	−	−	−	**22**	GI	2.2	−5.3	−4.4	−1.7
−	+	−	−	**55**	YA	−1.4	2.3	0.0	−1.7
+	+	−	+	**5**	RW	2.9	2.5	−4.8	3.7
−	−	+	−	**20**	VG	−2.7	−2.5	2.2	−5.4
+	−	+	+	**29**	GR	2.2	−5.3	2.9	2.5
−	+	+	+	**55**	FR	−4.9	1.3	2.9	2.5
+	+	+	−	**48**	DG	3.6	1.1	2.2	−5.4
0	0	0	0	**54**	AA	0.0	−1.7	0.0	−1.7

[a] The numbers correspond to those plotted in Fig. 4a and to the data given by Hellberg et al. [17].

The Statistical Design Approach

Prior to the construction of a statistical design, it was necessary to decide which descriptor variables might be of importance for biological activity. Here again, we have used the *z* scales for the amino acids (Sec. 3.2.3.1) as design variables, as they are independent of each other and summarize the information content of many different types of physico-chemical variables. We constructed a 2^{4-1} fractional factorial design for each of the two dipeptide positions, using only z_1 and z_2 as design variables (see Table 3). It was not possible to find dipeptides corresponding to a design with all three *z* variables. The design was also complemented with a center point. The nine dipeptides, which best corresponded to the settings in the FFD, were then selected. The training set of the nine peptides was modeled using PLS, including all cross and quadratic terms in the six variables, i.e. 27 descriptor variables (6 linear +6 quadratic +15 cross terms). The QSAR model ($R^2 = 0.97$ and $Q^2 = 0.53$) was then used to predict the biological activity of the remaining 49 dipeptides. In Fig. 4a, the observed activities are plotted against the activities predicted. The predictions can be compared with the results, when all dipeptides are included in the model.

The COST Approach

Among the 58 dipeptides, 34 contained glycine at either the first or the second position. Thus, these 34 compounds represent a training set, compiled according to the strategy to "systematically" change one separate feature at a time, i.e. the COST approach.

A QSAR for this training set was calculated ($R^2 = 0.64$ and $Q^2 = 0.52$), which was based on the three *z* scales for each one of the dipeptide positions. The cross and quadratic terms were not included. This QSAR was then used to predict the

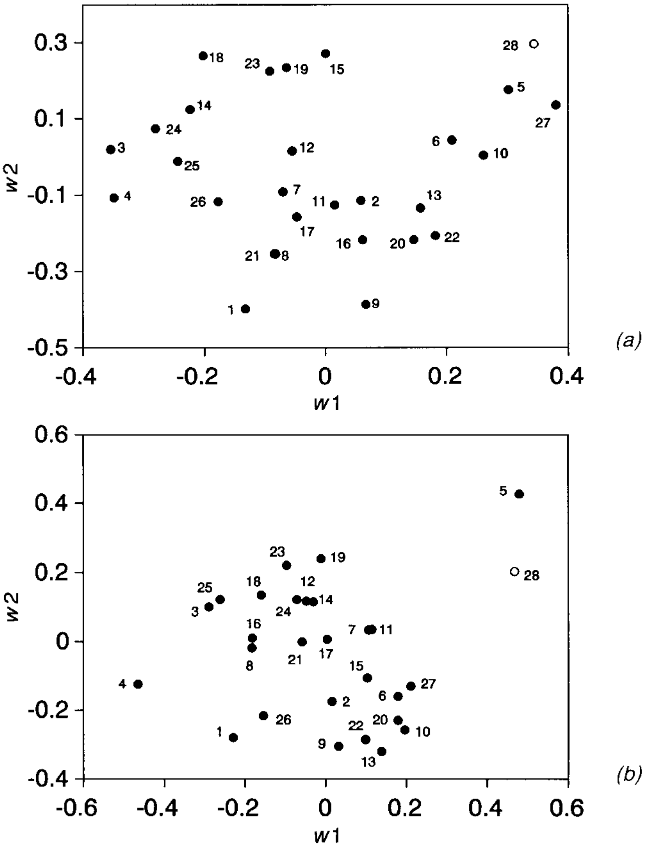


Fig. 5. a) PLS weights for the model, based on all 58 dipeptides, and in b), based on the designed training set. The numbering refers to the amino acid position and the scale used, i.e. 1 – 3 ($z_1 - z_3$) at position 1, 4 – 6 ($z_1 - z_3$) at position 2, 7 – 12 square terms and 13 – 27 cross terms. Variable 28 is the weight for the biological activity.

biological activity for the remaining 24 compounds. The predicted activites are plotted against the observed activities in Fig. 4b. A comparison of Fig. 4b with Fig. 4a, shows that the predictions from the COST design are much worse compared to the predictions from the FFD. The predictions from the COST design are even worse, if cross terms and quadratic terms are included (Fig. 4c).

Model Based on all 58 Compounds

In order to obtain a reference model, a QSAR was calculated based on the complete set of 58 dipeptides. As before each of the dipeptides was described by the three descriptors z_1, z_2 and z_3 at each amino acid position. In addition, to account for a weak non-linear behavior between the biological data and the physico-chemical characterization, quadratic and cross terms were added. PLS analysis resulted in a model with two significant latent variables (R^2 = 0.78 and Q^2 = 0.68). In this case,

Table 4. The 2^4 FD for z_1 and z_2 at position 1 and 2. Peptide analogs, approximately corresponding to the design matrix, were selected from the set of 48 bitter dipeptides

FD[b]				no.[a]	DP	Settings in PPs			
$z_1(1)$	$z_2(1)$	$z_1(2)$	$z_2(2)$			$z_1(1)$	$z_2(1)$	$z_1(2)$	$z_2(2)$
+	+	+	+		[c]				
+	+	+	−		[c]				
+	+	−	+		[c]				
+	+	−	−		[c]				
+	−	+	+		[c]				
+	−	+	−		[c]				
+	−	−	+	**18**	GW	2.23	−5.36	−4.75	3.65
+	−	−	−	**62**	SL	1.96	−1.63	−4.19	−1.03
−	+	+	+	**59**	WE	−4.75	3.65	3.08	0.39
−	+	+	−	**54**	FG	−4.92	1.30	2.23	−5.36
−	+	−	+	**60**	WW	−4.75	3.65	−4.75	3.65
−	+	−	−	**55**	FL	−4.92	1.30	−4.19	−1.03
−	−	+	+	**45**	IE	−4.44	−1.68	3.08	0.39
−	−	+	−	**47**	IS	−4.44	−1.68	1.96	−1.63
−	−	−	+	**41**	IW	−4.44	−1.68	−4.75	3.65
−	−	−	−	**39**	II	−4.44	−1.68	−4.44	−1.68

[a] The FD settings are not in the standard order.
[b] The numbers correspond to those plotted in Fig. 6a and to the data given by Asao et al. [20].
[c] Combinations of amino acid properties, which were not found in the set of 48 dipeptides.

all, except one of the dipeptides, were well described by the model (see Fig. 4d). The PLS weights for the two model dimensions are plotted against each other in Fig. 5a. A comparison with the corresponding plot for the FFD (Fig. 5b) reveals a general similarity of the weights for the two models. This strongly underlines the stability of the model obtained from the designed training set.

To conclude, we have, in practice, demonstrated the superior predictive capabilities of a QSAR model, which is based on an approximate statistical design, compared to a QSAR model, which is based on a poorly balanced design – the COST design. This is despite the fact that the COST design is based on 34 combinations of 19 different amino acids and the FFD design only consists of nine combinations of 9 amino acids. We have also noted that the model composed of all 58 dipeptides, is similar to the one based on a designed training set with only 9 dipeptides. Thus, we propose that the training set analogs for QSAR studies should *always* be selected according to an experimental plan. This should increase the information content in the training sets in comparison to arbitrary or COST designs. This example also shows that PPs are well suited as design variables in FD or FFD.

3.2.3.3 Dipeptides (Bitter Tasting)

Similarly to the example discussed in Sec. 3.2.3.2, we have investigated 48 bitter tasting dipeptides, which were compiled by Asao et al. [20]. A 2^4 factorial design, with z_1 and z_2 as design variables, was constructed, resulting in 16 different

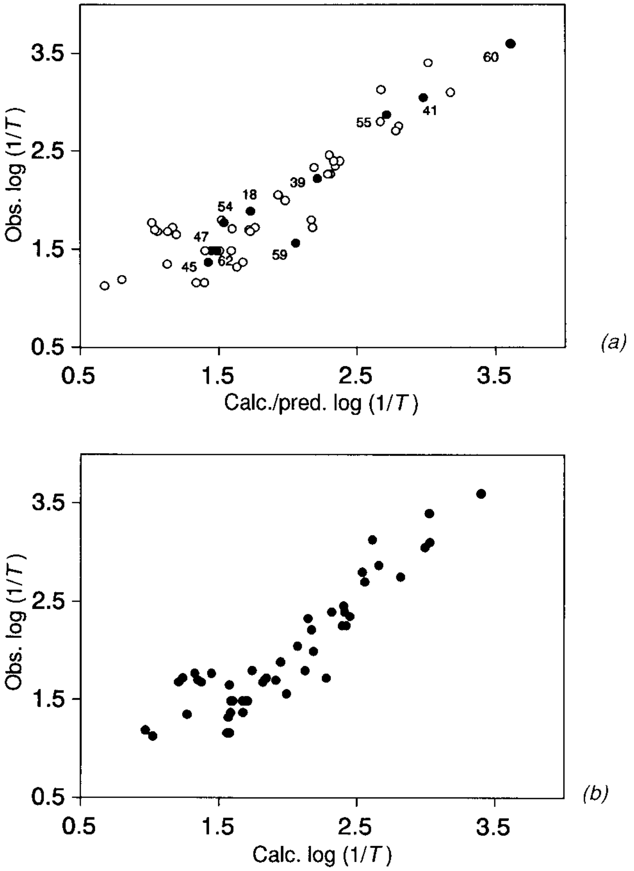


Figure 6. a) The observed bitterness threshhold activities plotted against the predicted bitterness threshold activities (open circles) for the QSAR model, based on a designed training set with ten dipeptides (filled circles) in z_1 and z_2 at position 1 and 2. b) The observed bitterness activities plotted against the calculated bitterness activities for a PLS model, based on all 48 dipeptides (see Table 4).

combinations for the two amino acid positions. In ten of these combinations, dipeptides were found (see Table 4). The missing dipeptides mainly corresponded to those with polar amino acids such as aspartate and arginine in the first N-terminal position. A QSAR based on all three z scales, with two PLS components, was able to predict the bitterness threshold (log $1/T$) for the remaining 38 dipeptides (Fig. 6a) with good accuracy (R^2 = 0.82 and Q^2 = 0.54). This plot can be compared with the corresponding relationship, (Fig. 6b) based on a two-component PLS model, with all 48 dipeptides included (R^2 = 0.82 and Q^2 = 0.76). Indeed, the training set, which was selected according to an approximate experimental design, was informative, and the loss of information was marginal with the designed set compared to the model, based on all of the dipeptides.

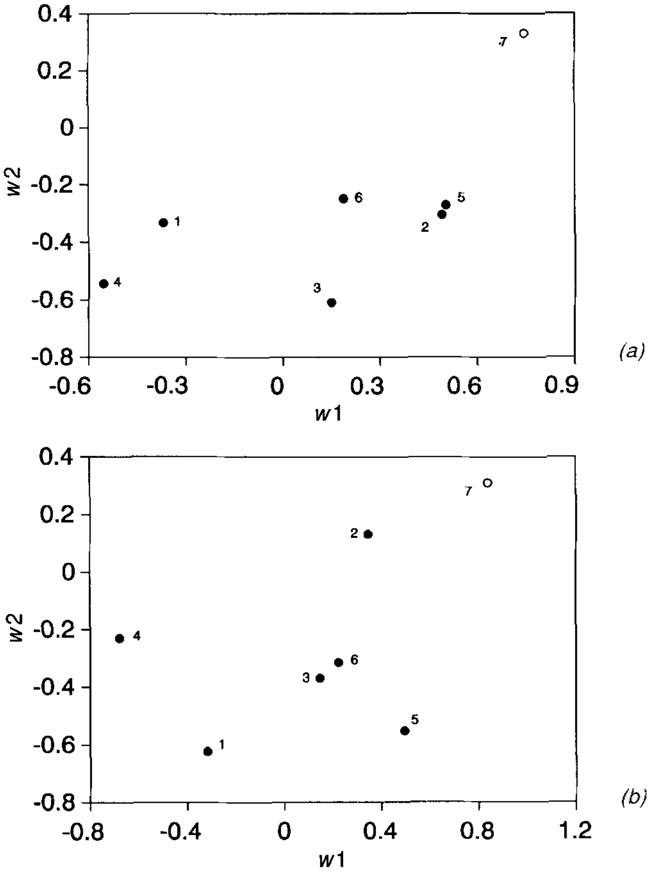


Figure 7. a) PLS weights for the model, based on all 48 dipeptides, and b) based on the designed training set. The numbering refers to the amino acid position and the scale used, i.e. 1−3 (z_1-z_3) for position 1, 4−6 (z_1-z_3) for position 2. Variable 7 is the weight for the biological activity.

A comparison of the PLS weights for the model with all the dipeptides included with the PLS weights for the designed set, also revealed similarities (Fig. 7a and 7b). Thus, the designed set summarizes information about the most important variables, and this information is scarcely affected by increasing the number of compounds in the training set.

3.2.3.4 Mimetics

The complex process of forming a non-peptide molecule from a peptide is also of considerable interest in drug design. One possibility would be to substitute one or more of the peptide bonds with so called isosteres (e.g. $-CH_2CH_2-$), to give a peptidomimetic, which would be more resistant to hydrolysis. There are few systematic QSAR studies involving the physico-chemical characterization of mime-

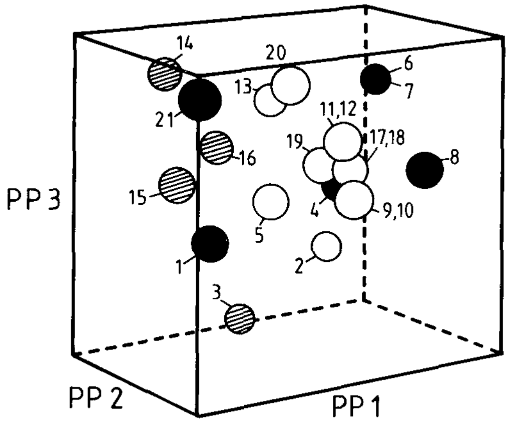


Figure 8. Scatter plot of the PPs for the peptidomimetics. Filled circles correspond to the training set and shaded circles correspond to the test set for the cholecystokinin antagonists (CCK-A and CCK-B). The mimetics are: **1**, $-COCH-$; **2**, $-CH_2O-$; **3**, $-COO-$; **4**, $-CH_2NH-$; **5**, $-COCH_2-$; **6**, $-CH=CH-$(*cis*); **7**, $-CH=CH-$(*trans*); **8**, $-CH_2CH_2-$; **9**, $-CHOHCH_2-$(*d*, *S*); **10**, $-CHOHCH_2$-(*l*, *R*); **11**, $-CHNH_2CH_2-$(*d*, *S*); **12**, $-CHNH_2CH_2-$(*l*, *R*); **13**, $-CH_2S-$; **14**, $-CSNH-$; **15**, $-CONCH_3-$; **16**, $-COS-$; **17**, $-CH(O)CH-$(*trans*); **18**, $-CH(O)CH-$(*cis*); **19**, $-CH_2NOH-$; **20**, $-CH_2C(NOH)-$; **21**, $-CHN(COCH_3)-$.

tics. One exception is the work of Fincham et al., [21] where a physico-chemical characterization of mimetics was used in structure-activity studies for a series of dipeptoids. Here, the amide bond was replaced with a series mimics. The biological effects studied for the dipeptoids were the IC_{50} values for cholecystokinin (CCK-A and CCK-B) antagonism. Recently, we extended their study, see Berglund et al., [22], by describing 21 mimetics with 26 physico-chemical variables. The physico-chemical characterization was then used to calculate three PPs for the mimetics. The PPs, plotted in Fig. 8, were used to construct a 2^{3-1} FFD (see Table 5), and dipeptoids approximately matching these specifications were found among the reported CCK antagonists. The design was complemented with an approximate center point. The series of five compounds was used to construct two QSAR models

Table 5. The 2^{3-1} FFD for the mimetic example. The PPs for the mimetics shown in Fig. 8 were used as design variables

FFD[a]			no.[b]	Mimetic	Settings in PPs		
*PP*1	*PP*2	*PP*3			*PP*1	*PP*2	*PP*3
−	−	+	**1**	$-COCH-$	−3.27	−2.94	1.08
−	+	−	**21**	$-CHN(COCH_3)-$	−5.35	3.21	−1.45
+	−	−	**8**	$-CH_2CH_2-$	4.29	−0.51	−1.76
+	+	+	**7**	$-CH=CH-$(*cis*)	4.96	2.19	2.18
0	0	0	**4**	$-CH_2NH-$	2.81	−0.85	0.57

[a] The FFD settings are not in the standard order.
[b] The numbers correspond to those plotted in Fig. 8.

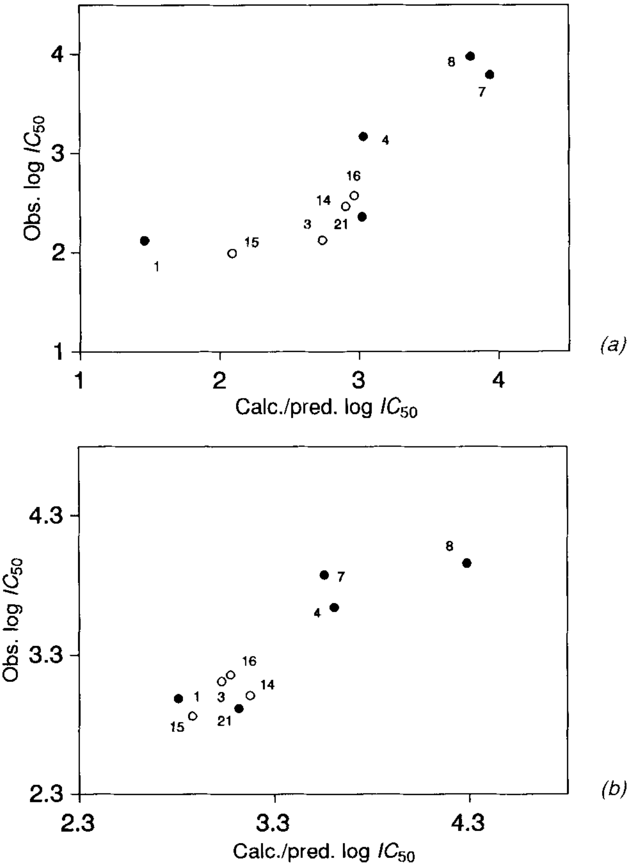


Figure 9. The observed activities plotted against the predicted activities (open circles) for the two mimetic test sets. The QSARs are based on the designed training set (filled circles) for the dipeptoids a) cholecystokinin (CCK-A) and b) (CCK-B).

which were used to predict the remaining reported CCK antagonists. Plots of the observed versus the predicted biological effects for the designed training set, indeed, showed good predictions for the remaining dipeptoids with known biological activites (see Figs. 9a and 9b).

3.2.3.5 Haloalkanes

In contrast to our first four examples, where the training sets were generated in retrospect, based on the existing literature data, the training sets in the following examples were generated before the biological testing. In the first example, the application of the QSAR strategy to a class of halogenated aliphatic hydrocarbons is discussed. This group of chemicals is of relevance for QSAR investigations from an environmental point of view. The class under consideration comprised

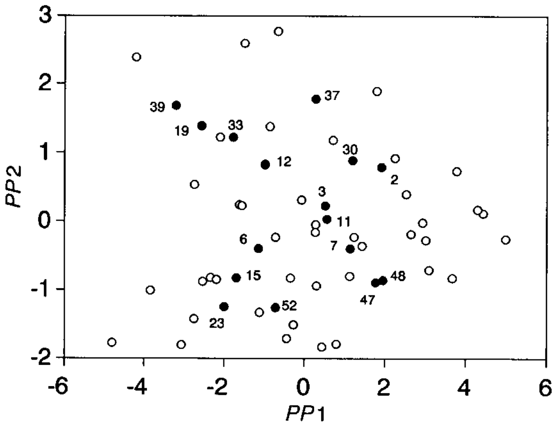


Figure 10. Scatter plot of *PP*2 vs *PP*1 for the 58 haloalkanes. The compounds submitted to biological testing (training and validation sets) are marked with filled circles. The compounds are numbered as in Table 6.

58 compounds with up to four carbon atoms and a varying degree of halogenation [5]. Interestingly, this class contained many chlorinated chemicals that are used on a large scale in industrial applications (as solvents, etc.), and also several chlorinated-fluorocarbons (CFCs).

In order to accomplish a multivariate characterization of the structural and physico-chemical properties of the 58 haloalkanes, a series of 13 descriptor variables was compiled [5]. These variables were subjected to PCA, and four significant principal components described 87% of the total variance. The scores of the first two PPs are plotted against each other in Fig. 10. These scores represented the PPs of the halogenated aliphatics and were used to construct a 2^{4-1} FFD, which is

Table 6. The 2^{4-1} FFD for the training set (top) and the 2^{3-1} FFD for the validation set (bottom). The haloalkane example

FFD				no.	Compound	Settings in PPs			
*PP*1	*PP*2	*PP*3	*PP*4			*PP*1	*PP*2	*PP*3	*PP*4
−	−	−	−	**52**	$CH_3(CH_2)_2Br$	−0.72	−1.26	−1.29	−0.51
+	−	−	+	**48**	$CH_3CHClCH_3$	1.96	−0.86	−0.81	0.15
−	+	−	+	**33**	CH_3CHBr_2	−1.77	1.22	−0.14	−0.08
+	+	−	−	**30**	CH_3CH_2Br	1.20	0.89	−0.90	−0.12
−	−	+	+	**15**	$CHCl_2CHCl_2$	−1.69	−0.83	0.92	0.70
+	−	+	−	**7**	CCl_3F	1.14	−0.40	0.95	−1.10
−	+	+	−	**39**	CBr_3F	−3.20	1.68	1.07	−1.90
+	+	+	+	**2**	CH_2Cl_2	1.92	0.79	0.13	0.70
0	0	0	0	**3**	$CHCl_3$	0.52	0.22	0.70	0.48
0	0	0	0	**11**	CH_2ClCH_2Cl	0.56	0.03	0.28	1.54
−	−	+		**23**	$CH_2ClCHClCH_2Cl$	−2.00	−1.25	0.28	
+	−	−		**47**	$CH_3CH_2CH_2Cl$	1.77	−0.89	−0.90	
−	+	−		**19**	CH_2BrCH_2Br	−2.56	1.39	0.67	
+	+	+		**37**	CH_2BrCl	0.28	1.78	−0.14	
0	0	0		**12**	CH_2BrCH_2Cl	−0.98	0.83	0.32	
0	0	0		**6**	CCl_4	−1.14	−0.40	1.58	

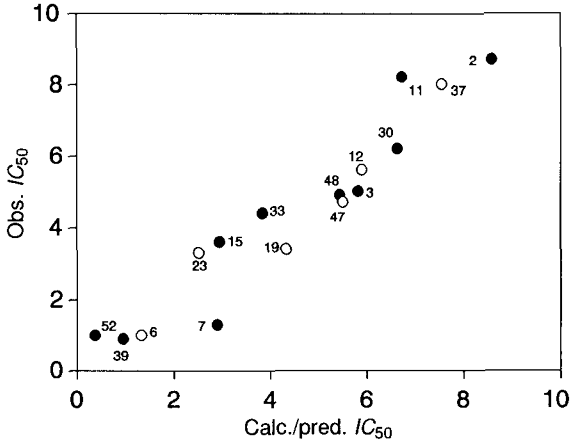


Figure 11. Correlation plot showing observed cytotoxicity (IC_{50}) plotted against the corresponding calculated (training set) and predicted (validation set) values. Training set compounds are marked with filled circles and validation set compounds with open circles. Notation as in Table 6.

summarized in Table 6. This design encodes eight compounds, which was supplemented with two compounds located in the interior part of the design. Thus, a training set, consisting of ten chemicals, was selected. In an analogous manner, a validation set of six chemicals was chosen (Table 6). This set was generated by a 2^{3-1} FFD, augmented with two center points.

The 16 compounds (training + validation sets), which were preferred as representatives for the whole class of halogenated aliphatics, were subjected to a broad range of biological tests, e.g. for acute and subacute toxicity, mutagenicity and cytotoxicity [5]. Here, we discuss the recent results that were obtained from cytotoxicity tests with human HeLa cells [23]. The cytotoxicities of the 16 compounds were expressed as the concentration, which inhibits cell growth by 50%, and is termed the IC_{50}. In order to account for the variation in cytotoxicity among the tested compounds, we found a subset of five predictor variables, which could sufficiently describe the biological endpoint. These five highly correlated descriptor variables were the molecular weight, the van der Waals volume, the octanol/water partition coefficient, and the log retention times from two HPLC systems. The PLS analysis, based on these five predictors, gave a one-dimensional model with $R^2 = 0.89$ and $Q^2 = 0.88$. As seen in Fig. 11, the QSAR accurately predicts the cytotoxicity for the compounds in the validation set. Thus, this QSAR may also be useful for predicting the cytotoxicities of the 42 non-tested halogenated aliphatic hydrocarbons.

3.2.3.6 Dibenzofurans

This example refers to a series of 87 polychlorinated dibenzofurans (PCDFs), for which a biological response concerning a rat enzyme induction potency was determined. Tysklind [24] first compiled a multivariate characterization of these compounds, consisting of 18 chemical descriptor variables. These descriptors were summarized by PCA (see score plot in Fig. 12). The resulting four PCs were used

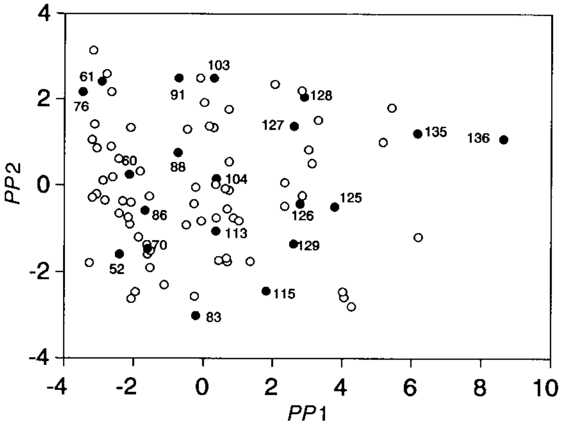


Figure 12. Scatter plot of *PP*2 vs *PP*1 for the 87 PCDFs. Compounds belonging to the training and validation sets are marked with filled circles and are numbered as in Table 7.

to construct a 2^{4-1} FFD, encoding ten compounds (eight corner compounds and two center points). A similar design, i.e. the other half-fraction of the full factorial design, was also formed to aid the selection of ten congeners for the validation set. In the validation set, an additional center point and octachlorodibenzofuran (OCDF) were included. OCDF was included, because it exhibited atypical chemical properties. In total, 20 compounds were selected out of the 87 PCDFs for biological testing (Table 7). The compounds were tested for ethoxyresorufin-O-deethylase (EROD)

Table 7. The 2^{4-1} FFD for the training set (top) and the 2^{4-1} FFD for the validation set (bottom). The dibenzofuran example

FFD				no.	Compound	Settings in PPs			
*PP*1	*PP*2	*PP*3	*PP*4			*PP*1	*PP*2	*PP*3	*PP*4
−	−	−	−	**52**	1,2,3,7-TCDF	−2.40	−1.5	−1.43	−1.18
+	−	−	+	**115**	2,3,4,7,8-PeCDF	1.82	−2.43	−0.82	1.23
−	+	−	+	**61**	1,2,6,9-TCDF	−2.78	2.59	−0.29	0.22
+	+	−	−	**135**	1,2,3,4,7,8,9-HpCDF	6.17	1.22	−1.07	−0.19
−	−	+	+	**86**	2,4,6,8-TCDF	−1.68	−0.57	1.46	2.94
+	−	+	−	**129**	1,3,4,6,7,8-HxCDF	2.60	−1.34	1.97	−0.86
−	+	+	−	**76**	1,4,6,9-TCDF	−3.46	2.17	1.60	−0.98
+	+	+	+	**128**	1,2,4,6,8,9-HxCDF	2.84	2.21	1.43	1.13
0	0	0	0	**104**	1,2,6,7,8-PeCDF	0.63	−0.07	0.67	0.46
0	0	0	0	**113**	2,3,4,6,7-PeCDF	0.36	−1.05	−0.27	0.70
−	−	−	+	**83**	2,3,6,8-TCDF	−0.21	−3.01	−1.43	2.14
+	−	−	−	**125**	1,2,3,7,8,9-HxCDF	3.79	−0.48	−3.27	−0.40
−	+	−	−	**91**	1,2,3,4,9-PeCDF	−0.71	2.50	−1.82	−0.22
+	+	−	+	**103**	1,2,4,8,9-PeCDF	0.30	2.50	−0.41	1.30
−	−	+	−	**70**	1,3,6,8-TCDF	−1.62	−1.37	1.43	−0.16
+	−	+	+	**126**	1,2,4,6,7,8-HxCDF	2.79	−0.42	2.19	0.70
−	+	+	+	**60**	1,2,6,8-TCDF	−2.12	0.26	1.06	1.66
+	+	+	−	**127**	1,2,4,6,7,9-HxCDF	2.61	1.38	1.46	−0.90
+ +	+	0	0	**136**	1,2,3,4,6,7,8,9-OCDF	8.64	1.09	0.17	−0.36
0	0	0	0	**88**	1,2,3,4,6-PeCDF	−0.73	0.76	0.38	0.20

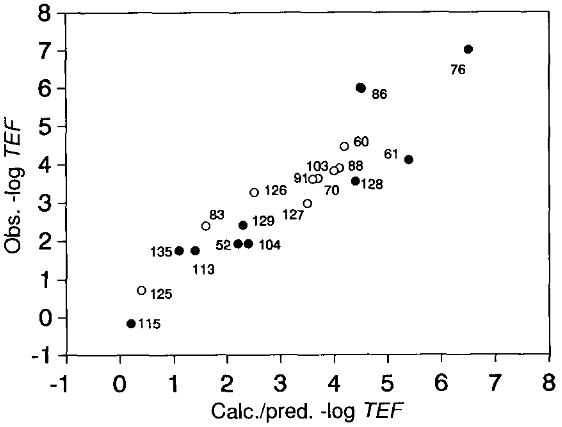


Figure 13. Observed toxic equivalency factor (−log *TEF*) plotted against calculated (filled circles) and predicted (open circles) *TEF* values. Notation as in Fig. 7.

induction in the H4IIE rat hepatoma cell bioassay [24]. The measured EROD induction potencies were converted to toxic equivalency factors (*TEF*) by calibration against the corresponding biological activity of the most potent known compound 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD). In the QSAR analysis, the negative logarithm of this *TEF* scale was used.

By means of PLS and using 37 chemical descriptors [25], a multivariate QSAR for the *TEF* scale was calculated, which was based on the ten training set compounds. This QSAR was effective in describing and predicting the variation in EROD induction potencies ($R^2 = 0.84$ and $Q^2 = 0.69$). Using the QSAR for predicting the biological activities of the validation set chemicals, resulted in an external $Q^2 = 0.81$, which strongly underpins the good predictive capability of this QSAR. Fig. 13 illustrates the relationship between the observed and calculated/predicted *TEF* values. It is evident that the model may be useful for predicting *TEF* values for the 67 non-tested congeners, which belong to this class of compounds.

3.2.3.7 Monosubstituted Benzenes

Skagerberg et al. [26] have determined PPs for one hundred monosubstituted aromatics. The compounds, which cover four types of electronically different substituents, i.e. electron acceptors and donors, alkyl groups and halogens, were multivariately characterized by means of nine physico-chemical descriptors. The descriptors used were π, *MR*, σ_m, σ_p, the Verloop parameters *L* and *B*1 − *B*4. PCA of the resulting 9 × 100 data matrix gave four principal components (PPs), reflecting 76% of the variance (Fig. 14). Tosato et al. [27] have used these PPs in statistical experimental design for setting priorities and conducting hazard assessments for monosubstituted benzene derivatives. The three first PCs were considered the most important and were used in a 2^3 full factorial design. This scheme encoded eight training set compounds as good representatives of all the other compounds (Table 8). Moreover, to allow for a verification of the QSAR models developed, a set of 11 additional compounds was selected to constitute the validation set. The validation

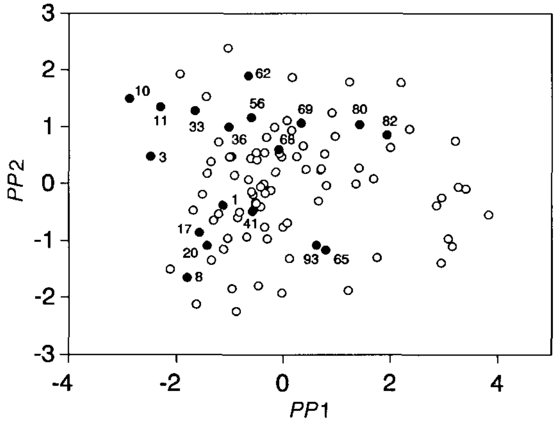


Figure 14. Scatter plot for the PPs of the 100 monosubstituted benzenes. Compounds, which have been biologically tested for EC_{50} values, are depicted with filled circles and are numbered according to Table 8.

set compounds were selected in such a way as "... to allow evaluation of the range of validity of the model and of its actual predictive capacity" [27]. Both the training set compounds and the validation set compounds were tested in the laboratory of Tosato and coworkers, using an assay called the Daphnia immobilization test (adjusted according to the relevant OECD guideline). The endpoint determined was the concentration causing a 50% effect (EC_{50}) and in the QSAR analyzes, the transformation log $1/EC_{50}$ was used. A PLS analysis of the training set, characterized by the nine aromatic descriptors, gave four latent variables with $R^2 = 0.99$ and

Table 8. The 2^3 FD for the training set (top) and the compounds in the validation set (bottom). The monosubstituted benzenes example

FD			no.	Compound	Settings in PPs		
*PP*1	*PP*2	*PP*3			*PP*1	*PP*2	*PP*3
−	−	−	**8**	Ph-NO_2	−1.80	−1.66	−0.03
+	−	−	**65**	Ph-$CO_2C_2H_5$	0.79	−1.16	−0.23
−	+	−	**10**	Ph-H	−2.88	1.49	−0.54
+	+	−	**82**	Ph-OC_4H_9	1.93	0.86	−0.31
−	−	+	**5**	Ph-Br	−0.47	−1.80	0.87
+	−	+	**93**	Ph-COC_6H_5	0.62	−1.09	1.70
−	+	+	**33**	Ph-CH_3	−1.66	1.28	0.28
+	+	+	**80**	Ph-*n*-C_4H_9	1.42	1.04	0.31
			41	Ph-SCH_3	−0.50	0.42	0.17
			56	Ph-C_2H_5	−0.60	1.16	0.26
			68	Ph-$CH(CH_3)_2$	−0.09	0.60	1.38
			69	Ph-C_3H_7	0.33	1.06	0.28
			20	Ph-CN	−1.43	−1.09	−0.85
			11	Ph-OH	−2.30	1.34	−0.60
			3	Ph-F	−2.48	0.48	−0.34
			62	Ph-$N(CH_3)_2$	−0.66	1.89	0.47
			17	Ph-CF_3	−1.57	−0.86	0.83
			36	Ph-OCH_3	−1.02	1.00	−0.34

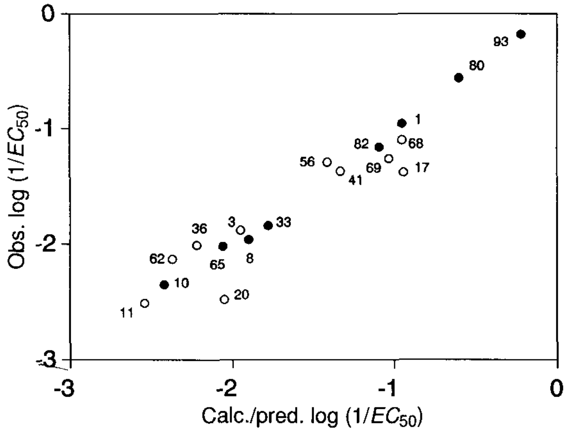


Figure 15. Scatter plot for the QSAR of the monosubstituted benzenes showing observed log (1/EC_{50}) plotted against the corresponding calculated/predicted values. The compounds are numbered as in Table 8. Open circles correspond to the validation set compounds, whereas filled circles correspond to the compounds used for model-building (training set).

$Q^2 = 0.64$. Fig. 15 shows the correlation plot between the observed and calculated/predicted biological activities. Evidently, the model predicts rather well, even though some of the validation set compounds were chosen deliberately, so as to rigorously test the predictive capability of the model.

3.2.3.8 Corrosive Carboxylic Acids

Our last example deals with a series of 45 aliphatic carboxylic acids and their corrosive effects towards rabbit skin. This study is of interest, because skin effects caused by corrosive chemicals are a frequently reported occupational hasard, and because many, suspectedly corrosive, compounds are commonly involved in industrial handling and transportation. Furthermore, this study is of relevance, because, ţo our knowledge, QSAR techniques have only been applied to a limited extent to modeling corrosivity endpoints [28]. In this particular investigation, the corrosive effects towards rabbit was selected as a representative biological model

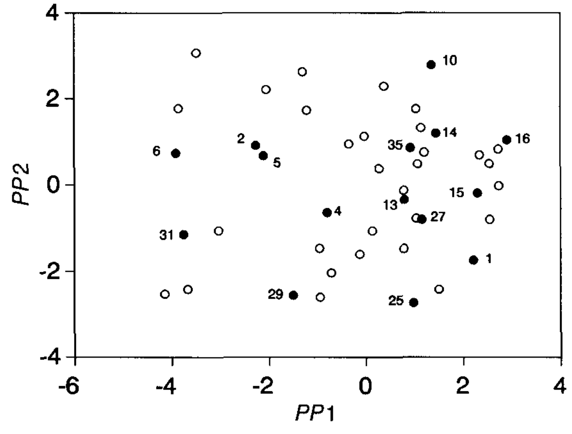


Figure 16. Scatter plot of *PP*2 vs *PP*1 for the corrosive carboxylic acids. The training and validation set carboxylic acids are designated by filled circles. Notation as in Table 9.

Table 9. The 2^3 FD for the training set (top) and the compounds in the validation set (below). The carboxylic acids example

FD			no.	Compound	Settings in PPs		
*PP*1	*PP*2	*PP*3			*PP*1	*PP*2	*PP*3
−	−	−	**31**	Maleic acid	−3.74	−1.14	1.06
+	−	−	**1**	Acetic acid	2.22	−1.74	−0.14
−	+	−	**6**	Trichloroacetic acid	−3.91	0.74	−1.55
+	+	−	**35**	4-Chlorobutyric acid	0.93	0.86	0.26
−	−	+	**4**	Chloroacetic acid	−0.79	0.63	0.09
+	−	+	**15**	Butyric acid	2.30	−0.18	−0.75
−	+	+	**2**	Bromoacetic acid	−2.26	0.92	0.61
+	+	+	**10**	Mercaptoacetic acid	1.37	2.79	2.21
0	0	0	**13**	3-Chloropropionic acid	0.80	−0.33	−0.07
			29	Malonic acid	−1.49	−2.56	1.17
			5	Dichloroacetic acid	−2.10	0.69	−1.11
			14	Methacrylic acid	1.46	1.20	0.37
			27	2-Hydroxybutyric acid	1.16	−0.80	0.53
			16	Vinylacetic acid	2.92	1.04	0.44

system. Initially, we described the 45 acids with a multivariate set of nine variables (molecular weight, melting point, density, refractive index, octanol/water partition coefficient, pK_a, energy of highest occupied and lowest unoccupied molecular orbital, and electronegativity). PCA of this multiproperty matrix yielded three PPs describing 74% of the variance, which were used to derive a 2^3 FD, supplemented with one interior point. In Fig. 16, the two first PPs are plotted against each other. Moreover, five compounds were selected for the verification set. The nine acids in the training set and the five acids in the validation set (Table 9) were tested biologically, to determine the lowest concentration at which signs of cutaneous corrosion could be found, i.e. the lowest-observed-effect-concentration (*LOEC*). Since strong non-

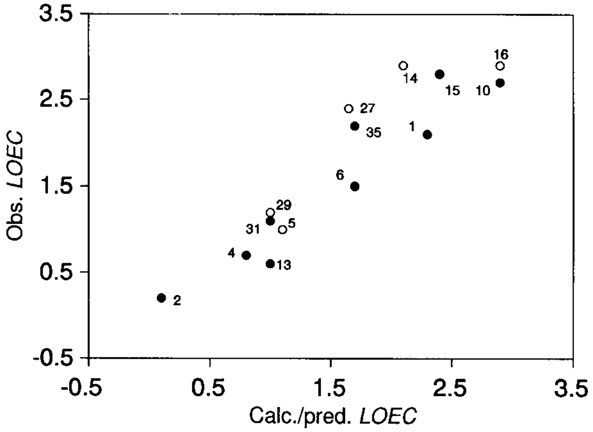


Figure 17. Observed *LOEC* plotted against calculated/predicted values for the corrosivity QSAR. The compounds used to calibrate the model (training set) are marked with filled circles, and the validation set is marked with open ones. See Table 9 for the numbering.

linearities were found in the existing relationship between the chemical structure and the biological activity, linear *and* quadratic terms (in total 18 variables) were used as predictor variables in the PLS modeling. This resulted in a QSAR with $R^2 = 0.83$ and $Q^2 = 0.59$. In Figure 17, the agreement between the observed and calculated/predicted LOEC values for the 15 tested compounds is shown. The external $Q^2 = 0.60$ is based on the performance of the validation set compounds. It is interesting to note, that in the ideal situation, both the external (validation set) and internal (cross-validation) Q^2 should be of a similar magnitude, since they mirror the same unknown predictive capability statistic. The conclusion from this study, is that it is possible to derive multivariate QSAR models which provide useful predictions for endpoints related to skin corrosion.

3.2.4 Discussion and Conclusions

Quantitative structure-activity relationships are valuable tools for modeling and predicting the biological responses of chemical compounds, and for the identification of potential structures with optimized biological properties. It is important to realize that QSAR modeling is not only restricted to small and semi-rigid molecules, but as shown here, it can be applied to long-chain and highly flexible chemical structures, such as peptides. Thus, QSAR modeling of flexible peptide sequences need not necessarily require a knowledge of the 3D structure of the compounds.

The use of projection methods, such as PLS, which can deal with multivariate data, will result in models with few descriptive components that are easy to interpret. However, the process, leading to useful models, consists of a number of important stages. Those aspects that deserve special attention were discussed in each one of the six consecutive steps of the multivariate strategy for QSAR. Here, it is particularly important that the compounds, which are used to calibrate the model, i.e. the training set, and the chemicals used to verify the predictive capability of the QSAR experimentally, are selected at least by approximate statistical experimental design. Such design can be achieved by using principal variables or PPs, defined for the class of compounds of interest. Finally, the compiled data, both chemical and biological, can be analyzed preferably using multivariate projection methods, such as PLS, which provides information about the structure of the data and the range of validity of the model. In this contribution, the main emphasis has been placed on demonstrating the benefits of using statistical experimental design in the selection of test series. This was discussed in connection with the dipeptide example in Sec. 3.2.3.2, where the quality of QSAR models, based on either a statistically or a COST-designed training set, are in sharp contrast to each other. The QSAR, based on the statistical design approach, was superior from the point of view of predictive capability. These results also revealed the consequences of not using statistical experimental designs in the compilation of the training set. Statistical experimental designs guarantee that many latent variables are varied systematically – in a balanced manner – and simultaneously, which is not the case if only one latent variable is modified at a time.

Since the main objectives of this chapter have been to shed some light on the use of statistical experimental design and PLS analysis, examples have been selected that best illustrate these important steps. Thus, the QSARs have been introduced and discussed, mostly from a statistical and technical point of view, and perhaps not so much in detail as regards to the interpretation of their meaning. It is, however, appropriate to stress that once a QSAR has been developed, it is important to interpret the significance of this relationship. Since QSARs can be interpreted as mathematical approximations of underlying fundamental relationships, their coefficients sometimes provide clues for mechanistic interpretations. A QSAR is well-founded, when the feature of the model is consistent with the mechanistic interpretation. As regards to the interpretations of the QSAR examples presented, the reader is referred to the original literature.

We regard QSAR modeling as a special case of semi-empirical modeling, typically leading to linear or low order polynomial expressions. Simple statistical rules for the validity of semi-empirical models, thus, also apply to QSARs, and can be used as guidelines to construct QSARs with valid predictive capabilities. These conditions have been considered in the development of the QSAR framework discussed in this chapter. To conclude, we propose that the training set of compounds for screening purposes and QSAR modeling should always be constructed according to a statistical experimental design. This, together with a multivariate representation of the chemical and biological properties of the studied substances, will strongly increase the information content in the training set series and will increase the efficiency, and the chances of success in drug development.

Software Used

In all the calculations SIMCA-4.41 for PC was used. SIMCA is available from Umetri AB, Box 1456, S-901 24 Umeå, Sweden.

Acknowledgements

Financial support from the Swedish Natural Science Research Council (NFR), the Centre for Environmental Research (CMF) in Umeå, the Swedish National Chemical Inspectorate (KemI), Räddningsverket and the Commission of the European Communities Environmental Technologies RTD Programme (DGXII/D-1) under contract number EV5V-CT92-0211 was gratefully received. The authors would like to thank Douglas Baxter for his helpful comments in the preparation of this manuscript.

References

[1] Pleiss, M. A., and Unger, S. H., The Design of Test Series and the Significance of QSAR Relationships. In: *Comprehensive Medicinal Chemistry. The Rational Design, Mechanistic Study and Therapeutic Applications of Chemical Compounds*, Vol **4**, *Quantitative Drug Design*, Hansch, C., Sammes, P. G., Taylor, J. B., and Ramsden, C. A. eds., Pergamon Press, Oxford (1990) p. 561–587

[2] Tosato, M. L., Marchini, S., Passerini, L., Pino, A., Eriksson, L., Lindgren, F., Hellberg, S., Johnsson, J., Sjöström, M., Skagerberg, B., and Wold, S., *Environ. Toxicol. Chem.* **9**, 265–277 (1990)

[3] Wold, S., and Dunn, III W. J., *J. Chem. Inf. Comp. Sci.* **23**, 6–13 (1993)

[4] Hellberg, S., *A Multivariate Approach to QSAR*, Ph.D. Thesis, University of Umeå, Umeå, 1986

[5] Eriksson, L., *A Strategy for Ranking Environmentally Occurring Chemicals*, Ph.D. Thesis, University of Umeå, Umeå, 1991

[6] Dunn, III, W. J., *Chemom. Intell. Lab. Syst.* **6**, 181–190 (1989)

[7] Carlson, R., *Design and Optimization in Organic Synthesis*, Elsevier, Amsterdam, 1992

[8] Baroni, M., Clementi, S., Cruciani, C., Kettaneth-Wold, M., and Wold, S., *Quant. Struct.-Act. Relat.* **12**, 225–231 (1993)

[9] Box, G. E. P., Hunter, W. G., and Hunter, J. S., *Statistics for Experimenters*, Wiley, New York, 1978

[10] Box, G. E. P., and Draper, N. R., *Empirical Model-Building and Response Surfaces*, Wiley, New York, 1987

[11] Wold, S., *Technometrics* **20**, 397–405 (1978)

[12] Ufkes, J. G. R., Visser, B. J., Heuver, G., and van der Meer, C., *Eur. J. Pharmacol.* **50**, 119–122 (1978)

[13] Ufkes, J. G. R., Visser, B. J., Heuver, G., and van der Meer, C., *Eur. J. Pharmacol.* **79**, 155–158 (1982)

[14] Hellberg, S., Sjöström, M., and Wold, S., *Acta Chem. Scand.* **B 40**, 135–140 (1986)

[15] Hellberg, S., Sjöström, M., Skagerberg, B., and Wold, S., *J. Med. Chem.* **30**, 1126–1135 (1987)

[16] Hellberg, S., Sjöström, M., Skagerberg, B., Wikström, C., and Wold, S., *Acta Pharm. Yugosl.* **37**, 53–65 (1987)

[17] Hellberg, S., Eriksson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B., Wold, S., and Andrews, P., *Int. J. Peptide Protein Rws.* **37**, 414–424 (1991)

[18] Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M., and Wold, S., *Quant. Struct.-Act. Relat.* **8**, 204–209 (1989)

[19] Cheung, H.-S., Wang, F.-L., Ondetti, M. A., Sabo, E. F., and Cuchman, D. W., *J. Biol. Chem.* **255**, 401–407 (1980)

[20] Asao, M., Iwamura, H., Akamatsu, M., and Fujita, T., *J. Med. Chem.* **30**, 1873–1879 (1987)

[21] Fincham, C. I., Higginbottom, M., Hill, D. R., Horwell, D. C., O'Toole, J., Ratcliffe, G. S., Rees, D. C., and Roberts, E., *J. Med. Chem.* **35**, 1472–1484 (1992)

[22] Berglund, A., Eriksson, L., Johansson, E., and Sjöström, M., manuscript in preparation

[23] Eriksson, L., Sandström, M., Sjöström, M., Tysklind, M., and Wold, S., *Quant. Struct.-Act. Relat.* **12**, 124–131 (1993)

[24] Tysklind, M., *Multivariate Chemical Characterization and Modelling of Polychlorinated Dioxines and Dibenzofurans*, Ph.D. Thesis, University of Umeå, Umeå, 1993

[25] Tysklind, M., Tillitt, D., Eriksson, L., Lundgren, K., and Rappe, C., *Fundam. and Appl. Toxicol.* **22**, 227–285 (1994)

[26] Skagerberg, S., Bonelli, D., Cruciani, G., and Ebert, C., *Quant. Struct.-Act Relat.* **8**, 32–38 (1989)

[27] Tosato, M. L., Pino, A., Passerini, L., Marachini, S., Vigano, L., and Hoglund, M. D., *Sci. Total. Environ.*, submitted 1992

[28] Eriksson, L., Berglind, R., and Sjöström, M., *Chemom. Intell. Lab. Syst.*, **23**, 235–245 (1994)

3.3 Total Response Surface Optimization

Lowell H. Hall

Symbols

${}^m\chi_t^v$	Molecular connectivity *chi* index of order, *m*, and type, *t*; with the 'v' supercript, the valence type index is meant; without the 'v', the simple index is intended
pC	Negative logarithm of the concentration required to achieve a standard biological effect, commonly stands for activity
u_i	Any structure descriptor which may be used in a QSAR equation
$\Sigma x0$	$= {}^0\chi + {}^0\chi^v$, the sum of zeroth order *chi* indexes
$\Delta x0$	$= {}^0\chi - {}^0\chi^v$, the difference between zeroth order *chi* indexes; called delta *chi* zero
$\Sigma x1$	$= {}^1\chi + {}^1\chi^v$, the sum of first order *chi* indexes
$\Delta x1$	$= {}^1\chi - {}^1\chi^v$, the difference between first order *chi* indexes; called delta *chi* one

3.3.1 Background

It has been observed, that, for a series of biologically active molecules, the difference in structure from one molecule to another corresponds to a change in the biological response. When there is systematic variation in molecular structure, there is also systematic variation in biological activity. The QSAR paradigm is based on the assumption that there is a relationship between molecular structure and biological activity, which arises from this systematic variation. This relation is a specific manifestation of the form-function relation which is well-known in science. Capturing the meaning of systematic variation has been the chief problem in QSAR pursuits.

There are two broad classes of QSAR approaches. In one class, the relationship is derived from linear free energy relations. In this approach physical properties are used to represent molecules and to relate to an experimental measure of the activity in a linear fashion [1]. In a methodology developed over the past 18 years, more direct representation of the molecular structure have been achieved. In this structure-based approach, descriptors represent important features of molecular structure. A set of structure indexes has been developed from chemical graph theory and will be described in a subsequent section [2–6].

In the usual QSAR method, a structure variable, u_i (physical property or structure descriptor) is assumed to have a linear relation to activity:

$$\mathrm{p}C = a_1 u_1 + c \tag{1}$$

It has been observed, however, that the relation between biological activity and structure is not always linear, especially over a wide range of activity. As molecular size increases, for example, the biological effect often increases to a maximum value, and then decreases. This effect may be attributed to several phenomena, including decreasing solubility and chemical activity, differential lipid transport, or size effects at the receptor or enzyme active site. Such non-linear effects are often represented by a simple quadratic expression:

$$\mathrm{p}C = a_1 u_1 + a_2 u_1^2 + c \tag{2}$$

Many useful non-linear equations of this type have been reported [1, 7, 11].

3.3.2 Representation of a Response Surface

When two or more structure variables are required to represent the variation in the list of molecules, the non-linear equation becomes more complicated. In addition to linear and squared terms, there are cross terms in the structure variables. For a two-variable case:

$$\mathrm{p}C = a_1 u_1 + b_2 u_1^2 + a_2 u_2 + b_2 u_2^2 + c u_1 u_2 + d \tag{3}$$

The term "response surface" arises because of the contours generated in a plot of u_1 versus u_2. The right hand side of Eq. (3) expresses a general parabolic surface; one can draw contours at levels of constant activity. Examples are given in Fig. 1 and 2. The contours are elliptical in shape; each ellipse may be described by two axes. Because Eq. (3) represents a parabolic shape, the surface possesses either a maximum or minimum, corresponding to the extremum value of pC. In general, the extremum point does not coincide with the u_1, u_2 origin. Further, the ellipse axes are not parallel to the u_1, u_2 axes. The investigator may choose either to transform the data or to change the origin and to align the axes; such changes are not necessary for some aspects of QSAR, but do provide some additional information.

If it is desired to put the contour ellipses into a standard form, it is possible to perform two transformations on the general quadratic form of Eq. (3). By obtaining the position of the extremum (maximum or minimum), one can translate the origin to the location of the extremum point. The mathematical consequence of this transformation, is that the linear terms disappear from the equation. Let $u_{1,\mathrm{ext}}$ be the position of the optimum activity (extremum point) for variable 1, and for variable two, $u_{2,\mathrm{ext}}$. Then, let the transformed variables be $q_1 = u_1 - u_{1,\mathrm{ext}}$ and $q_2 = u_2 - u_{2,\mathrm{ext}}$. It, thus, follows that,

$$\mathrm{p}C = b_1 q_1^2 + b_2 q_2^2 + b_3 q_1 q_2 + d \tag{4}$$

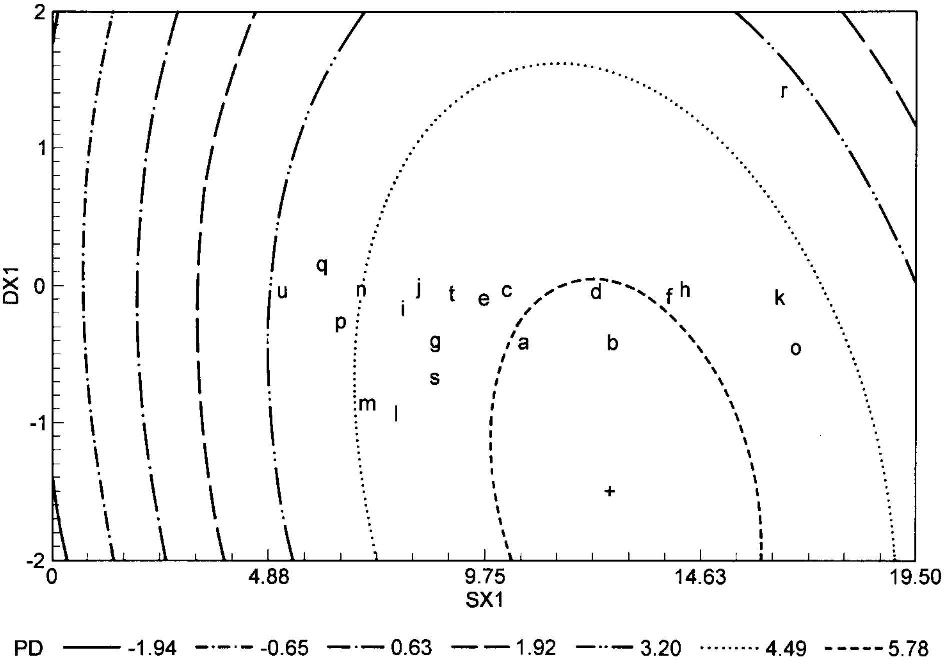


Figure 1. A plot of $\Delta x1$ (labeled Dx1) versus $\Sigma x1$) for the neurotoxicity data showing the relation between the measure of molecular size ($\Sigma x1$) and molecular electronic structure ($\Delta x1$) with contour intervals for pD. The contours intervals are for pD = −1.94, −0.65, 0.63, 1.92, 3.20, 4.49, and 5.78 (for the smallest oval). The points are labeled for the compounds as shown under plot symbol in Table 1. The point of maximum activity is labeled with a + sign.

Finally, the general parabolic surface is characterized by the axes of the elliptical contours. In general, the axes are not parallel to the u_1, u_2 axes. These axes are rotated by an angle θ. It is possible to find a rotation matrix in order to rotate this surface, so that the axes are parallel to the u_1, u_2 axes. This transformations $\boldsymbol{B}$ is based on the eigenvectors of the matrix of coefficients in Eq. (5):

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{b}_1 & 1/2\boldsymbol{b}_3 \\ 1/2\boldsymbol{b}_3 & \boldsymbol{b}_2 \end{pmatrix}$$

The mathematical consequence of this operation, is that the cross term disappear. The final form is then:

$$\mathrm{p}C = b_1 q_1'^2 + b_2 q_2'^2 + d \tag{5}$$

in which q_1' and q_2' are the rotated coordinates with the origin at the surface extremum point.

This analysis permits a simplified representation on the data space. The structure-activity relationship under consideration, which is assumed to be quadratic in two variables, actually has the appearance of two squared variables, shown in Eq. (5), as compared to the more complicated form in Eq. (3). This analysis may seem

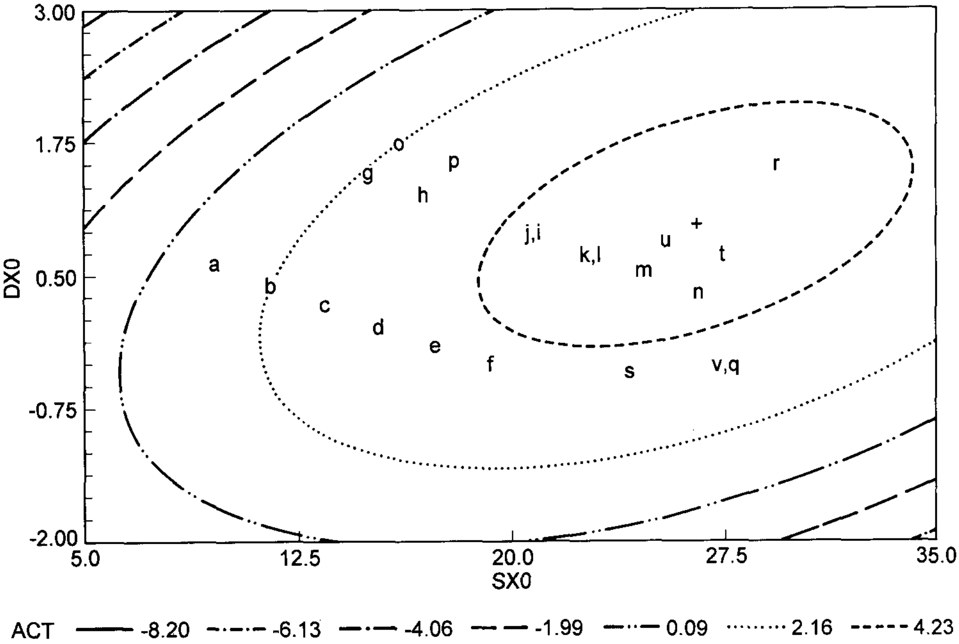


Figure 2. A plot of $\Delta x0$ (labeleld $Dx0$) versus $\Sigma x0$ (labeleld $Sx0$) for the bioconcentration data, showing the elliptical contours for pBCF with the contour intervals shown at -8.20, -6.13, -4.06, -1.99, 0.09, 2.16, and 4.23 (the smallest oval). The data points are shown as lower case letters corresponding to the plot symbol in Table 2. The maximum point is labeled as '+'. The pair of points for **i** and **j** occur at the same coordinates as **k** and **l** as well as **q** and **v**.

to be rather cumbersome and, if it must be undertaken manually, it may not appear worth the effort. Fortunately, most current statistical packages offer routines which perform the complete analysis along with other diagnostic information. For example, SAS carries out all the necessary operations in what is called proc RSREG [8]. This algorithm determines whether the optimum point is a minimum or a maximum point. Further, the eigenvectors may also be given along with the transformation equations. The statistical package may also warn the user, when the extremum point is actually a saddle point, or that there is insufficient information to establish the extremum point. All this information is determined, along with the usual statistical information, including regression coefficients, correlation coefficients, standard deviation and, so forth.

3.3.3 Structure Descriptors from Chemical Graph Theory

In the analysis of a response surface, any of the indexes from chemical graph theory can be used as the structure variables, including molecular connectivity *chi* indexes, *kappa* shape indexes (See Chap. 2.1), electropological state indexes (See Chapt. 2.2), and others [10]. For this present discussion, we shall be using molecular connectivity *chi* indexes. There are several detailed presentations of the development of these indexes and reviews of their uses [2, 3, 5, 6]. A very brief presentation is given in Chap. 2.1.

3.3.4 Examples

3.3.4.1 Neurotoxicity of Fluorophosphorous Compounds

Mager [10] has reported the neurotoxicity of 21 fluorophosphates, fluorophosphonates and fluorophosphorodiamides. In this data set, there are cyclic and non-cyclic alkyl portions of esters and amides along side one aromatic substituent. The activity to produce ataxia in adult white hens was reported as mg/kg. The doses were expressed on a molar basis and converted to the negative logarithm, p*D*, which ranges from 3.20 to 5.92. The data are shown in Table 1. The molecular connectivity indexes were computed in the standard manner with the aid of the computer program Molconn-X, Version 2.0 [9]. All statistical analyzes were carried out using the SAS statistical system [8].

The first order simple and valence *chi* indexes, ${}^1\chi$ and ${}^1\chi^v$, gave an excellent QSAR equation:

$$\begin{gathered} \mathrm{p}D = 1.265\,{}^1\chi - 0.286({}^1\chi)^2 + 1.124\,{}^1\chi^v - 0.188({}^1\chi^v)^2 + 0.275\,{}^1\chi\,{}^1\chi^v + 0.661 \\ r^2 = 0.959 \qquad s = 0.177 \qquad F = 70 \qquad n = 21 \end{gathered} \tag{6}$$

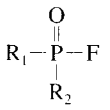
The observed, calculated and residual toxicities are given in Table 1. A partial report on this analysis has already been given [11].

It is important to be able to obtain significant structure information from the QSAR equation. It was possible to give a structure interpretation for this data set directly from the two first order *chi* indexes. In this case, however, a somewhat different form of analysis will be developed. The ${}^1\chi$ and ${}^1\chi^v$ variables encode information about both molecular size and molecular electronic structure. At this point in the analysis, it is useful to separate these two types of information. Such a separation may be accomplished by a simple transformation:

$$\begin{aligned} \Sigma x1 &= {}^1\chi + {}^1\chi^v \\ \Delta x1 &= {}^1\chi - {}^1\chi^v \end{aligned} \tag{7}$$

These sum and diffierence variables have the following properties [5]. In simple molecules, such as alkyl alcohols, halides, amines, etc., it can be shown that ${}^1\chi^v$ may be partitioned into two terms, one arising from the functional group X_{fg}, and the other from the alkyl portion, X_{ag}. Such a partition is possible because of the additive nature of the *chi* index. In the ${}^1\chi$ index, on the other hand, all the terms are calculated, as if the whole molecule were an alkane. Thus, ${}^1\chi$ can be written as a sum of two terms, X_{ag}, for the skeletal portion of the functional group, and X_{alk}, for the remainder of the molecule. The term, X_{alk} is the same for the two indexes.

Table 1. The observed, calculated and residual neurotoxicity values for the data set of fluorophosphate compounds



Obs.	Compound substitutents (R_1, R_2)	Plot Symbol[a]	$^1\chi$[b]	$^1\chi$[c]	p*D*[d]	*Calc*[e]	*Res*[f]	*Pres*[g]
1	Di(propylamino)	**a**	5.121	5.518	5.86	5.76	0.10	0.12
2	Di(butylamino)	**b**	6.121	6.518	5.92	5.87	0.05	0.06
3	Di(propoxy)	**c**	5.121	5.142	5.66	5.56	0.10	0.12
4	Dibutoxy	**d**	6.121	6.142	5.63	5.71	−0.08	−0.09
5	Diisopropoxy	**e**	4.833	4.931	5.57	5.49	0.08	0.09
6	Di-*sec*-pentoxy	**f**	6.909	7.007	5.43	5.58	−0.15	−0.18
7	Diethylamino)	**g**	4.121	4.518	5.11	5.25	−0.14	−0.15
8	Dipentoxy	**h**	7.121	7.142	5.26	5.46	−0.20	−0.26
9	*N*,*N'*-dimethyl, ethoxy	**i**	3.887	4.044	5.00	4.92	0.08	0.09
10	Diethoxy	**j**	4.121	4.142	5.07	5.01	0.06	0.07
11	Dicyclohexoxy	**k**	8.157	8.255	5.20	4.81	0.39	0.82
12	Isopropoxy, methyl	**l**	3.417	4.335	4.94	4.92	0.02	0.03
13	Ethoxy, methyl	**m**	3.061	3.940	4.62	4.55	0.07	0.13
14	*N*,*N'*-Dimethyl, methoxy	**n**	3.504	3.505	4.67	4.45	0.22	0.25
15	Dicyclohexamino)	**o**	8.157	8.607	4.72	4.88	−0.16	−0.68
16	Di(methylamino)	**p**	3.121	3.397	4.09	4.25	−0.16	−0.22
17	Dimethoxy	**q**	3.121	2.967	4.03	3.91	0.12	0.21
18	Di(2-Methylphenylamino)	**r**	8.978	7.551	3.23	3.25	−0.002	−1.60
19	Isopropoxy, ethyl	**s**	3.977	4.649	5.29	5.28	0.01	0.01
20	Isopropoxy, ethoxy	**t**	4.477	4.536	5.07	5.27	−0.20	−0.23
21	Hydroxy, methoxy	**u**	2.561	2.611	3.20	3.38	−0.18	−0.31

[a] Symbol used in plot in Fig. 1.
[b] First order molecular connectivity index.
[c] First order valence molecular connectivity index.
[d] Negative logarithm of the neurotoxicity (mg/kg) [9].
[e] p*D* calculated from Eq. (6) or Eq. (10).
[f] p*D* − *Calc*.
[g] The predicited residual, obtained by deleting the observation and then predicting its value from the remaining $n - 1$ observations.

Using the sum and difference variables and the partitions, described above, we have the following equations

$$\Sigma x1 = {}^1\chi + {}^1\chi^{\mathrm{v}} = (X_{\mathrm{ag}} + X_{\mathrm{alk}}) + (X_{\mathrm{fg}} + X_{\mathrm{alk}}) = 2X_{\mathrm{alk}} + (X_{\mathrm{ag}} + X_{\mathrm{fg}}) \qquad (8)$$

$$\Delta x1 = {}^1\chi - {}^1\chi^{\mathrm{v}} = (X_{\mathrm{ag}} + X_{\mathrm{alk}}) - (X_{\mathrm{fg}} + X_{\mathrm{alk}}) = X_{\mathrm{ag}} - X_{\mathrm{fg}} \qquad (9)$$

These two transformed structure variables may be understood as follows. First, consider the difference variable, $\Delta x1$. The terms for the saturated portions of a molecule, X_{alk}, disappear entirely from the $\Delta x1$ variable. Furthermore, any functional group or unsaturated portion is represented by $X_{\mathrm{ag}} - X_{\mathrm{fg}}$. This difference accentuates the electronic contribution of heteroatoms or non-sigma bonding. X_{ag} arises from the calculation of $^1\chi$, as if the atoms were saturated carbon atoms, that is, as if they possessed only sigma electrons. Since X_{fg} encodes both the sigma and the non-sigma eletrons, the difference between X_{ag} and X_{fg}, the $\Delta x1$ variable, encodes only the presence of lone pair and pi electrons. Thus, the difference term encodes electronic effects due to heteroatoms and unsaturation in the molecule.

Now, let us consider the sum variable, $\Sigma x1$. In the sum variable $\Sigma x1$, the common portion of the two *chi* indexes, X_{alk}, is emphasized by a multiplication by 2, and the functional group portion, X_{fg}, is augmented by X_{ag}. These terms encode much size information and, augmented as they are in contrast to $\Delta x1$, this variable is largely a measure of the molecular size. The heteroatom contribution to size is expressed by ${}^{1}\chi^{\mathrm{v}}$ in the X_{fg} term. Skeletal size is encoded in both X_{alk} and X_{ag}.

In terms of molecular electronic structure, the difference between sp^3 carbon atoms and heteroatoms or unsaturated carbon atoms, is in the nature of the electron distribution. In saturated carbon atoms, valence electrons are found in only sigma-type orbitals, whereas in the functional groups, valence electrons are also involved in pi and lone pair orbitals. The difference variable, $\Delta x1$, is a descriptor of the structural contribution of the electrons in pi and lone pair orbitals.

The sum variable, $\Sigma x1$, encodes the whole molecular skeleton, including functional groups. Based on the summation of all skeletal contributions and reflecting contributions of all atoms, $\Sigma x1$ is expected to correlate with molecular size. To whatever extent size is an important factor in biological relationships, such a sum variable was found to be important. The difference variable encodes non-sigma electrons, which are those electrons, which exert a major influence on chemical interactions. For singly bonded nitrogen, oxygen and fluorine, $\Delta x1$ encodes lone pair electrons.

For multiply bonded nitrogen, oxygen, phosphorous and for aromatic rings, $\Delta x1$ encodes electrons in pi orbitals. Thus, use of these two variables permits the factoring of structure information into molecular size effects and molecular electronic effects.

There is another important characteritic of these two transformed variables: $\Sigma x1$ and $\Delta x1$ are orthogonal. Orthogonality here depends upon two factors. The original variables, the *chi* indexes, must not be collinear, and they must be of a similar magnitude. A simple scaling factor can always ensure that the two *chi* indexes are approximately of the same magnitude. This orthogonality is very useful in QSAR because it eliminates some statistical problems. A similar transformation can be performed on physical property variables such as log P and MR (molar refraction), as was achieved by Mager in his analysis of this neurotoxicity data set. However, the transformed variables, linear combinations of log P and MR, have no physical meaning. Thus, use of this sum and difference transformation aids the statistical analysis, but confuses the interpretation. *Chi* indexes enhance both statistical analysis and structure interpretation.

The biological activity may be regressed against the two transformed structure variables in the full quadratic model, which represents the biological response surface. Since the transformation is linear, the statistical results are the same, but the coefficients are different:

$$\begin{aligned} \mathrm{p}D = {} & 1.195\,\Sigma x1 - 0.049\,\Sigma x1^2 + 0.071\,\Delta x1 - 0.187\,\Delta x1^2 \\ & - 0.049\,\Sigma x1\,\Delta x1 - 1.477 \end{aligned} \tag{10}$$

$$r^2 = 0.959 \qquad s = 0.177 \qquad F = 70 \qquad n = 21$$

Table 1 lists the compounds along side the observed, calculated and residual pD. These results clearly indicate the quality of the analysis. The largest residual occurs for compound **11** which contains one of the two cyclic substituents. No residuals are greater than two standard deviations.

This equation represents a parabolic surface with an extremum point, either a maximum or minimum in the activity and the analysis of the response surface can now be performed. Of course, this analysis could also be performed on the equations obtained by directly using the *chi* indexes. We were analyzing the equation based on $\Sigma x1$ and $\Delta x1$, because we wished to emphasize the usefulness of these two variables. The response surface equation is simplified by moving the origin to the coordinates of the extremum point. That point corresponds to the point at which the first derivatives are zero. The expressions for $\Sigma x1$ and $\Delta x1$ at the extremum point are as follows:

$$\Sigma x1_{ext} = (2a_2b_1 - ca_2)/(c^2 - 4b_1b_2) \tag{11}$$

$$\Delta x1_{ext} = (2a_1b_2 - ca_1)/(c^2 - 4b_1b_2) \tag{12}$$

The symbols a, b, and c refer to Eq. (3).

In this prepsent study, $\Sigma x1_{ext} = 12.757$, and $\Delta x1_{ext} = -1.491$, In the new coordinate system, with the origin shifted, $q_1 = \Sigma x1 - \Sigma x1_{ext}$, and $q_2 = \Delta x1 - \Delta x1_{ext}$. When this substitution is made, the linear terms disappear and the remaining equation has only three terms. If one uses ordinary regression on this three-variable equation, essentially the same statistics result, but with a higher F value, due to a reduced number of variables:

$$\mathrm{p}D = -0.0497q_1^2 - 0.187q_2^2 - 0.0493q_1q_2 + 6.090 \tag{13}$$

This QSAR analysis can be discussed in terms of the variables, $\Sigma x1$ and $\Delta x1$. $\Sigma x1$ is much larger than $\Delta x1$ as shown in Table 1, as a direct result of its definition. $\Sigma x1$ also spans a wider range of values, from 2.992 to 15.058, compared to 1.099 to 3.132 for $\Delta x1$. $\Delta x1$ does not have as large a variance as $\Sigma x1$; many compounds lie in a narrow range, because they possess only two heteroatoms (in addition to the phosphate group), either nitrogen or oxygen (See Fig. 1). Two compounds are significantly different. Compound **18** is the only one with an aromatic ring and compound **21** is the only phosphinic acid. It can easily be seen that the electronic structure in this data set does not vary much. This electronic variation is directly represented by the variation in $\Delta x1$.

On the other hand, there is significant variation in molecular size in this data set. This variation, encoded by $\Sigma x1$, is largely responsible for the toxicity variation, with the exception of compounds **18** and **21**, which possess the lowest values.

In a plot of the data on the $\Sigma x1$, $\Delta x1$ axes (Fig. 1), most compounds fell within a band, but compounds **18** and **21** were exceptions. Furthermore, it was noted that compounds with an alkyl substituent, directly on the phosphorus, were situated in the region around $\Delta x1 = -0.8$ to -0.9, amides around -0.4, and esters near $\Delta x1 = -0.1$. Using this $\Sigma x1$, $\Delta x1$ form of analysis, it could clearly be seen that size variation was very important in this data set. Some useful clustering of compounds can be achieved with *chi* index plots based on $\Sigma x1$ and $\Delta x1$, or ${}^1\chi$ and ${}^1\chi^v$.

Based on the regression model, the maximum toxicity and the values for the variables at the maximum point may be obtained from the condition, $q_1 = q_2 = 0$. This extremum point is a maximum in this data set, with $pD_{\max} = 6.09$, $\Sigma x1_{\max} = 12.76$, and $\Delta x1_{\max} = -1.49$. Based on the definition of $\Sigma x1$ and $\Delta x1$, the *chi* index values at the maximum are, ${}^1\chi_{\max} = 5.63$, and ${}^1\chi^{\mathrm{v}}_{\max} = 7.13$. No compound in the data set corresponds to these values although compound **2** is the closest. In fact, the maximum point does not lie within the actual data set but just outside as shown in Fig. 1. If one wishes to use this data set as an aid in designing compounds of lower toxicity, then one should design compounds with *chi* values, which are far from these values at the maximum. Based on the significance of $\Sigma x1$ and $Dx1$, one would look for a greater size, resulting from increased alkyl portions, or for smaller molecules with a greater electronic contribution from heteroatoms. Also, introducing more size, with a much greater contribution from pi and lone pair electrons, would result in a molecule being further away from the maximum point on the response surface. Compound **18** illustrates this point.

3.3.4.2 Bioconcentration of Chlorinated Phenyls and Biphenyls

Another area of significant interest, besides the toxicity of chemicals, is the ability of chemical substances found in the environment to accumulate in organisms. This concern is especially important for aquatic organisms. Sabljić and Protić [12] have published an analysis of a set of chlorinated organic molecules, using molecular connectivity indexes. The measured bioconcentration values were expressed on a molar basis and then converted to the negative logarithm $pBCF$.

In their analysis, Sabljić and Protić used a simple quadratic relation between bioaccumulation and structure, which was based on the second order valence *chi* index. They included 17 compounds in their test set an then predicted four other compounds. Their *chi* equation gave predictions which was in very good agreement with experimental values.

We expanded the investigation of this data set to a full two-variable parabolic relation, so that a response surface could be generated and analyzed, and we included all 21 compounds. We examined the zero, first, and second order valence *chi* indexes. The *chi* indexes were computed using Molconn-X [9]. We found that a very good full quadratic relation could be built using both the simple and valence zero order *chi* indexes, as follows:

$$\log BCF = -0.464\,{}^0\chi - 0.560({}^0\chi)^2 + 1.872\,{}^0\chi^{\mathrm{v}} - 0.788({}^0\chi^{\mathrm{v}})^2 + 1.327\,{}^0\chi\,{}^0\chi^{\mathrm{v}} - 3.833$$

$$r^2 = 0.972 \qquad s = 0.21 \qquad F = 110 \qquad n = 22 \qquad (14)$$

When the sum and difference transformations were performed, similar to Eqs. (8) and (9), the following relation was obtained:

$$\log BCF = 0.704\,\Sigma x0 - 0.015\,(\Sigma x0)^2 - 1.168\,\Delta x0 - 0.678\,(\Delta x0)^2 + 0.094\,\Sigma x0\,\Delta x0 - 3.833 \qquad (15)$$

$$r^2 = 0.972 \qquad s = 0.21 \qquad F = 110 \qquad n = 22$$

Table 2. The observed, calculated and residual biconcentration factors for the data set of chlorinated organic compounds

Obs.	Compound Name	Plot Symbol[a]	$^0\chi^b$	$^0\chi^{v\,c}$	log *BCF*[d]	*Calc*[e]	*Res*[f]	*Pres*[g]
1	Chlorobenzene	**a**	5.113	4.521	1.08	1.16	−0.08	−0.19
2	1,4-Dichlorobenzene	**b**	5.983	5.577	2.33	2.16	0.17	0.24
3	1,2,3-Trichlorobenzene	**c**	6.853	6.634	2.69	2.92	−0.23	−0.29
4	1,2,4,5-Tetrachlorobenzene	**d**	7.724	7.690	3.65	3.47	0.18	0.23
5	Pentachlorobenzene	**e**	8.594	8.747	3.70	3.78	−0.08	−0.11
6	Hexachlorobenzene	**f**	9.464	9.803	3.93	3.87	0.06	0.09
7	Biphenyl	**g**	8.226	6.773	2.53	2.27	0.26	0.35
8	4-Chlorobiphenyl	**h**	9.096	7.830	2.77	3.23	−0.46	−0.54
9	2,4,4′-Trichlorobiphenyl	**i**	10.836	9.943	4.69	4.48	0.21	0.25
10	2,2′,5-Trichlorobiphenyl	**j**	10.836	9.943	4.69	4.48	0.21	0.25
11	2,2′,4,4′-Tetrachlorobiphenyl	**k**	11.707	11.000	4.86	4.76	0.10	0.11
12	2,2′,5,5′-Tetrachlorobiphenyl	**l**	11.707	11.000	4.86	4.76	0.10	0.11
13	2,2′,4,5,5′-Pentachlorobiphenyl	**m**	12.577	12.056	4.66	4.82	−0.16	−0.19
14	2,2′,4,4′,5,5′-Hexachlorobiphenyl	**n**	13.447	13.113	4.66	4.66	0.00	0.00
15	Diphenyloxide	**o**	8.933	7.182	2.29	2.14	0.15	0.27
16	4-Chlorodiphenyloxide	**p**	9.803	8.238	2.87	3.14	−0.27	−0.36
17	Endrin	**q**	13.533	13.883	3.61	3.63	−0.02	−0.03
18	Methoxychlor	**r**	15.458	13.914	4.79	4.75	0.04	0.54
19	Heptachlor	**s**	11.878	12.267	3.76	3.90	−0.14	−0.19
20	DDT	**t**	14.044	13.366	4.79	4.84	−0.05	−0.07
21	DDD	**u**	13.121	12.309	4.81	4.92	−0.11	−0.12
22	Dieldrin	**v**	13.533	13.883	3.76	3.63	0.13	0.20

[a] Symbol used to identify compound in the plot in Fig. 2.
[b] Zero order molecular connectivity index.
[c] Zero order valence molecular connectivity index.
[d] logarithm of the experimental bioconcentration factor [12].
[e] p*BCF* calculated from Eq. (14) or Eq. (15).
[f] p*BCF* − *Calc*.
[g] The predicted residual, obtained by deleting the observation and then predicting its value from the remaining $n - 1$ observations.

The observed, calculated and residual log *BCF* values are given in Table 2 along side the compound names. The QSAR model is very good; there is only one observation having a residual greater than two standard deviations. There are no trends in the plots of residuals versus the observed log *BCF* values. Analysis of the response surface revealed that the extremum point is a maximum and the coordinates of the maximum are $\Sigma x0_{\text{ext}} = 26.530$, and $\Delta x0_{\text{ext}} = 0.971$. In the original *chi* index the values are $^0\chi_{\text{ext}} = 13.750$, and $^0\chi^{\text{v}}_{\text{ext}} = 12.780$. A contour plot of the response surface is given in Fig. 2 along with the positions of the 22 compounds in the data set.

In the contour plot, the families of compounds are clearly visible. The biphenyls are arrayed along a straight line, as are the phenyls, which are parallel to the line of biphenyls. Furthermore, the two diphenyloxides are on an adjacent parallel line. These lines follow a direction, in which size is increasing due to the addition of chlorine atoms, and the line also follows along a direction of increasing number of lone pair electrons, due to the increase of chlorine atoms. As number of chlorine atoms increases, $\Delta x0$, becomes more negative, because $^0\chi^{\text{v}}$ becomes larger relative to $^0\chi$: $^0\chi^v > {}^0\chi$ for atoms beyond fluorine in the periodic table. The four compounds which do not clearly fall into these three structure classes, fall in to a somewhat different part of the plot, depending upon their size and number of lone pair electrons. This display of molecular structures gives a clear picture of the structure attributes which influence the bioconcentration property.

3.3.5 Conclusions

Representation of biological data on a parabolic surface is possible, when the data set reveals quadratic non-linearity. In this case, it is useful to analyze the data, in order to determine the extremum point. The analysis leads to the determination of the extremum point in a straightforward manner, determining whether the extremum is a maximum or minimum.

When the data are displayed as a contour plot, as in Fig. 1 and 2, the structure information is more clearly visible. Further, it is easier to discern just how the structure descriptors relate to the biological activity, so that molecular design information may be obtained. It is not usually necessary to actually carry out the transformations described above, in order to obtain the equation in the simplified form. The typical computer output provides the information necessary for analysis. If a particular computer program does not include a response surface feature, the same information may be obtained by ordinary multiple linear regression.

It should be pointed out, that the number of observations in these two data sets, 21 and 22, is somewhat small for the five variables in the QSAR models, Eqs. (6) and (14). As a general rule, it is desirable to have at least five compounds for each variable. The QSAR equations contain five variables, including the linear, squared and cross terms. There is a somewhat higher possibility for chance correlation in these cases. If more data is available, these models could be improved by the addition of such data. We have included the predicted residual for both cases, that is, the residuals predicted for a given observation, when that observation is deleted from the data set and then predicted from the remaining observations. As can be seen in Tables 1 and 2, the predicted residuals (pres) are not poorly behaved, suggesting that these models may still have predictive power, despite the less than ideal number of observations.

In this presentation, we have included an additional strategy for structure analysis. The use of the *chi* indexes, especially in the sum/difference transformation, reveals important structure features. The two examples described here make use of the zero and the first order simple and valence *chi* indexes. Further, an enhancement in structure information is developed with particular linear combinations of these simple and valence indices. The sum variables ($\Sigma x0$, $\Sigma x1$) are highly, related to molecular size. The difference variables ($\Delta x0$, $\Delta x1$) are strongly related to molecular electronic structure, especially the role of the pi and lone pair electrons. We have described these difference indices also as delta chi indices [3, 5, 13, 14]. Higher order indexes may also be used in this same manner. This form of structure analysis provides a basis for further design of molecules, to improve activity or diminish undesirable effects, such as toxicity or bioconcentration.

References

[1] Van Valkenburg, W., ed., *Biological Correlations – The Hansch Approach* (Advances in Chemistry Series **114**), American Chemical Society, 1972

[2] Kier, L. B., and Hall, L. H., *Molecular Connectivity in Structure-Activity Analysis*, John Wiley & Sons, Research Studies Press Chichester, England, 1986

[3] Hall, L. H., *Computational Aspects of Molecular Connectivity and its Role in Structure-Property Modeling*. In: *Computational Chemical Graph Theory*, Rouvray, D. H., ed., Nova Press, New York (1990) p. 202–233

[4] Kier, L. B., Indexes of Molecular Shape from Chemical Graphs. In: *Computational Chemical Graph Theory*, Rouvray, D. H., ed., Nova Press, New York, 1990, p. 152–174

[5] Hall, L. H., and Kier, L. B., *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Relations*. In: *Reviews of Computational Chemistry*, Boyd, D., and Lipkowitz, K., eds., VCH Publishers (1991) p. 367–422

[6] Kier, L. B., and Hall, L. H., *An Atom-Centered Index for Drug QSAR Models*. In: *Advances in Drug Research*, Vol. **22**, Testa, B., ed., Academic Press, 1992, p. 2–38

[7] Cavallito, C. J., ed., *Structure-Activity Relationships*, Vol. **I**, Pergamon Press, Oxford, 1973

[8] SAS Institute, Cary, NC

[9] Molconn-X Version 2.0 from Hall Associates Consulting, Quincy, MA, 02170 USA

[10] (a) Mager, P. O., in The MASCA Model of Pharmacochemistry. In: *Drug Design*, Ariens, E. J., ed., Academic Press, New York, Vol. IX, 187–236; Vol. X, 343–401, 1980.
(b) Mager, P. O., *Toxicol. Lett.*, **ii**, 67 (1982)

[11] Hall, L. H., and Kier, L. B., *J. Molec. Struct. (Theochem)* **134**, 309–316 (1986)

[12] Sabljić, A., and Protić, M., *Chem.-Biol. Interactions* **42**, 301–310 (1982)

[13] Kier, L. B., and Hall, L. H., *Pharm. Res.* **6**, 497–500 (1989)

[14] Kier, L. B., and Hall, L. H., *Quant. Struct.-Act. Relat.* **10**, 134–140 (1991)

[15] Van de Waterbeemd, H., ed., *Advanced Computer-Assisted Techniques in Drug Discovery*, Methods and Principles in Medicinal Chemistry, Vol. **3**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1995.

3.4 Disjoint Principal Properties of Organic Substituents

Han van de Waterbeemd, Gabriele Costantino, Sergio Clementi, Gabriele Cruciani and Roberta Valigi

Abbreviations

CCD	Central composite design
COMFA	Comparative molecular field analysis
DOD	D-optimal design
DPPs	Disjoint principal properties
FD	Factorial design
FFD	Fractional factorial design
GOLPE	Generating optimal PLS estimations
MLR	Multiple linear regression
PC	Principal component
PCA	Principal component analysis
PLS	Partial least squares
PPs	Principal properties

Symbols

$\mathrm{p}ED_{50}$	Dose at which 50% effect is observed
σ_p	Hammett constant for para substitution
V_{T}	Steric constant related to molar volume
n	Number of compounds
r	Correlation coefficient
r^2	Explained variance
Q^2	Cross-validated correlation coefficient
w_1	First principal component from *w*-scales
l_1	First DPP from set of lipophilic descriptors
e_1	First DPP from set of electronic descriptors
s_1	First DPP from set of steric descriptors
h_1	First DPP from set of H-bonding descriptors

3.4.1. The Design of Molecular Diversity

3.4.1.1 Combinatorial Chemistry

The strategy of structure-based molecular design has been proven to be very successful in the pharmaceutical industry [1]. However, when structural information about the biological target is lacking, the strategy of lead finding involves the synthesis and testing of widely diverse compounds. In the field of peptide chemistry the generation of large peptide libraries has given new impetus to lead finding programs [2]. Increasingly, the interest is being focused on non-peptidic small molecules in combinatorial chemistry projects. Therefore, the definition of structural diversity is of considerable interest.

The systematic variation of substituents in a molecule has been the subject of various studies in the past. Besides synthetic feasibility and economic considerations, substituents are chosen on the basis of properties, such as polarity, size and H-bonding capacity. Although quantitation scales have been developed for such substituent properties, it is still not straightforward to select a representative subset of substituents that adequately covers the multidimensional parameter space. In this chapter, we will illustrate how principal properties (PPs) and disjoint principal properties (DPPs), derived from a large set of property descriptors, can be used to make rational choices. Using an experimental design approach, we have also introduced a set of twelve representative organic substituents.

3.4.1.2 Statistical Experimental Design

As described in the previous Chap. 3.1 and 3.2, statistical experimental designs should be used for complete coverage of the descriptor space by a minimum number of compounds. Such strategies include factorial designs (FD), fractional factorial designs (FFD), central composite designs (CCD), or D-optimal designs (DOD) [3,4]. The scope of such design plans is to limit the number of compounds to be synthesized and to guarantee statistically sound structure-property correlations. Recently, some of the authors of this paper have reported on the use of D-optimal design schemes in QSAR studies [4]. Briefly, D-optimal designs are more general than FD and are particularly appropriate for handling constraint problems, such as reducing of polysubstitution on an aromatic ring to only a few sites.

3.4.2 Substituent Descriptors

Several important compilations of aliphatic and aromatic substituent descriptors have been made [5 – 11]. These collections contain experimental and calculated substituent descriptor values. Van de Waterbeemd et al. [5 – 7, 11] at the University of Lausanne have compiled up to 121 variables for a set of 59 selected substituents,

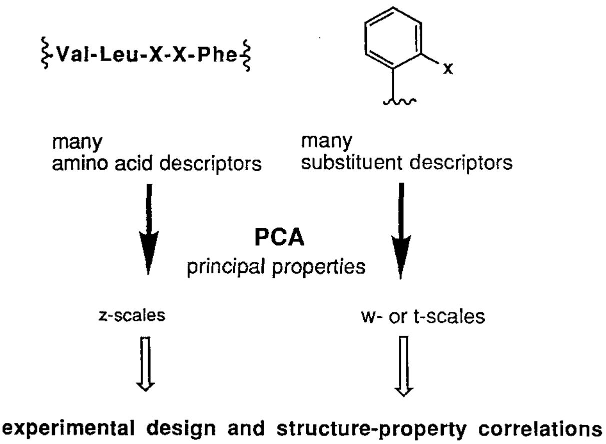


Figure 1. The use of principal properties in the design and in structure-property correlations of peptides and non-peptidic bioactive compounds.

[5–7, 11] while Clementi et al. [8, 9] at the University of Perugia evaluated a longer list of substituents, but for only 9 descriptors. Both approaches have their limitations. The larger set of substituents is described by only a few descriptors and, thus, only partly covers the descriptor space. The smaller set of substituents adequately covers the descriptor space, but offers limited choices for substitution.

3.4.2.1 Principal Properties (PPs)

The information content of large data tables can be reduced to less dimensions by pattern recognition techniques, such as principal component analysis (PCA) (see Chap. 4.1). The latent variables, obtained as statistical scores of a PCA, are called principal properties (PPs). These have been derived for amino acids, called *z*-scales

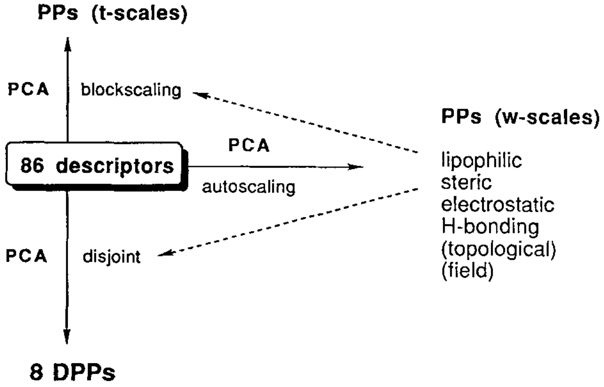


Figure 2. Strategies to derive principal properties (PPs) and disjoint principal properties (DPPs) from a set of 86 descriptors for 40 common organic substituents.

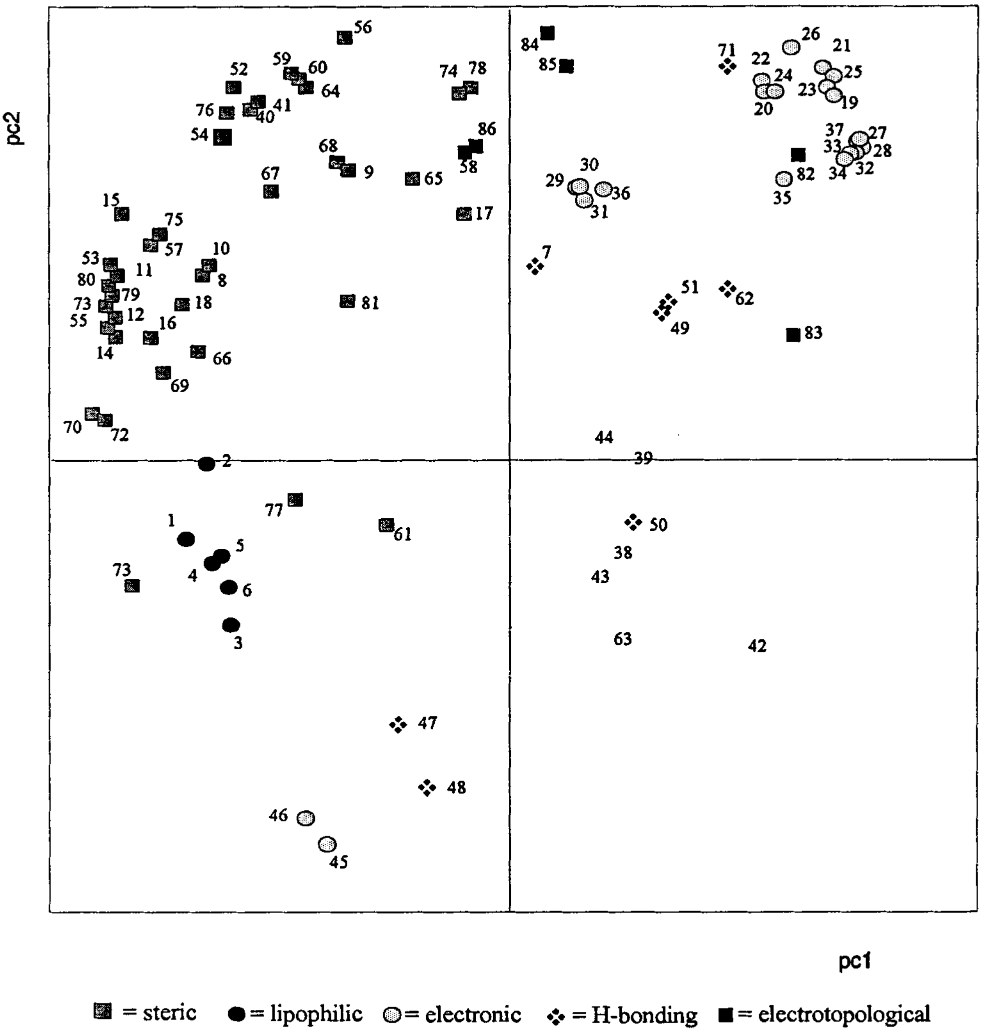


Figure 3. Loadings plot of the first two principal properties using the 86 descriptor 40 substituent data set. Descriptor numbering according to van de Waterbeemd et al. [6,7].

(see Chap. 3.2), as well as for organic substituents, called *w*-scales [5, 6] or *t*-scales [8, 9]. Such PPs can be used to describe the substituents or amino acids in structure-property correlation studies. Furthermore, they are of great interest in experimental design strategies (Fig. 1). They have the interesting property, in that they are orthogonal and can, thus, be used in multiple linear regression (MLR) studies. The w_1 and t_1 scales, already mentioned, describe mainly steric features, and w_2 and t_2 encode electronic aspects of the substituents. However, these PPs contain the generally recognized substituent properties (lipophilic, steric, electronic, and H-bonding) with contributions of each in each PP.

Recently we have reconsidered these data sets and performed a study, which is aimed at obtaining a unique set of PP scales for use in physical organic and medicinal chemistry [12]. The strategy described below is outlined in Fig. 2. These studies include the set of 86 descriptors, described by van de Waterbeemd et al. [7] with slight modifications to the Verloop parameters, $B2-B4$. The analyzes were restricted to a selected set of 40 substituents, for which most experimental data were available. The PC scores or PPs for the other 19 substituents were obtained by projection to the models of the training set. PCA on the autoscaled data yielded five principal components, describing 32%, 27%, 10%, 8% and 4% (cumulative 81%) of the data respectively. It might be questioned as to whether this is, indeed, sufficient for the calculation of significant and informative PPs. The loadings plot of the first two components (PPs explaining 59% of the variance) are reported in Fig. 3. In this plot five groups of descriptors were identified, which is partly consistent with our previous findings. These include the already mentioned lipophilic, steric, electronic/electrostatic and H-bonding properties of the substituents, as well as a diffuse group of topological descriptors, which are difficult to interpret. The previously mentioned *w*-scales were obtained in the same way [5, 6], which also include 3D CoMFA-derived field properties [7].

In order to eliminate the effect of the different size of each group, blockscaling of the descriptors can be applied, using an approach, which has previously been used to derive *t*-scales [9, 13]. Since PPs, derived either by autoscaling or blockscaling, include mixed contributions from all five main groups of variables, we followed a conceptually different approach by considering disjoint descriptor matrices.

3.4.2.2 Disjoint Principal Properties (DPPs)

Leaving out the topological descriptors, PCA was performed on four sets of descriptors. The first two significant components of each, called disjoint principal properties (DPPs), are reported in Table 1. In contrast to PPs, these DPPs are not orthogonal to each other, and are partly intercorrelated (see Table 2). In particular, the first lipophilic and H-bonding DPPs are collinear, which can be understood in the light of recent work on the information content in log *P* values [14, 15]. Orthogonality of DPPs is not a problem, as long as partial least squares (PLS) is used for data analysis instead of multiple linear regression (MLR).

3.4.2.3 Selection of Representative Substituents

In Table 3, the substituents are grouped according to their subspace in the DPP space. The sequence of signs, defined with respect to the mean value of each column in Table 1 for the training set, is steric (s), electronic (e), lipophilic (l) and H-bonding (h), and refers to the first property only. Owing to the collinearity among some of the scales, several subspaces are void and some subspaces are more populated than others. Thus, although the selected 59 substituents seem to be optimally chosen,

Table 1. Disjoint principal properties of common organic substituents

Explained variance %		62	13	77	18	94	2	69	10
	DPP	s_1	s_2	e_1	e_2	l_1	l_2	h_1	h_2
				a)					
1	–Br	7.308	–4.285	– 4.853	–3.677	1.159	0.776	0.659	–0.408
2	–Cl	5.344	–3.553	– 4.828	–3.699	0.810	0.724	0.625	–0.400
3	–F	2.935	–2.385	– 4.364	–5.045	–0.266	0.559	0.276	0.034
4	–I	9.859	–4.859	– 4.479	–3.315	1.810	0.818	0.696	–0.411
5	$-NO_2$	8.622	–5.375	– 9.929	–3.029	–1.318	0.546	3.804	–1.121
6	–H	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	–OH	3.840	–2.066	– 0.528	–5.170	–2.080	0.639	4.200	1.573
9	$-NH_2$	4.519	–2.110	1.976	–5.203	–2.623	0.178	4.919	1.566
10	$-SO_2NH_2$	2.765	–6.756	– 7.014	–1.380	–3.885	0.194	8.309	0.297
11	$-CF_3$	10.138	–7.693	– 6.288	–1.415	1.321	0.540	0.816	–1.099
13	$-SO_2CF_3$	16.950	–8.764	–11.121	–3.140	0.111	–0.536	6.377	–2.997
15	–CN	6.003	–2.385	– 8.476	–2.406	–1.621	0.498	3.045	–0.582
17	–NCS	10.212	–1.801	– 6.954	–4.035	1.612	0.744	2.354	–0.739
18	–CHO	6.764	–3.235	– 5.343	–1.117	–1.713	0.401	3.247	–0.673
19	–COOH	8.541	–3.910	– 5.081	–0.904	–1.254	0.556	5.041	1.343
20	$-CONH_2$	9.356	–3.653	– 5.014	–0.789	–3.443	0.311	7.042	1.084
22	$-CH_3$	5.217	–2.336	1.035	–0.686	1.125	0.064	0.013	–0.055
23	–OCH3	7.352	–1.812	– 1.103	–4.761	–0.911	1.004	3.084	–0.641
24	$-CH_2OH$	7.235	–2.300	– 0.203	–0.499	–2.267	–0.178	4.228	1.547
26	$-SO_2CH_3$	13.825	–6.616	– 8.994	–2.657	–3.159	–0.054	6.147	–1.910
28	$-NHCH_3$	7.477	–1.542	3.527	–5.081	–1.494	0.558	4.809	1.284
32	$-COCH_3$	10.033	–4.151	– 5.426	–0.628	–1.343	0.463	3.756	–0.826
33	$-COOCH_3$	12.236	–4.746	– 5.125	–0.959	–0.490	0.565	4.198	–0.953
39	$-C_2H_5$	8.660	–1.877	1.042	–0.515	1.970	–0.038	–0.154	–0.072
41	$-N(CH_3)_2$	10.572	–2.640	2.687	–6.346	–0.379	0.781	3.640	–0.905
42	$-C_3H_5$	11.076	–3.711	0.982	–1.215	2.164	0.022	–0.292	–0.002
43	$-COOC_2H_5$	14.572	–3.697	– 5.140	–1.023	0.496	0.632	4.124	–0.896
44	$-C_3H_7$	11.350	–1.278	0.845	–0.516	2.968	0.036	–0.237	–0.030
45	$-CH(CH_3)_2$	12.263	–4.399	1.000	–0.646	2.919	0.078	–0.209	–0.044
48	$-C_4H_9$	14.330	–0.254	1.064	–0.621	3.995	0.132	0.023	–0.034
49	$-C(CH_3)_3$	15.305	–6.877	1.354	–0.528	3.848	0.177	–0.106	0.031
50	$-OC_4H_9$	15.554	0.962	– 0.895	–5.077	2.335	1.043	3.660	–0.784
51	$-NHC_4H_9$	16.283	0.917	4.063	–0.492	1.348	0.908	5.231	1.206
52	$-N(C_2H_5)_2$	17.661	–4.007	3.734	–5.862	1.611	0.948	4.366	–1.035
53	$-C_5H_{11}$	16.652	0.907	1.166	–0.307	5.001	0.211	0.023	–0.034
54	$-C_6H_5$	16.333	–2.924	– 0.763	–1.067	3.626	0.545	0.349	–0.211
55	$-OC_6H_5$	17.396	–1.283	– 2.625	–4.764	1.649	1.260	2.893	–0.473
57	$-C_6H_{11}$	18.812	–1.009	1.201	–0.634	5.112	0.136	–0.095	0.215
58	$-COC_6H_5$	19.703	–2.742	– 5.321	–1.292	0.944	0.844	4.308	–0.999
59	$-CH_2CH_2Ph$	20.018	0.251	0.929	–0.994	5.034	0.249	0.004	–0.025
				b)					
8	–SH	6.428	–2.899	– 3.070	–2.390	0.351	0.607	1.745	1.460
12	$-OCF_3$	11.299	–3.682	– 5.358	–2.681	–0.069	1.321	3.897	–2.057
14	$-SCF_3$	13.233	–4.157	– 6.275	–1.935	1.248	0.873	2.454	–1.942
16	–SCN	10.138	–2.651	– 7.318	–3.069	0.086	0.767	3.239	–0.981
21	$-OCONH_2$	10.530	–2.121	– 3.093	–0.946	–2.526	0.248	6.912	1.026
25	$-NHCONH_2$	11.396	–2.970	– 0.313	–3.547	–2.987	0.494	7.314	1.250
27	$-SCH_3$	9.575	–2.388	– 1.694	–3.296	0.655	0.514	1.296	–0.728
29	$-C_2H$	6.879	–2.291	– 3.014	–1.177	0.478	0.283	0.420	0.236
30	$-CH_2CN$	9.580	–2.521	– 2.554	–2.594	–1.506	–0.096	3.590	–0.812
31	$-C_2H_3$	7.902	–2.388	– 0.809	–1.042	1.249	0.034	–0.118	0.005
34	$-OCOCH_3$	10.899	–1.694	– 4.561	–2.969	–1.715	0.855	5.473	–1.202
35	$-CH_2COOH$	11.514	–2.723	– 2.577	–1.066	–1.396	–0.191	5.615	1.031
36	$-OCH_2COOH$	12.550	–1.575	– 1.464	–0.770	–1.928	0.235	6.932	0.708
37	$-NHCOCH_3$	12.149	–2.341	– 2.794	–4.039	–2.421	0.523	6.450	0.827
38	$-NHCOOCH_3$	14.053	–2.832	– 0.895	–2.997	–0.838	–0.063	6.204	0.872
40	$-OC_2H_5$	10.100	–1.099	– 1.278	–4.198	0.209	0.812	3.302	–0.686
46	$-OC_3H_7$	12.927	–0.134	– 1.193	–4.448	1.267	0.932	3.454	–0.721
47	$-OCH(CH_3)_2$	13.233	–1.690	– 0.588	–6.218	0.995	0.788	3.454	–0.721
56	–NHPh	17.863	–1.159	1.082	–4.848	0.829	0.981	5.126	1.242

a) Calculated; (b) Projected. Substituent numbering as given by van de Waterbeemd et al. [6, 7]

Table 2. Correlation matrix (r) of disjoint principal properties

	s_1	s_2	e_1	e_2	l_1	l_2	h_1	h_2
s_1								
s_2	0.02							
e_1	0.12	**0.58**						
e_2	0.07	−0.08	−0.05					
l_1	**0.46**	0.27	0.38	0.20				
l_2	0.12	0.21	−0.04	**−0.54**	0.06			
h_1	0.14	−0.17	−0.27	−0.24	**−0.77**	0.08		
h_2	−0.23	0.37	**0.51**	0.12	−0.20	−0.19	0.14	

we were quite confident that they would cover a large part of the descriptor space. Using D-optimal design implemented in the DESDOP program [4], a selection of twelve representative substituents were made [12]. These included: H, Br, OH, CN, $COCH_3$, CH_3, SO_2CH_3, $N(CH_3)_2$, $C(CH_3)_3$, C_6H_5, COC_6H_5 and OC_3H_7. Alternatives may be selected from the various subgroups in Table 3.

3.4.3 An Example of DPPs in Design and Analysis

In a previous study on substituent descriptors the potential use of w-scales (PPs) was illustrated with a series of tricyclic neuroleptics [7] (Fig. 4). Using MLR and the classical Hansch approach the following equation between ataxia in mice and a steric (V_T) and an electronic descriptor (σ_p) was found:

$$\mathrm{p}ED_{50} = 0.533\sigma_p + 0.0366V_T - 0.00062V_T^2$$

$$n = 16\,; \quad r^2 = 0.76\,; \quad Q^2 = 0.42 \qquad (1)$$

Using the PPs, obtained as w-scales, and again MLR for the analysis, a significant equation was obtained, but which was difficult to interpret,

$$\mathrm{p}ED_{50} = -0.085\,(w_1)^2 + 0.134(w_2)^2 - 0.297(w_3)^2 + 1.343 \qquad (2)$$

$$n = 16\,; \quad r^2 = 0.79\,; \quad Q^2 = 0.59$$

where Q^2 is the cross-validated correlation coefficient (r_{cv}^2).

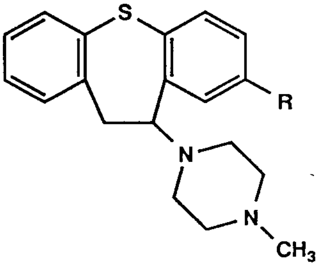


Figure 4. Design and data analysis of tricyclic neuroleptics using the DPP approach. R = Me, Et, i-Pr, t-Bu, H, OH, $-NH_2$, F, Cl, Br, $-CF_3$, $-OMe$, $-COMe$ or $-SMe$.

Table 3. Classification of substituents in the DPP space. The signs represent deviations from the mean values in Table 1 for s_1, e_1, l_1, h_1, respectively. The underligned substituents form a representative set of substituents with broad coverage of substituent properties

− − − −	− − − +	− − + −	− − + +	− + −	− + +	+ +	+ + +	+ − −
−F	$-NO_2$	**−Br**		**−H**	**−OH**	**$-CH_3$**		
−SH	**−CN**	−Cl			$-NH_2$	$-C_2H_5$		
$-C_2H$	−CHO	−I			$-OCH_3$	$-SCH_3$		
	−COOH	$-CF_3$			$-CH_2OH$	$-C_2H_3$		
	$-CONH_2$	−NCS			$-NHCH_3$			
	$-COCH_3$				**$-N(CH_3)_2$**			
	−SCN				$-OC_2H_5$			
	$-OCONH_2$							
	$-CH_2CN$							
	$-OCOCH_3$							

By reanalyzing these biological data, using a PLS model obtained by GOLPE [16] with the present DPPs, and after transformation to pseudo-regression coefficients [17], one obtains the following:

$$pED_{50} = -0.336\, l_2h_2 - 0.081\, h_2 + 0.032\, l_1s_2 - 0.030\, h_1h_2 - 0.004\, h_2s_1 + 1.021$$
$$n = 16\,; \qquad Q^2 = 0.75 \qquad (3)$$

It is remarkable that there are mainly cross terms in this equation, and furthermore, that H-bonding properties are quite important. This was not apparent from Eq. (1).

By using a D-optimal design approach it can be shown that a similar equation can be derived by using only nine substituents, namely, *t*-Bu, OH, NH_2, H, CH_3, F, SCH_3, Cl and $COCH_3$. This illustrates the way, in which the number of compounds in a series can be reduced to the strict minimum, thus, saving on resources for other subseries of compounds.

3.4.4 Conclusions

The present chapter describes a set of 8 new descriptors for 59 common substituents, which have been derived from 86 original experimental and calculated variable sets. These new descriptors are obtained from a disjoint analysis of four blocks of different aspects of substituent effects, namely lipophilic, electronic, steric and H-bonding effects. For each block, two new significant descriptors were derived. Applying a D-optimal design strategy, these four pairs of disjoint principal properties (DPPs) have been used to define a well-balanced set of substituents, covering the descriptors space as well as possible. These include H, Br, OH, CN, $COCH_3$, CH_3, SO_2CH_3, $N(CH_3)_2$, $C(CH_3)_3$, C_6H_5, COC_6H_5 and OC_3H_7.

It must be stressed that these DPPs are not orthogonal to each other and cannot be used in multiple linear regression (MLR) data modeling, whereas they can be

+ − − +	+ − + −	+ − + +	+ + − −	+ + − +	+ + + −	+ + + +
$-SO_2NH_2$	$-OC_6H_5$			$-NHCONH_2$	$-C_3H_5$	$-OC_4H_9$
$-SO_2CF_3$	**$-COC_6H_5$**			$-OCH_2COOH$	$-C_3H_7$	$-NHC_4H_9$
$-SO_2CH_3$	$-SCF_3$			$-NHCOOCH_3$	$-CH(CH_3)_2$	$-N(C_2H_5)$
$-COOCH_3$					$-C_4H_9$	**$-OC_3H_7$**
$-COOC_2H_5$					**$-C(CH_3)_3$**	$-OCH(CH_3)_2$
$-OCF_3$					$-C_5H_{11}$	$-NHPh$
$-CH_2COOH$					**$-C_6H_6$**	
$-NHCOCH_3$					$-C_6H_{11}$	
					$-CH_2CH_2Ph$	

used, without any problem, in PLS modeling. The DPP approach might not be considered satisfactory from a rigorous chemometric point of view. However, work is in progress to explore how the present DPPs will compare with some rotations of the *w*-scales, which cover the four main substituent effects.

References

[1] Gubernator, K., ed., *Structure-Derived Ligand Design. Methods and Principles in Medicinal Chemistry*, Vol. **4**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1995

[2] Moos, W. H., Green, G. R., and Pavia, M. R., *Ann. Rep. Med. Chem.* **28**, 315–324 (1993)

[3] Pleiss, M. A., and Unger, S. H., The Design of Test Series and the Significance of QSAR Relationships. In: *Quantitative Drug Design* (Comprehensive Medicinal Chemistry, Vol. **4**). Hansch, C., Sammes, P. G., and Taylor, J. B., eds., Pergamon Press, New York (1990) p. 561–587

[4] Baroni, M., Clementi, S., Cruciani, G., Kettaneh-Wold, N., and Wold, S., *Quant. Struct. Act. Relat.* **12,** 225–231 (1993)

[5] van de Waterbeemd, H., and Testa, B., *Adv. Drug Res.* **16**, 85–225 (1987)

[6] van de Waterbeemd, H., El Tayar, N., Carrupt, P. A., and Testa, B., *J. Comput.-Aided Mol. Des.* **3**, 11–132 (1989)

[7] van de Waterbeemd, H., Carrupt, P. A., El Tayar, N., Testa, B., and Kier, L. B., Multivariate Data Modeling of New Steric, Topological and CoMFA-Derived Substituent Parameters. In: *Trends in QSAR and Molecular Modeling 92*, Wermuth, C. G., ed., Escom, Leiden (1993) p. 69–75

[8] Alunni, S., Clementi, S., Edlund, U., Johnels, D., Hellberg, S., Sjöström. M., and Wold, S., *Acta Chem. Scand.* **37**, 47–53 (1983)

[9] Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G., and Ebert, C., *Quant. Struct.-Act. Relat.* **8**, 32–38 (1989)

[10] Hansch, C., and Leo, A. J., *Correlation Analysis in Chemistry and Biology*, Wiley, New York 1979

[11] van de Waterbeemd, H., Clementi, S., Costantino, G., Carrupt, P. A., and Testa, B., CoMFA-Derived Substituent Descriptors for Structure-Property Correlations. In: *3D QSAR in Drug Design: Theory, Methods and Applications*, Kubinyi, H., ed., Escom, Leiden (1993) p. 697–707

[12] Costantino, G., Clementi, S., Cruciani, G., Valigi, R., van de Waterbeemd, H., and Wold, S., *Quant Struct.-Act. Relat.*, submitted

[13] Alunni, S., Clement, S., Ebert, C., Linda, P., Musumarra, G., Sjöström, M., and Wold, S., *Chem. Soc. Perkin II*, 485–490 (1985)

[14] El Tayar, N., Testa, B., and Carrupt, P. A., *J. Phys. Chem.* **96**, 1455–1459 (1992)

[15] El Tayar, N., Tsai, R. S., Carrupt, P. A., and Testa, B., *J. Chem. Soc. Perkin Trans II*, 79–84 (1992)

[16] Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R., and Clementi, S., *Quant. Struct.-Act. Relat.* **12**, 9–20 (1993)

[17] Clementi, S., Cruciani, G., Curti, G., and Skagerberg, B., *J. Chemom.* **3**, 499–509 (1989)

Chemometric Methods in Molecular Design
edited by Han van de Waterbeemd
© VCH Verlagsgesellschaft mbH, 1995

4 Multivariate Data Analysis of Chemical and Biological Data

4.1 Principal Component and Factor Analysis

Rainer Franke and Andreas Gruska

Abbreviations

A	Anilines
B	Benzenes
BA	Benzoic acids
B. fr.	*Bacillus fragilis* A 22862
CNS	Central nervous system
DBH	Dopamine β-hydroxylase
E. cl.	*Enterobacter cloacae* A 9656
E. co.	*Escherichia coli* A 15119
E. fa.	*Enterococcus faecalis* A 9809
K. pn.	*Klebsellia pneumoniae* A 9664
M. mo.	*Morganella morganii* A 15153
NB	Nitrobenzenes
P	Phenols
PAA	Piperidinoacetanilides
P. ae.	*Pseudomonas aeruginosa* A 9843
PCRA	Principal component regression analysis
PCMM	Principal component analysis and multidimensional mapping
PhAA	Phenylacetic acids
PhOAA	Phenoxyacetic acids
P. mi.	*Proteus mirabilis* A 9900
PP	Principal property
S. au.	*Staphylococcus aureus* A 9537
S. ma.	*Serratia marcescens* A 20019
S. pn.	*Streptococcus pneumoniae* A 9585
TMIC	Two-dimensional mapping of intraclass correlation matrices

Symbols

A	Loading matrix (also factor pattern)
a_{kj}	Loading of the j-th variable in the k-th principal component or factor
α_k	Corresponding normalized eigenvector
$BC(DEF)$	PPs for organic compounds
$BRDa$	Decrease of diastolic (D) blood pressure in Wistar rats at dose a (after logarithmic transformation)

BRDb	Decrease of diastolic (D) blood pressure in Wistar rats at dose b (after logarithmic transformation)
BRSa	Decrease of systolic (S) blood pressure in Wistar rats at dose a (after logarithmic transformation)
BRSb	Decrease of systolic (S) blood pressure in Wistar rats at dose b (after logarithmic transformation)
CSA	Cavity surface area
$\boldsymbol{E}$	Matrix of the e_k
e_k	Basic effects operating in a system of biological tests
ε_h	Error of data reproduction
F	Fisher's F value
F	Inductive electronic substituent constant
$\boldsymbol{f}_k$	Scores of the k-th factor
h_j^2	Communality of the j-th variable
IND	Indicator function
$\log k'$	Hydrophobicity parameter from HPLC
λ_k	Eigenvalue of the k-th principal component
Max	Maximal potency = $\log (1/ID_{50})_{\max}$
MIC	Minimal inhibitory concentration
MR	Molar refractivity
μ	Dipole moment
P	Partition coefficient n-octanol/water
$\boldsymbol{P}_k$	Scores of the k-th principal component
pI_{50}(I)	Inhibition of DBH, Cu^2 excess
pI_{50}(II)	Inhibition of DBH, no Cu^{2+} excess
pK_a	pK_a value of fusaric acids
pK_b	pK_b value of fusaric acids
π	Hydrophobic substituent constant
ψ_h	ψ value according to Exner
$\boldsymbol{R}$	Correlation matrix
R	Resonance polar electronic substituent constant
r	Correlation coefficient
$\boldsymbol{R}^+$	Reduced correlation matrix
RE	Real error
s	Standard deviation
σ	Hammett constant
$t1$	Analgesic potency log = $(1/ID_{50})$ of fentanyl derivatives measured after $t = 1/32$ h
$t2$	Analgesic potency log = $(1/ID_{50})$ of fentanyl derivatives measured after $t = 1/16$ h
$t3$	Analgesic potency log = $(1/ID_{50})$ of fentanyl derivatives measured after $t = 1/8$ h
$t4$	Analgesic potency log = $(1/ID_{50})$ of fentanyl derivatives measured after $t = 1/4$ h
$t5$	Analgesic potency log = $(1/ID_{50})$ of fentanyl derivatives measured after $t = 1/2$ h

$t6$	Analgesic potency log $= (1/ID_{50})$ of fentanyl derivatives measured after $t = 1$ h
$t7$	Analgesic potency log $= (1/ID_{50})$ of fentanyl derivatives measured after $t = 2$ h
$t8$	Analgesic potency log $= (1/ID_{50})$ of fentanyl derivatives measured after $t = 4$ h
$t9$	Analgesic potency log $= (1/ID_{50})$ of fentanyl derivatives measured after $t = 6$ h
$t10$	Analgesic potency log $= (1/ID_{50})$ of fentanyl derivatives measured after $t = 8$ h
$t_1 - t_3$	PPs for amino acids
$\boldsymbol{W}$	Matrix of the w_k
w_k	Corresponding weight for the k-th test
$\boldsymbol{Y}$	Matrix of biological data
y_{ij}	Measurement for the i-th compound in the j-th test
$z_1 - z_3$	PPs for amino acids
$z_1 - z'_3$	PPs for amino acids

4.1.1 Introduction

If measurements are made on a number of objects, the results are usually arranged into a matrix, which is called a data table. The measurements are traditionally placed in the columns of this matrix and called variables, and the objects are associated with the rows. We shall follow this convention, although, from a purely mathematical point of view, there is no need for such a decision and, in addition, objects may be regarded as variables in the same right as measurements.

As long as data tables are two-dimensional (two rows or two columns), they can be quickly visualized using, for example, two-dimensional plots in a cartesian coordinate system. For multidimensional data tables, this is no longer possible, not only because of the abundance of entries, but also due to the complexity of data structure, as these entries depend on variables, objects, and the interactions between them. In order to understand such data in their entirety and to adequately deal with their mathematical properties, methods of multivariate statistics are required. Factor analysis methods, such as principal component analysis, factor analysis, canonical correlation and (multiple) correspondence analysis, which all have been applied to biological or chemical problems, (for reviews, see [1 – 13]) play an important role here. Their main objectives are to display multidimensional data in a space of lower dimensionality with a minimal loss of information, to extract the basic features "behind" the data, and to visualize data tables into some pictorial form, if possible, with the ultimate goal of interpretation and/or prediction. In this chapter, principal component and factor analysis and their application in the field of medicinal chemistry will be considered. Emphasis will be placed on practical aspects, which will be demonstrated with selected examples; additional basic mathematical treatments can be found, for example, in [14 – 20]. Some typical applications of principal component and factor analysis in medicinal chemistry are summarized in Table 1.

A number of regression equations will be presented in the following text. The terms in brackets after the regression coefficients are the 95% confidence intervals. n is the number of data points, r the correlation coefficient, s the standard deviation, and F is Fisher's F value.

4.1.2 Basic Principles

4.1.2.1 Principal Component Analysis

If for n chemical compounds ($i = 1, \ldots, n$) biological potencies are measured in m biological tests ($j = 1, \ldots, m$) the results can be arranged in a matrix which we shall call the biological data matrix. If the tests are put into the columns (variables) and the compounds into the rows (objects), then the matrix has the following form:

$$\boldsymbol{Y} = (y_{ij})_{n,m} \tag{1}$$

where y_{ij} is the biological potency of the i-th compound in the j-th test. In order to give all variables (which may be on quite different scales) the same importance, they are usually **standardized by autoscaling** according to,

$$y_j = (y_{j,\text{original}} - \bar{y}_{j,\text{original}})/s_{j,\text{original}} \tag{2}$$

where the index "original" refers to the original measurements which have a standard deviation of $s_{j,\text{original}}$, and a mean of $\bar{y}_{j,\text{original}}$.

Autoscaled variables have a mean of zero and unity variance. If we refer to measured values or measurements in this chapter, it is always tacitly assumed that the measurements have already been autoscaled according to Eq. (2).

If the biological tests considered are similar from a biological point of view, the following assumptions can be made:

1. The observed biological response in each test depends on a number of fundamental effects, termed here "basic effects", for example, transport through a membrane, binding to a biological target, etc.
2. These basic effects are present in all tests, but to varying degrees.
3. The biological response, y_{ij}, may be expressed as a linear combination of these effects.

If there are p such effects, e_k, ($k = 1, \ldots, p$), we then obtain:

$$y_{ij} = \sum_{k=1}^{p} e_{ik} w_{kj} \tag{3}$$

or, in vector and matrix notation,

$$\boldsymbol{y}_j = \sum_{k=1}^{p} \boldsymbol{e}_k w_{kj} \tag{4}$$

$$\boldsymbol{Y} = (e_{ik})_{n,p} \, (w_{kj})_{p,m} = \boldsymbol{E} \, \boldsymbol{W} \tag{5}$$

where the variable, y_j, represents the results from the j-th test, and $\boldsymbol{Y}$ represents the biological data matrix. The value of e_{ik} reflects how strongly the k-th basic effect is affected by the i-th compound. Thus, e_k is characteristic of the compounds and their properties. The weights, w_{kj}, are a measure of how important the corresponding e_k are in each biological test. They are characteristic of the biological tests, since it depends on the properties of the biological systems, which basic effects operate and to what extent they in operation.

Let us consider a simple example, where only two basic effects ($p = 2$) operate in three biological tests ($m = 3$). Then, Eq. (3) takes the following form,

$$\begin{aligned} y_{i1} &= e_{i1}w_{11} + e_{i2}w_{21} \\ y_{i2} &= e_{i1}w_{12} + e_{i2}w_{22} \\ y_{i3} &= e_{i1}w_{13} + e_{i2}w_{23} \end{aligned} \qquad (6)$$

and the matrices $\boldsymbol{W}$ and $\boldsymbol{E}$ become:

$$\boldsymbol{W} = \begin{array}{cc} & \textit{tests} \\ \begin{array}{c}\text{basic}\\ \text{effects}\end{array} & \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix} \end{array} \qquad (7)$$

$$\boldsymbol{E} = \begin{array}{cc} & \textit{basic effects} \\ \text{compounds} & \begin{bmatrix} e_{11} & e_{12} \\ \cdot & \cdot \\ e_{i1} & e_{i2} \\ \cdot & \cdot \\ e_{n1} & e_{n2} \end{bmatrix} \end{array} \qquad (8)$$

With the matrices $\boldsymbol{W}$ and $\boldsymbol{E}$, the features of the chemical compounds are completely separated from those of the biological tests, since $\boldsymbol{E}$ solely depends on the properties of the molecules, and $\boldsymbol{W}$ solely on the characteristics of the tests. Such a separation may provide a much deeper insight into the data structure and its underlying effects, than would be obtained from the global response data. In addition, the dimensionality of the data space is reduced. While the original data matrix has n rows and 3 columns, the matrices $\boldsymbol{E}$ and $\boldsymbol{W}$ have only two columns or two rows, respectively. Geometrically, the objects and test systems can now be represented in two-dimensional coordinate systems, spanned by the column vectors of $\boldsymbol{E}$ or the row vectors of $\boldsymbol{W}$, respectively, while, originally, the respective coordinate systems would have had 3 (columns of Y) or n (rows of Y) axes. If, in addition, the two effects could indeed be labeled "transport" and "target binding" by a suitable procedure, we would have gained considerably more information about the data and would have actually obtained information about the basic effects, which are operating and which underly the entire data, without having measured them.

The model underlying principal component analysis exactly corresponds to Eqns. (3) and (4): the elements of the (standardized) data matrix $Y$are described by a sum of product terms where, in each term, one factor is characteristic of the objects (compounds), and the other factor is characteristic of the variables (measurements, such as for example, from biological testing). In the terminology of principal component analysis this model becomes:

$$y_{ij} = \sum_{k=1}^{p} P_{ik}a_{kj} + \sum_{k=p+1}^{m} P_{ik}^{(0)}a_{kj}^{(0)} \tag{9}$$

The P_{ik} in Eqn. (9) are called principal components (PCs) and are sometimes also referred to as "scores". They correspond to the e_{ik} in Eqn. (3) and characterize the features of the compounds. Mathematically speaking, the $\boldsymbol{P}_k$ are orthogonal vectors, which are so determined that the original data matrix is reproduced. Analogously, a_{kj} in Eqn. (9) corresponds to the weight, w_{kj}, in Eq. (3) and characterizes the test systems. The "weight", a_{kj}, is a measure of the contribution of the k-th PC to the j-th variable, $\boldsymbol{y}_j$ (j-th column of $\boldsymbol{Y}$). As a consequence, a high value of $|a_{kj}|$ signifies a high importance of the k-th PC for the j-th variable. The j-th variable is then said to be highly "loaded" in the k-th PC, and the a_{kj}s are, therefore, also called "loadings".

Mathematically, the number of PCs, which can be extracted from a data matrix, is usually equal to m, the number of original variables, y_j. With this number of components, the data matrix can be exactly reproduced. This, however, is not a desired result, since it would not lead to a reduction of the dimensionality of the data space. What one wants to find is the minimum number, p, of components, such that, in the space which they span, the original variables can be represented without loss of relevant information. It is only then that the components will truly reflect the basic effects "behind" the data, in keeping with Eq. (3). These p components are represented by the first term in Eq. (9), while the components with the superscript "O" in the second term represent "irrelevant" or "residual" information comprising errors of measurement and, possibly, some error in the model. The objective of principal component analysis is to only evaluate the first sum in Eq. (9); the resulting components will then reproduce the data matrix, Y, within residual error. The following relation, which will not be derived here, serves as a starting point for evaluating the loadings and components:

$$\boldsymbol{R} = AA^{\mathrm{T}} \tag{10}$$

$\boldsymbol{R}$ denotes the correlation matrix of the data, and A is the loading matrix:

$$A = (a_{kj})_{p,m} \tag{11}$$

In order to solve Eq. (10), an additional condition is introduced, whereby the components are determined in sequence, in such a way that the first component accounts for the largest amount of correlation in $\boldsymbol{R}$, the second component for the next largest amount of correlation in $\boldsymbol{R}$, and so on. An eigenvalue problem, then, ensues according to the following,

$$\boldsymbol{R}\alpha_k = \lambda_k \alpha_k \tag{12}$$

where λ_k are eigenvalues, and α_k the corresponding orthonormalized eigenvectors. This equation is solved by diagonalizing $\boldsymbol{R}$ using standard procedures [21, 22]. Scores and loadings are obtained from the resulting eigenvectors and eigenvalues. The eigenvalues represent the variance contributions of the components. As a consequence of the procedure, the first component will have the largest, and the last component the smallest eigenvalue. Variances and correlations in the principal component model are related to the loadings and are defined as follows:

1. Variance of variable $\boldsymbol{y}_j$, extracted by the k-th principal component, is equal to a_{kj}^2
2. The variance of $\boldsymbol{y}_j$ extracted by p components is equal to

$$\sum_{k=1}^{p} a_{kj}^2$$

3. The variance contribution of the k-th principal component is equal to

$$\lambda_k = \sum_j a_{kj}^2$$

4. The total variance extracted by p components is equal to

$$\sum_{k=1}^{p} \sum_j a_{kj}^2$$

5. The correlation between variables y_q and y_r is given by the following:

$$r_{qr} = \sum_k a_{kq} a_{kr}$$

In order to find the minimum number of components, p, necessary for data reproduction within residual error, the components are added step by step to the model according to Eq. (12). After each step, the data matrix is reproduced, and the procedure is continued until only non-systematic "noise" remains. A criterion, which was sometimes used to recognize this salient feature was to accept only components with eigenvalues >1. However, this criterion seems to be much too narrow [23, 24] and may lead to the rejection of components, which are important for explaining the data. A better alternative is the Scree plot [25], where the residual percent variance (or simply the eigenvalues) are plotted against the number of components. The resulting curve should descend steeply and level off, if a limit corresponding to residual variation is approached. This point is used to deduce the number of components. Some other useful criteria are given in the following points.

1. So many components are extracted that the average error of the reproduced data becomes equal to the average experimental error. The average error, ε_h, of data reproduced using h principal components is:

$$\varepsilon_h = (1/nm)\left[\sum_j \sum_i \left(y_{ij} - \sum_{k=1}^{h} P_{ik}a_{kj}\right)\right] \tag{13}$$

   This criterion requires that the average experimental error of the variables is known and that there is no model error.

2. If the experimental error is not known, the so-called φ value, according to Exner [26], can be used:

$$\varphi_h = \sqrt{\frac{\sum_{i=1}^{m} \sum_{j=1}^{n} \left(y_{ij} - \sum_{k=1}^{h} a_{ik}x_{kj}\right)^2}{\sum_{i=1}^{m} \sum_{j=1}^{n} (y_{ij} - y_{..})^2} \cdot \frac{mn}{(mn) - h}} \tag{14}$$

   where $y_{..}$ is the total mean. For precise chemical or physical measurements, so many components should be added that $\varphi_h < 0.1$. In the case of less precise biological data, φ_h should be within the range of about 0.2 to 0.3.

3. The indicator function, IND, introduced by Malinowski [4, 27, 28], is also very useful:

$$IND = RE/(m - h)^2 \tag{15}$$

   RE is the so-called "real error" and is given by the following equation:

$$RE = \left\{[1/mn(m - h)]\left[\sum_{k=h+1} \lambda_k \sum_j \sum_i y_{ij}^2\right]\right\}^{1/2} \tag{16}$$

   The indicator function passes through a minimum for the correct number of components ($h = p$). The occurrence of a minimum also proves that the data analyzed can be correctly expressed by the model of principal component analysis.

4. Cross-validation is strongly recommended and is extensively used, especially by Wold et al. [29]. In this technique modified data sets are generated by eliminating small groups of objects, until each object has been left out once, and only once. For each modified set a principal component model is generated, which is then used to predict the value of the left out compounds. Then the quantity "*PRESS*" (prediction error sum of squares) is computed:

$$PRESS = \sum_i \sum_j (y_{ij} - y_{ij}\,(\text{predicted}))^2 \tag{17}$$

   The number of components is choosen so that *PRESS* is minimized.

If the minimum number of principal components, necessary to reproduce the data within experimental error, has been found, the principal component analysis, as such, is essentially complete. The practical gains so far is a reduced dimensionality of data space and that the number of relevant components reflects the true complexity of the data in terms of basic effects. Further evaluation and interpretation of the results can be achieved in several ways:

1. Matrix A will give information about the internal relatedness of the variables (e.g. biological tests). Variables, having the same information content, will give high values for $|a_{kj}|$ in the same component. A grouping of variables is, thus, obtained. If the first two components already account for a sufficiently high proportion of the data variance, this can then be visualized by a simple two-dimensional plot with the loadings of the first two components, $\boldsymbol{a}_1$ and $\boldsymbol{a}_2$, as axes (loading or factor weight plot). In such a plot each variable appears as a point, the points of related variables being close together (correlation coefficient r approaches 1) or symmetrical with respect to the center of the plot (r approaches -1)[1]. To obtain a numerically comprehensive picture of the relatedness of the variables, the loading matrix A can be rotated by multiplying with a rotation matrix. This corresponds to a rotation of the axes of the original plot, so that they pass through clusters of points, representing closely related variables and, thus, basic effects (see Fig. 1). Usually, an orthogonal rotation (VARIMAX rotation) is used, where all axes are rotated about the same angle. For mechanistic reasons, the basic effects need not necessarily be uncorrelated, i.e. in some cases, oblique rotations (different angles of rotation for different axes) can produce the simplest and most interpretable structure of the loadings.
2. The objects (compounds) can be plotted in a coordinate system with $\boldsymbol{P}_k$ as axes (score plot). If the two first principal components account for a sufficiently large proportion of the data variance, this plot is then two-dimensional. Objects may be clustered with respect to a certain property (which allows classification) or other characteristic patterns.
3. Identification of the "abstract" components with physically meaningful parameters will give an indication about the physical nature of the basic effects which underly the components and will eventually lead to multivariate RSARs. To this end, a special target rotation procedure introduced by Weiner and Malinowski (for details, see Ref. [4]) can be used. First, the parameters which are believed to be related to the components must be selected, either from a theoretical model, or from educated guesses (analogous to the Hansch analysis). The components are then rotated into vectors of these parameters (e.g. π, σ, etc.), also autoscaled, which are called test vectors; the elements of the test vectors are the values of the corresponding parameters for the objects (compounds) of the data matrix. If the target rotation has been successfully performed, the data can be reproduced by the test vector $\boldsymbol{t}_k$. Eq. (9) is then transformed into,

$$Y_{ij} = \sum_{k=1}^{p} t_{ik} a_{kj}^{R} + \varepsilon_{ij} \tag{18}$$

where t_{ik} is the value of the k-th vector for the i-th compound, ε_{ij} is the residual including the second term of Eq. (9) plus the error of the target rotation, and a_{kj}^{R} are the elements of the rotated loading matrix, A^{R}, obtained from:

$$A^{R} = (a_{kj}^{R})_{p,m} = AQ^{-1} \tag{19}$$

[1] Such correlations may be difficult to detect. It is, therefore, recommended to calculate a correlation matrix, prior to factor or principal component analysis, in order to obtain a clear picture of all pairwise correlations to start with.

The elements of the rotation matrix, $\boldsymbol{Q}$, are the results of least squares fits as in multiple regression analysis, in which the test vectors are dependent variables, and the components $\boldsymbol{P}_k$ are independent variables. Eq. (19) represents a system of "multivariate" equations (QSARs) as in multivariate regression analysis, with the original variables, y_j (autoscaled measurements in the biological tests) as dependent variables and the test vectors as independent variables.

In many cases, the primary objective is not a complete replacement of all components by test vectors, but rather the interpretation or identification of individual components with physico-chemical parameters. A much simpler approach can then be used. The components (dependent variable), are correlated with descriptor variables in a standard regression procedure. As a result, regression equations, for each of the components considered, are obtained.

4. Principal components may also be used within the framework of more complex methods such as PLS or SIMCA. Such methods will not be dealt with in this text; instead the reader is referred to Chaps. 4.3 and 4.4.

4.1.2.2 Factor Analysis

Factor analysis (FA) is very similar to principal component analysis (for reviews, please refer, for example, to Lewi [1], Rummel [14] and Harman [15]). The only but essential difference is that in FA only a proportion of the data variance is considered to be common to all variables. The remaining proportion is attributed to unique properties of one variable at a time. With this in mind, Eq. (9) may be rewitten as the following in order to obtain the model for FA.

$$y_{ij} = \sum_{k=1}^{p} f_{ik} a_{kj} + q_{ij} d_j \tag{20}$$

In this equation, the so-called common factors, f_k, which span the common factor space, replace the components in Eq. (9), and the a_{kj} are, once again, the loadings, representing the relations between factors and original variables. The $\boldsymbol{q}_j$ are called unique factors; their squared loadings, d_j^2, which are called uniquenesses, comprise the proportion of the data variance, which can be attributed to the unique variable properties not involved in the common correlation structure. Factor analysis is the method of choice in all cases, where such unique properties of variables occur, and this is to be expected, when a data matrix contains variables, which are quite different in nature and meaning and which are only loosely interrelated. In such matrices, error variances differing in size are to be expected, even for autoscaled data, which is another reason for applying factor analysis instead of principal component analysis.

The common features (the basic effects), sought after in the data space, are represented by the factors, f_k. Their extraction from the given data is based on the general assumption that only a certain proportion of the variability of, for example, a biological test can be explained by the basic effects also present in the other variables under consideration. As a consequence, only so much of the variance is considered that a minimum of common factors results.

The first objective of factor analysis is the evaluation of the loading matrix,

$$A = (a_{kj})_{p,m} \tag{21}$$

which is also called factor pattern in FA. This procedure is nearly the same as in principal component analysis, with the only difference being, that instead of the correlation matrix, $\boldsymbol{R}$, the so-called reduced correlation matrix, $\boldsymbol{R}^+$, is diagonalized. $\boldsymbol{R}^+$ differs from $\boldsymbol{R}$ in that the l's in the diagonal of the latter have been replaced by the communalities h_j^2. As a result, the reduced correlation matrix only contains that proportion of the data variance, which can be assigned to the common factor space. Representing this proportion of the data variance, the communality for the j-th variable is defined as,

$$h_j^2 = \sum_{k=1}^{p} a_{kj}^2 \tag{22}$$

and the variance of the j-th standardized variable then becomes,

$$s_j^2 = h_j^2 + d_j^2 = 1 \tag{23}$$

where d_j^2 represents the uniqueness. The uniqueness consists of the error variance and the so-called specificity, the latter representing mechanistically meaningful specific properties and systematic divergencies of the j-th variable.

Communalities must be estimated prior to the analysis. This can be accomplished in several ways which will not, however, be discussed here. During the computations, these estimates can be improved through iterative cycles. The number of relevant factors, p, is usually determined from the corresponding eigenvalues. Only those factors are considered significant, whose eigenvalues exceed a given borderline value and which, therefore, account for more than a given minimum variance. The borderline value is usually taken as representing an eigenvalue $\geq 5\%$ of the sum of all positive eigenvalues.

Prior to interpretation, the factors are usually rotated in such a way that the factor pattern simplifies as much as possible (Thurstone's simple structure). This structure is characterized by the property that a maximum number of variables lies close to the coordinate axes when presented in common factor space (axes = loadings), so that the largest possible number of factor loadings becomes zero (Fig. 1). Thus, in the presence of a simple structure, the variables are divided into mutually exclusive groups with, in an ideal case, non-zero loadings only in one factor. Whereas the original factors are always orthogonal due to the method of their extraction, the rotation can be orthogonal (VARIMAX rotation) or oblique. In some cases, a simple structure is achieved only by oblique methods, which are also justified by the fact that the "basic effects", underlying the data, must not necessarily be uncorrelated.

Results from factor analysis can be evaluated and interpreted in much the same way as outlined above for principal component analysis. The evaluation of the factor scores, f_{ik}, however, is not as straightforward as in principal component analysis. Since the rank of the factor matrix (common + unique factors) generally

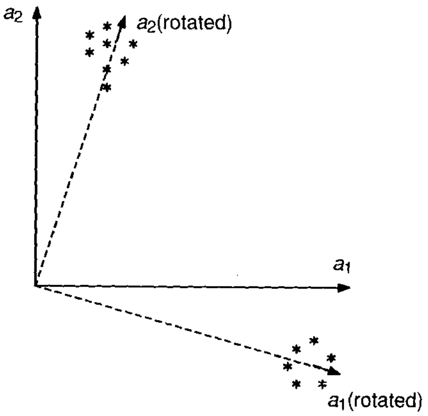


Figure 1. VARIMAX rotation of a two-factor solution towards Thurston's simple structure. Variables are presented as points in the space spanned by the loadings of the first and second factor (factor loading plot). In the unrotated coordinate systems, all variables have non-zero loadings in both factors. Rotation gives rise to the variables being divided into two clearly mutually exclusive groups, which have non-zero loadings in only one factor.

exceeds the number of variables, the f_{ik} must be estimated in an indirect way (see, e.g. Harman [15]). Although it is somewhat laborious to estimate f_{ik}, it is worthwhile, since the factors (scores) characterize the features of the objects of the data matrix and can be handled in much the same way as the principal components (provided that the proportion of the data variance, represented by the common factor space, is large enough).

The decision as to whether principal component analysis or factor analysis is to be used, depends on the nature of the data. If the variables are of a similar nature and reflect the same mechanisms of interactions between objects and systems, and if, in addition, the error variance of all variables is uniform and of comparable size, principal component analysis can be applied. Factor analysis is the method of choice, if the variables reflect very different processes with error variances of different sizes. From a chemical or biological point of view, principal component analysis can be used, if there is good reason to assume that the data can be described by Eq. (9) with no unique contributions of single variables; this requires that a corresponding theoretical model exists or that it is possible to derive one. If such a model does not exist and nothing is known about the behavior of variables, factor analysis probably is a good first move, which then operates as a model generator. If then the communalities of all variables are grater than approximately 0.8, principal component analysis can also be applied, which has the advantage that the scores can be exactly calculated [30].

4.1.3 Applications of Principal Component and Factor Analysis in Medicinal Chemistry

Some typical situations for the application of factor and principal component analysis in medicinal chemistry are summarized in Table 1. The examples, to be

Table 1. Selected applications of factorial methods in medicinal chemistry

Objects	Measurements	Applicatons
Compounds	Potency in a set of parallel biological tests with similar objects (e.g. bacteria, fungi)	Relations between tests: recognition of redundancies and of tests with high information content
Compounds	Potency in biological tests, believed to be mechanistically related	Relations between tests (especially: in vivo/in vitro): test of mechanistic hypotheses, relevance of in vitro tests, separation of pharmacodynamic and pharmacokinetic effects
Compounds	Potency in a biological screen	Relations between tests: redundancy, selectivity and mapping with respect to pharmacological profiles
Compounds	Potency at different times	Separation of different pharmacokinetic processes (e.g., distribution/elimination); Separation of pharmacokinetic and pharmacodynamic effects
Compounds	Various pharmacokinetic parameters	Separation of different pharmacokinetic processes
Compounds/ Substituents	Various physico-chemical parameters	Relations between parameters or between properties of compounds/substituents
Compounds/ Substituents	Various physico-chemical parameters	Design of optimal training series
Compounds/ Substituents	Various physico-chemical parameters	Mapping/classification with respect to biological properties
Compounds/ Substituents	Various physico-chemical parameters/measurements	Derivation of principal properties

discussed in the following subsections, have been selected to demonstrate the utility of these methods in practical applications to real data and problems in medicinal chemistry.

4.1.3.1 Data from Parallel Biological Tests

Frequently, a series of compounds is investigated in a battery of parallel tests with similar organisms, looking for the same type of biological activity. The main issues in this case concern redundancy and specificity. If redundancy is large, certain tests can be dispensed with to save experimental work, while tests with a high specific information content must be retained. In addition, if QSARs are to be derived, it is usually not necessary to consider all tests, if a well-defined data structure exists. One can then select representative key tests, or perform QSAR analyses directly with principal components or factors. A typical situation would be the screening of potential antibacterial agents against several bacterial strains; other examples are the screening for pesticides or for antitumor compounds. Antibacterial data have been successfully treated by principal component or factor analysis, usually showing high redundancy and a well-defined data structure [30–40]. Principal component analyses of herbicidal piperidinoacetanilides [42] and benzonitriles [43a] showed in both cases a distinct grouping of tests in accordance with the biological properties of test objects. A principal component analysis of allylamine antimycotics, which were screened against seven strains of fungi, revealed three significant components, which could then be submitted to a subsequent QSAR analysis [43b]. Examples for the application of principal component analysis to antitumor tests can be found in [44] and [45]. The use of simple model organisms to evaluate the toxic, carcinogenic, and mutagenic potential of chemical compounds has become a very important issue, in order to be able to cope with the ever increasing number of chemicals in the human environment. The resulting batteries of parallel tests can also be effectively investigated by principal component or factor analysis. Systematic multivariate analyses of mutagenicity short term tests have been performed by Benigni and coworkers [46–49]. In one of the analyses, it was found, for example, that the results from 20 tests (42 compounds tested) could be described by 6 factors. The first factor, accounting for 58% of the data variance, was interpreted to represent "intrinsic genotoxicity" [47]. Nendza and Seydel [50, 51] investigated the toxic effects of phenols and anilines, measured in 11 in vitro tests (bacteria, yeast, protoplasts and algae), by means of principal component analysis. The first principal component extracted almost 80% of data variance (indicating high redundancy) and correlated with lipophilicity. A similar result was obtained by factor analyzing cytotoxicity data (9 cell lines) and in vivo toxicities (LD_{50} in rats and mice) of a structurally heterogeneous set of 19 compounds: two factors accounted for 90% of the data variance (Gruska, A., Halle, W., and Franke, R., unpublished results). In a similar investigation for 9 endpoints, Eriksson et al. [52] obtained 3 significant principal components from principal component analysis. Redundancy was also found for the acute toxicity of 267 chemicals on six species of biota [53]. Two principal components were obtained from a series of 30 structurally diverse compounds screened in 4 tests, which were related to the induction of anesthesia and spindle disturbances [54]; the first component showed a correlation with log *P*.

Frequently, a certain pharmacological effect is investigated in different models, with the assumption that each model reflects the desired potency under more or less specific conditions. Typical cases are receptor binding assays in vitro (receptors, enzymes) in combination with pharmacological in vivo tests. The fundamental problem in this case would be to prove the internal relationship between the models (whether they have the same mechanism of action), and/or to separate pharmacodynamic and pharmacokinetic factors. A simple example of this type was provided by a principal component analysis of the inhibitory potency of a series of nine 4-hydroxyquinoline-3-carboxylic acids against respiration of Ehrlich ascites tumor cell suspensions and three respiratory enzymes in vitro (muscle lactate dehydrogenase (M4-LDH), pig heart cytoplasmic malate dehydrogenase (s-MDH) and pig heart mitochondrial malate dehydrogenase (m-MDH) [55]. Two components accounted for 91% of the data variance. The first principal component was loaded for the three enzymes, which had loadings close to zero in the second one, while the ascites test was highly loaded in the second, but not in the first component. It, thus, follows that the enzyme tests are closely interrelated and the ascites test is completely separated, so that the enzyme inhibitory potency is not connected with the inhibition of Ehrlich ascites respiration. As the first component obviously reflects enzyme inhibition in vitro, it may be regarded as an average expression for the inhibition of all three enzymes. Correlating P_1 (the first principal component) with molecule parameters yielded the following relationships (the SO_3^- derivative was not included because of difficulties in calculating μ):

$$P_1 = 0.34(\pm 0.7)\mu + 1.38(\pm 0.78)B_4 - 4.86(\pm 2.22) \qquad (24)$$
$$n = 8 \qquad r = 0.902 \qquad s = 0.520$$

In this equation, μ is the dipole moment, and B_4 is Verloop's width parameter. It, thus, follows that enzyme inhibition depends on both steric and electronic effects, with inhibitory potency increasing with increasing dipole moment or electron-donating power of the substituents as well as with increasing substituent width. In the case of the second component for the Ehrlich ascites test, the following relation was obtained (π = hydrophobic substituent constant calculated from apparent partition coefficients):

$$P_2 = 0.79(\pm 0.24)\pi + 0.51(\pm 0.33) \qquad (25)$$
$$n = 8 \qquad r = 0.931 \qquad s = 0.415$$

In this case, potency is dominated by hydrophobicity; no relationships with significant contributions by steric or electronic parameters could be found. As a result of principal component analysis, a clear picture of the data structure, as well as of the physical nature of the two "basic effects" as reflected by the two components (QSARs), were obtained. A further example of this type will be discussed in somewhat more detail on page 130.

Sometimes compounds are also investigated in parallel tests with quite different biological actions. This may happen, for example, with a general screening in a pharmaceutical company, or if a synthetic chemist wants to obtain as much biological information about new compounds as possible. In addition to redundancy and the general interrelatedness (grouping) of tests, aspects which may be of importance here, are selectivity (separation of desired effect(s) from undesired toxic or side effects) and the evaluation of pharmacological profiles. For such data, factor analysis will often be the method of choice. A typical example is provided by the work of Weiner and Weiner [56], who introduced principal component analysis into QSAR work and investigated the results of a series of diphenylaminopropanols, screened in 11 different pharmacological tests. As these tests are related to quite different mechanisms of action high uniquenesses are to be expected. It was, thus, not surprising that principal component analysis yielded as many as 8 components, so that the dimensionality reduction achieved was not very impressive. If, however, the same data was submitted to factor analysis, only 3 relevant factors were obtained, accounting for 80% of the data variance [30]. A clear grouping of tests resulted, and factor scores could be related, to physico-chemical parameters. The effect of *ortho*-, *meta*- and *para*-substituted phenyls (12 substituents in each set) on 24 biological activities, such as antibacterial, antitumor, enzyme inhibition, and others was investigated by Codarin et al. [41]. Considering the first two components the authors concluded "that the major behavior of the activity data, expressed by the first two components, is related to the known descriptors π, σ, and E_s". Results for the induction of various enzymes by polychlorinated biphenyls have been presented by Franke et al. [57].

A central issue in drug research is to decide whether, within a given series of compounds, toxic effects can be minimized, while maintaining a desired potency. How factorial methods can aid in such cases can be demonstrated with a simple example, concerning a series of antiinflammatory phenylglycin esters with branched and unbranched alkyl groups in the alcoholic part. The compounds were synthesized and tested for antiinflammatory potency (against carrageenin and dextran edema) as well as for anticonvulsant potency (antagonism against histamine, $BaCl_2$ and acetylcholine) and for toxicity by Schulz and coworkers [58–60]. Principal component analysis afforded 2 components [30], accounting for 93% of the data variance. As, in a loading plot, toxicity is not separated from the other tests, it was to be concluded that with the type of structural variation present in the series, a pronounced decrease in toxicity, while maintaining the desired effects, is not possible. The first component could be related to π probably reflecting the transport to the site of action [30]:

$$P_1 = -0.79(\pm 0.11)\pi^2 + 6.87(\pm 0.88)\pi - 13.95(\pm 1.66) \qquad (26)$$

$$n = 13 \qquad r = 0.984 \qquad s = 0.193$$

Another interesting aspect of such an analysis is to derive pharmacological profiles by mapping substances according to measurements in relevant biological tests. A number of representative examples, mostly concerning CNS active drugs, have been published by Lewi [1–3].

Drug interactions in model systems have also been investigated by principal component analysis. Seydel et al. [61] extracted one component from data of benzylamines interacting with phospholipids, which showed a non-linear dependence on π; if more tests were incorporated, then a second component was obtained. The interaction of monoamino oxidase (MAO) inhibitors with amino acids, studied by charge-transfer chromatography, led to 3 principal components and the conclusion that MAO inhibitory drugs interact only with dicarboxylic acids via electrostatic forces [62].

Example: Antibacterial Naphthyridines in Different Bacterial Strains

Data on the antibacterial potency of naphthyridines (structure see Fig. 2) taken from the literature [63] were submitted to factor analysis (Franke et al. [63a]) using the statistical program package STATGRAPHICS [64]. The data presented in Table 2 yielded three significant factors accounting for 82.7%, 7.9%, and 4.9% of the variance in common factor space, respectively. As the first two factors together already represent 90.6% of the variance, a weight plot of these factors can provide enough information about the relatedness of tests. Such a plot is shown in Fig. 3. Although an ideal simple structure was not achieved, even after VARIMAX

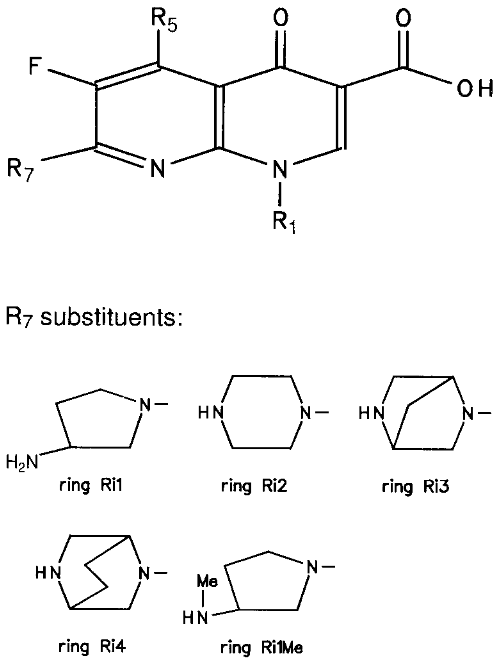


Figure 2. Antibacterial fluoroquinolones.

rotation of the axes, a distinct clustering of points can still be seen. The largest cluster contains the tests, *P. ae.*, *P. mi.*, *E. cl.*, *K. pn.*, *E. co.*, *M. mo.* and *S. ma.* (see Table 2 for abbreviations). Obviously, these tests are very similar, at least with respect to the substances investigated. Situated fairly close to this cluster is another cluster with the tests *E. fa.* and *S. au.*, again indicating similarity, but at the same time, suggesting that these two tests show some special behavior. The tests *S. pn.* and *B. fr.* are situated much further away and the distance of *B. fr.* is, in fact,

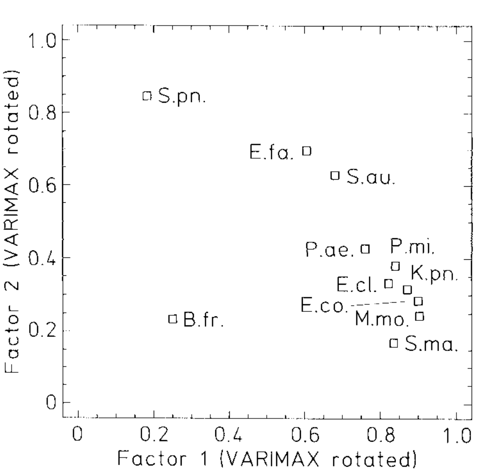


Figure 3. Loading plot of the antibacterial tests after VARIMAX rotation. For abbreviations of variables, see Table 2.

Table 2. Antibacterial potency of naphthyridines (log 1/*MIC*, *MIC* values taken from Bouzard et al. [63]) against various bacterial strains[1], and scores of the first factor (f_1)

	log 1/*MIC*											
No.	*S. pn.*	*E. fa.*	*S. au.*	*E. co.*	*K. pn.*	*E. cl.*	*P. mi.*	*M. mo.*	*S. ma.*	*P. ae.*	*B. fr.*	f_1
6	0.30	−0.30	1.79	2.69	1.79	2.09	0.88	1.52	0.88	0.00		
7	0.30	0.30	2.69	3.00	2.69	2.69	1.52	1.79	1.79	0.60	−0.30	1.074
8	0.00	0.00	1.22	1.69	1.52	1.22	0.88	1.22	1.22	0.30	−0.90	−0.215
9	0.00	−0.30	1.52	1.22	1.22	0.60	0.30	0.60	0.60	−0.60		
10	0.88	0.30	2.52	2.09	1.69	1.69	1.22	1.22	1.22	0.60	−0.60	0.41
11	0.60	0.30	1.69	1.52	1.22	0.88	0.60	0.88	0.88	−0.30	−1.50	−0.50
12	2.09	1.22	2.39	1.69	1.52	1.52	0.88	0.88	0.60	0.60	0.00	0.39
13	0.60	0.88	2.39	2.39	1.52	1.52	1.22	1.22	0.88	0.00	−0.60	0.25
14	1.52	−0.60	1.52	1.52	1.52	0.88	0.30	1.22	0.60	0.00	−1.20	−0.44
15	−0.30	−0.30	0.60	1.79	1.52	0.60	0.30	0.30	0.30	−0.30	−0.30	−0.90
16	−0.90	0.60	1.52	1.22	1.22	0.60	1.22	0.60	0.30			
17	0.00	0.00	1.52	2.09	1.69	1.52	0.88	1.22	0.88	0.30	−1.50	−0.15
18	0.60	−1.20	0.30	0.30	0.00	0.00	−0.90	−0.60	0.00	−1.20	0.00	−2.05
19		0.00	0.88	1.22	1.22	1.52	1.22	1.22	0.60	0.60		
20	0.88	0.88	2.69	2.69	2.69	2.69	1.79	1.82	1.52	0.60	0.00	1.20
21	0.88	0.88	1.52	2.09	1.79	1.22	0.88	1.22	0.88	0.30	−1.20	0.07
22	0.60	0.60	1.79	2.69	2.39	1.79	1.52	2.39	1.79	0.60	−0.30	0.90
23	0.88	0.88	2.39	2.69	2.39	2.69	2.09	1.79	2.39	0.60	0.60	1.36
24	0.00	−1.20	0.00	0.60	0.60	0.60	−0.30	0.30	0.30	−0.60	0.00	−1.52
25	−0.90	−1.80	−0.30	0.00	−0.30	−0.30	−1.20	−0.90	−1.20	−1.50	−2.09	−3.09
26	0.30	0.00	1.52	1.79	1.52	1.52	1.22	0.88	0.88	0.30	−0.30	−0.08
27	0.60	0.00	1.79	1.79	1.52	1.52	0.60	0.88	0.60	0.00	−0.60	−0.24
28	0.60	0.30	1.79	2.39	2.09	1.79	1.52	1.22	1.22	0.60	0.00	0.50
29	1.22	0.88	2.09	2.09	1.22	1.52	1.22	0.88	0.88	0.30	−0.60	0.20
30	1.22	0.60	2.09	2.09	2.09	2.09	0.88	0.88	0.30	0.30	−0.60	0.24
31	1.22	0.88	2.09	2.09	2.09	1.79	1.52	1.22	1.22	0.60	0.30	0.68
32	1.22	0.60	1.79	2.39	1.79	1.52	1.52	1.22	0.88	0.60	0.00	0.47
33	1.52	1.22	2.69	2.69	2.69	2.69	2.09	1.52	1.52	0.60	0.60	1.38
34	1.22	0.60	2.09	1.79	1.79	1.52	0.88	1.22	0.60	0.00	−0.30	0.13
35	1.52	1.22	2.39	2.69	2.69	2.69	2.09	2.09	1.52	0.60		

[1] *S. pn.*, *Streptococcus pneumoniae* A 9585; *E. fa.*, *Enterococcus faecalis* A 9809; *S. au.*, *Staphylococcus aureus* A 9537; *E. co.*, *Escherichia coli* A 15119; *K. pn.*, *Klebsellia pneumoniae* A 9664; *E. cl.*, *Enterobacter cloacae* A 9656; *P. mi.*, *Proteus mirabilis* A 9900; *M. mo.*, *Morganella morganii* A 15153; *S. ma.*, *Serratia marcescens* A 20019; *P. ae.*, *Pseudomonas aeruginosa* A 9843; *B. fr.*, *Bacillus fragilis* A 22862.

greater than is shown on the two-dimensional plot in Fig. 3, as this test is the only one which also has a high loading in the third factor. This clustering is in good agreement with phylogenetical properties of the bacterial strains considered. The conclusions, with respect to the compounds investigated, thus, are: (i) the set of tests is redundant, so that the same information could have been obtained with less bacterial strains; (ii) the tests *S. pn.* and *B. fr.* yield specific information, not contained in any other test and cannot, therefore, be dispensed with; (iii) it is sufficient to consider one test from each of the two clusters and, in addition, the tests *S. pn.* and *B. fr.* for SAR considerations or QSAR analyses.

The first factor mainly represents the tests in the large cluster (with some influence of the two tests in the smaller cluster). This factor may, thus, be regarded as an average representation of the results in these tests and can be directly submitted to a QSAR analysis, providing an overall SAR picture in these tests (factor scores are included in Table 3). Using Free-Wilson analysis [65] in the Fujita-Ban variant [66] (standard substituents: R_1 = *t*-Bu; R_5 = H; R_7 = ring Ril; see also Fig. 2), a statistically high significant result was obtained, which is represented in Table 3 together with the Free-Wilson results for *S. pn.* and *B. fr.* As was to be expected from the results of the factor analysis, the de novo activity contributions for these two strains differ markedly from those obtained for the first factor. When Free-Wilson analysis was applied to the original activity data of the tests, residing in the large cluster, the results (not shown) were compatible with the activity contributions from the first factor, as would be expected.

Table 3. Significant activity contributions from the Free-Wilson analysis of the factor 1 scores as well as for *S. pn.* and *B. fr.* (P = 95%; for the scores in italics, 90% < P < 95%)

Substituent	f_1	*S. pn.*	*B. fr.*
const.	*0.40*	0.43	1.11
t-Bu (R_1)	standard		
Et (R_1)	—		—
FEt (R_1)	*0.75*		
cPr (R_1)	0.73		
F_2Ph (R_1)	*0.43*		0.67
FPh (R_1)	—		—
F-t-Bu (R_1)	−1.02		
H (R_5)	standard		
Me (R_5)			
Et (R_5)	−2.48		
Ph (R_5)	−4.21	−1.89	−2.12
Ri1 (R_7)	standard		
Ri2 (R_7)	−0.88	−0.72	−0.78
Ri3 (R_7)			*−0.51*
Ri4 (R_7)	−0.74		−1.03
Ri1Me (R_7)			
r	0.965	0.788	0.838
s	0.271	0.403	0.361
F	16.2	1.8	2.8

The use of principal components or factors as dependent variables in QSAR in order to represent a set of biological tests is very convenient, if the data structure enables this (see also Eqs. (24) – (26) and (29)). Other examples where this has been achieved for antibacterial potencies have been presented the literature [34, 40].

Example: Antihypertensive Fusaric Acids In Vitro and In Vivo

Another typical example, which will be discussed briefly here, concerns antihypertensive fusaric and picolinic acids (Fig. 4). These compounds are believed to act by inhibiting catecholamine biosynthesis, via blockade of the enzyme dopamine β-hydroxylase (DBH) that converts dopamine into noradrenaline. However, other mechanisms of action have also been discussed, and the question that remains to be answered, is whether the blood pressure decreasing effect of these compounds is due to DBH inhibition or not (see, e.g. Dove et al. [67] and references cited therein). The following measurements have been made for a number of analogs [68, 69]:

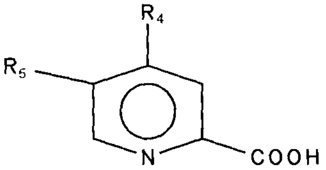


Figure 4. Fusaric (R_5 = *n*-Bu) and picolinic (R_5 = H) acids.

- Inhibition of DBH in vitro:

  pI_{50}(I): incubation with an excess of Cu^{2+} ions (copper complex formation may play an important role),

  pI_{50}(II): without Cu^{2+} excess,
- Decrease of systolic (index "*S*") and diastolic (index "*D*") blood pressure in male Wistar rats with renal hypertension at doses of 0.25 mmol/kg (index "*a*") and 0.5 mmol/kg (index "*b*"): *BRSa*, *BRSb*, *BRDa* and *BRDb* (transformed into logarithms),
- pK_a and pK_b values.

All data (taken from Ref. [67]) are summarized in Table 4 together with the values of $\Sigma\,\pi$ and $\Sigma\,\sigma$. In addition, a classification of compounds with regard to their in vivo potency is also included [67]: class 0 = "inactive" compounds; class 1 = "active" compounds. Missing pK_a and pK_b values were estimated from the following relationships, obtained from experimentally available values and Swain-Lupton's *F* and *R* values:

$$\begin{aligned} \mathrm{p}K_a &= -6.77(\pm 1.39)R(\mathrm{R}_4) - 4.06(\pm 1.17)F(\mathrm{R}_5) \\ &\quad - 0.91(\pm 0.86)R(\mathrm{R}_5) + 5.33(\pm 0.50) \\ n &= 15 \quad r = 0.977 \quad s = 0.291 \quad F = 52.9 \end{aligned} \tag{27}$$

$$\begin{aligned} \mathrm{p}K_b &= 2.53(\pm 1.00)F(\mathrm{R}_4) + 1.81(\pm 0.46)F(\mathrm{R}_5) \\ &\quad + 1.41(\pm 0.30)R(\mathrm{R}_5) + 13.01(\pm 0.19) \\ n &= 18 \quad r = 0.974 \quad s = 0.126 \quad F = 60.8 \end{aligned} \tag{28}$$

Table 4. Pharmacological data of antihypertensive fusaric and picolinic acids. pI_{50} = inhibition of DBH in vitro (I: incubation with on excess of Cu^{2+} ions; II: no excess of Cu^{2+} ions), *BR* = log (decrease blood pressure), male Wistar rats with renal hypertension: *S* = systolic, *D* = diastolic, *a* = 0.25 mmol/kg, *b* = 0.5 mmol/kg; class. = classification for in vivo activity (0 = inactive, 1 = active)

No.	R_4	R_5	pI_{50}(I)	pI_{50}(II)	*BRSa*	*BRDa*	*BRSb*	*BRDb*	class.	Σ π	Σ σ
1	H	H	5.35	5.35	0.98	1.02	0.63	1.17	0	0.00	0.00
2	H	Et	5.59	4.77			1.87	1.60	1	1.02	−0.15
3	H	n-Bu	6.05	5.96					1	1.98	−0.20
4	H	OH	5.77	5.74	1.39	1.35	1.23	0.94	0	−0.67	−0.37
5	H	OMe	–	5.64	0.94	0.93	0.92	0.94	0	−0.02	−0.27
6	H	O-*n*-Pr	–	5.89	1.20	0.78	1.37	1.39	1	1.05	−0.25
7	H	O-*n*-Bu	6.60	5.76	1.42	0.98	1.33	0.79	1	1.55	−0.32
8	H	O-Benzyl	–	6.28	1.60	1.48	1.51	1.26	1	1.66	−0.33
9	H	COOH	4.92	5.28	0.63	0.97		0.98	0	−0.32	0.45
10	H	Cl	5.62	4.39	0.92	0.54	1.48	1.11	0	0.71	0.23
11	H	Br	–	4.80	0.94	1.07	1.11	1.24	0	0.86	0.23
12	H	I	5.35	4.96	1.56	1.46	1.67	1.72	1	1.12	0.18
13	H	NH_2	–	5.06	0.58	0.74	1.26	1.16	0	−1.23	−0.66
14	H	NHOH	4.88	4.74	1.10	0.89	1.17	0.70	0	−1.34	−0.34
15	H	NHCOMe	7.00	5.14					0	−0.97	0.00
16	H	NHCOEt	–	5.47	1.02	0.70			0	−0.43	0.00
17	H	NO_2	3.30	4.05	0.57	0.36	0.72		0	−0.28	0.78
18	H	CN	3.30	3.39						−0.57	0.71
19	OMe	*n*-Bu	5.00	5.04	1.44	1.50	1.77	1.76	1	1.96	−0.08
20	OEt	*n*-Bu	5.02	5.47	1.49	1.24	1.70	1.54	1	2.36	−0.10
21	Cl	*n*-Bu	6.02	6.82	1.31	1.22	1.16	1.03	1	2.69	0.17
22	NO_2	*n*-Bu	5.10	4.96	1.09	0.96		0.86	1	1.70	0.51

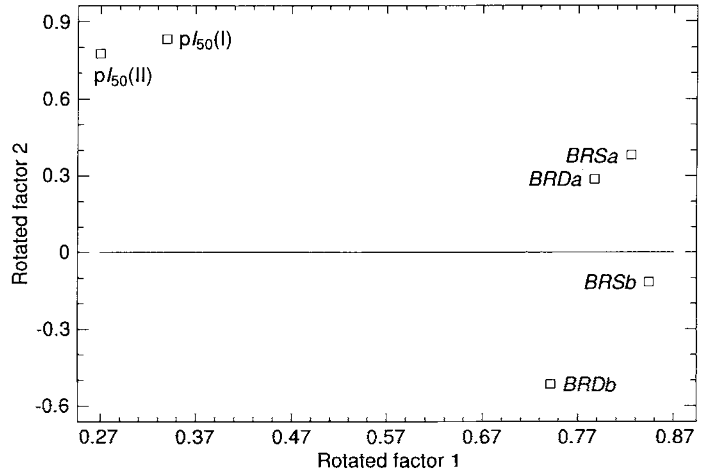


Figure 5. Loading plot of variables related to the inhibition of DBH and to the decrease of blood pressure in vivo. For abbreviations of variables, see text.

A factor analysis of the 6 biological potencies extracted 93.3% of the data variance with two factors (factor 1: $\lambda_1 = 2.99$, % variance = 61.5; factor 2: $\lambda_2 = 1.55$, % variance = 31.8). The loading plot in Fig. 5 shows that the in vivo results afford factor one, while factor two is mainly afforded by the two in vitro tests. It then follows that there is no simple correlation between in vivo and in vitro tests so that the in vivo potency does not reflect DBH blocking activity in vitro. The classification into classes 0 and 1, however, can be described by a discriminant function with $\Sigma\,\pi$ and pI_{50}(II) as variables [67], which can be interpreted to mean that the compounds are active in vivo, if the DBH inhibition is sufficiently high, and hydrophobicity allows for efficient transport to the site of action.

Fig. 6 shows a scatter plot of compounds obtained from a principal component analysis with the variables pK_a, pK_b, $\Sigma\,\pi$ and $(\Sigma\,\pi)^2$. This analysis afforded two significant components accounting for 65.4% ($\lambda_1 = 1.83$) and 34.6% ($\lambda_2 = 0.97$) of the data variance, respectively, (total: 100%). Compounds were labeled according to their class membership for in vivo potency, and as can be seen, the two classes are clearly separated. Thus, scatter plots, obtained from principal component (or factor) analysis of physico-chemical variables, can produce patterns, in which compounds are clustered according to some biological property which is not included in the analysis. This is of some practical importance, as such plots may be of use in selecting compounds for further investigations. If in such an analysis compounds with known biological properties and new compounds with unknown biological properties are included, those new compounds which are in the vicinity of (or within) clusters containing the already tested analogs possessing the desired property, are the best candidates for synthesis (or testing, if already synthesized). This problem can, of course, also be solved by applying classification methods, including classification, which is based on principal components (SIMCA; see Chap. 4.3). However, such scatter plots have the advantage, in that they not only

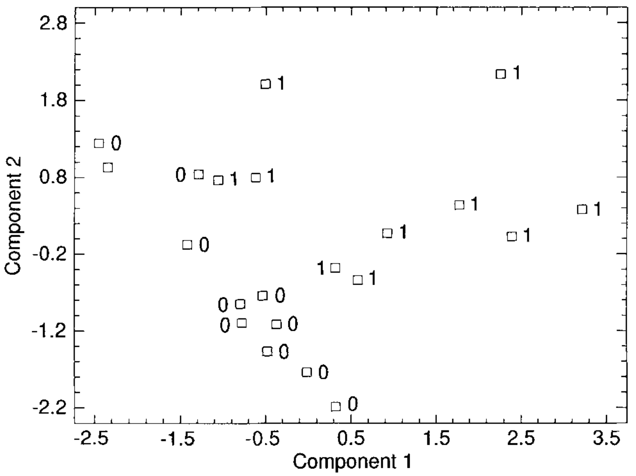


Figure 6. Scatter plot of fusaric acids in the space spanned by the first two principal components. Variables: pK_a, pK_b, $\Sigma\,\pi$, $(\Sigma\,\pi)^2$; 0 = inactive compounds; 1 = active compounds.

allow a selection of compounds with respect to a desired potency but that aspects of a series design can also be included (see page 153).

The application of principal component analysis in QSAR work often ends at the stage of such clustering, i.e. the method serves as a cluster analysis approach. In order to obtain a clearer picture, cluster analysis can subsequently be applied to further analyze such scatter plots [70]. Some examples of using simple principal component analysis for cluster analysis are given in the literature [71 – 74]. The objective is similar to that of classification methods (pattern recognition). If classes of compounds are known before the analysis, however, this approach works better with a separate PC model for each class as in the SIMCA method of Wold et al. (see Chap. 11). A special classification procedure based on principal component analysis has been recently proposed by Rose et al. [75]. A somewhat different application of principal component analysis can be found in the work of Cammarata and Menon [76, 77]. They derived a data matrix for compounds with different modes of action, by coding the presence or absence of groups at designated positions and weighing these codes by the molar refractivity of the groups present. A coordinate system, with the resulting components as axes, was then used to plot the data, leading to a certain clustering of compounds with similar types of biological action.

Example: Sulfones and Sulfonamides in Whole-Cell and Cell-Free Systems

The antibacterial effects of 17 4′-substituted 4-aminodiphenylsulfones in 7 cell-free folate synthesizing enzyme extracts and in 2 whole cell cultures of various mycobacterial strains and strains of *E. coli* sensitive and resistant to sulfones, have been determined by Seydel and coworkers [34] and submitted to principal component analysis. Missing data 19% were estimated in an iterative process

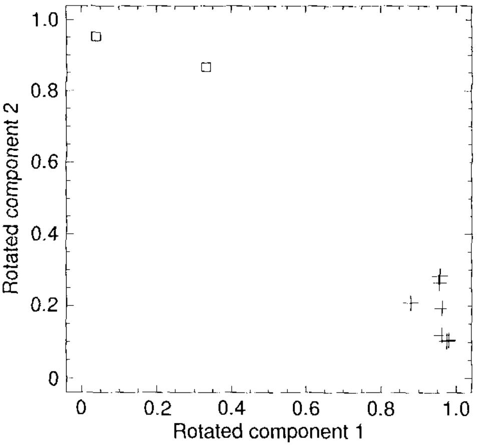


Figure 7. Loading plot of variables related to antibacterial potency (after VARIMAX rotation). Whole cell systems are represented by squares, and cell-free systems are represented by crosses in accordance with Ref. [34].

within principal component analysis. Two significant principal components were obtained ($\lambda_1 = 6.94$ and $\lambda_2 = 1.45$), accounting for 77.1% and 16.1% of the data variance, respectively (total: 94.2%).

The loading plot in Fig. 7 shows that the cell-free test systems essentially afford the first component, while the second component represents whole cell activities. This means that potency in cell-free and whole cell systems are governed by different factors. The first component obviously reflects an "average" of enzyme inhibition, while the second could be related to transport through the cell membrane. If this is true, scores of the first component should be related to physico-chemical parameters in the same way as found for the cell-free system data, while the second component should show a relationship to hydrophobicity, as would be typical for transport processes. This is indeed the case, as P_1 correlates with the electronic demand of substituents and with the fraction ionized, while P_2 shows a bilinear dependence on lipophilicity, which is expressed by the HPLC parameter, $\log k'$, with the optimum at $\log k' = 0.83$:

$$P_2\,(\text{rotated}) = 1.40(\pm 0.52)\log k' - 3.49(\pm 1.32)\log[0.098(\pm 0.173)k' + 1] + 0.507(\pm 0.726) \quad (29)$$

$$n = 17 \qquad r = 0.934 \qquad s = 0.396 \qquad F = 22.27$$

This is another example of how principal component analysis can be of use to order data, so that the intrinsic potency at the site of action can be separated from transport phenomena.

4.1.3.2 Pharmacokinetic Data and Time Series

Pharmacokinetic properties are characterized by a variety of parameters, which reflect different aspects of pharmacokinetics. Many of these parameters are intercorrelated and, thus, lend themselves to multivariate analyses, such as principal component and factor analysis, with the aim of finding components (factors), characteristic of fundamental pharmacokinetic processes. If pharmacokinetic and response data are combined, there is the opportunity to separate pharmacodynamic and pharmacokinetic effects. The same goal may be achieved, furthermore, if time series of pharmacological response measurements are subjected to principal component or factor analysis. In any case, data will be simplified and, thus, be easier to interpret and subsequent QSAR analyses are made considerably easier.

Illustrative examples for applying principal component analysis to pharmacokinetic data can be found in a paper by Schaper and Seydel [78]. Considering five examples, it was possible to represent complex data by relatively simple principal component models. The components could be identified with basic processes such as elimination, protein binding, and distribution and also showed correlations with hydrophobicity, as was to be expected for these processes. Pharmacodynamic effects could be separated from pharmacokinetic effects for 11 morphine-like analgesics by analyzing the following variables using principal component analysis (after logarithmic transformation): the times for the onset of maximum activity and $1/ED_{10}$ values, respectively, after intravenous and intraventricular administration, $1/\mathrm{ED}_{50}$ in the hot plate assay, and binding to rat brain homogenates. The first of two significant components represented intrinsic activity, while the second could be attributed to pharmacokinetic processes and showed a bilinear relationship with log P. Similar results were reported for pyrethroids in insects by Ford et al. [79, 80] using canonical correlation analysis.

The first principal component analysis of a time series was performed by Franke and co-workers [81, 82] with data for the antiinflammatory potency of 14 disubstituted salicylic acids (against carageenin edema in Wistar rats; data supplied by Bekemeyer [83]) measured 3, 4, and 5 hours after administration. Two components were obtained, with the first component reflecting the pharmacodynamic part of the observed effect and the second, pharmacokinetics. The first component could be related to substituent constants, indicating that steric effects and the presence of a free carboxyl group are important for antiinflammatory potency. This example shows the potential of such an analysis, but suffers from the limited number of times considered and a low data variance. One cannot expect that a separation of pharmacokinetic and pharmacodynamic effects will always be possible by analyzing a time series. If, for example, pharmacokinetic and pharmacodynamic effects depend similarly on hydrophobicity (which may frequently happen for unspecific effects), principal component analysis of a time series may simply produce one component which represents both effects. An example for this is given by Schaper and Seydel [78] in the non-specific cardiodepressive effects of β-blockers:

Analysis of a time series can aid in dissecting more complex pharmacological processes into components, with the final objective being to derive complex quantitative structure-time-activity relationships (QSTAR); a further example will be presented in the next section. If, in a series of compounds, the concentrations at the site of action at different times all relate similarly to log P according to the bilinear model of Kubinyi [84], or as a parabolic relationship [85], and pharmacodynamic effects are absent, only one component will be obtained.

Table 5. Analgesic potencies (log $(1/ID_{50})$ in rats, tail withdrawal test [86]; $t1 - t10$) of fentanyl $t = 1/32$ h; $t2$: $t = 1/16$ h; $t3$: $t = 1/8$ h; $t4$: $t = 1/4$ h; $t5$: $t = 1/2$ h; $t6$: $t = 1$ h; $t7$: $t = 2$ h; $t8$: *Max* = maximal potency = log $(1/ID_{50})_{max}$. f_1: factor scores of factor 1; f_2: factor scores

No.	R_1	R_2	R_3	R_4	R_5	$t1$	$t2$	$t3$	$t4$
1	H	H	CH_2OCH_3	C_2H_5	C_6H_5	8.46	8.72	8.70	8.66
2	H	H	CH_2OCH_3	C_2H_5	C_4H_3S	8.54	8.64	8.76	8.64
3	H	H	$COOCH_3$	C_2H_5	C_6H_5	8.72	9.00	9.00	9.03
4	H	H	$COOCH_3$	C_2H_5	C_4H_3S	8.77	8.95	9.06	8.96
5	CH_3	H	$COOCH_3$	C_2H_5	C_6H_5	8.25	8.66	8.70	8.76
6	H	H	$COOCH_3$	C_2H_5	C_6H_5	8.27	8.46	8.61	8.67
7	H	CH_3	$COOCH_3$	C_2H_5	C_6H_5	8.27	8.57	8.62	8.72
8	H	H	$COCH_3$	C_2H_5	C_6H_5	8.54	8.68	8.75	8.78
9	H	H	$COCH_3$	c-C_3H_5	C_6H_5	8.44	8.69	8.64	8.71
10	H	H	COC_2H_5	C_2H_5	C_6H_5	8.34	8.45	8.48	8.48
11	H	H	COC_2H_5	C_2H_5	C_4H_3S	8.45	8.79	8.77	8.72
12	H	H	H	C_2H_5	C_6H_5	7.41	7.45	7.48	7.38
13	H	$-CH_3$	H	C_2H_5	C_6H_5	8.65	8.73	8.75	8.75

Example: Decomposition of Time-Dependent Response Data by Factor Analysis

For analgesic potencies of a series of fentanyl derivatives (for structure see Fig. 8), measured after 10 different times [86] (see Table 5), Balaz et al. [87] derived the following relationship starting from a model-based disposition function:

$$\log (1/ID_{50}) = -\log (BP + 1) - Dt/(BP + 1) + A \qquad (30)$$

$$A = 8.923\,, \qquad B = 1.492 \cdot 10^{-3}\,, \qquad D = 5.461$$

Eq. (30) is statistically highly significant and describes the entire data (compounds **6** and **16** not included). The first term is supposed to describe transport and the second term is supposed to describe elimination. Both transport and elimination depend on hydrophobicity.

As an alternative to starting from a well-defined model, the same data (taken from Ref. [87]) were submitted to factor analysis [87a] using the statistical program package STATGRAPHICS [64].

Two significant factors were obtained with eigenvalues of $\lambda_1 = 7.22$ and $\lambda_2 = 1.60$, accounting for 80.6% and 17.9% of the variance, respectively (total: 98.5%). Factor scores (VARIMAX rotated) are included in Table 5. All vatriables have high communalities and are, thus, well represented in common factor space. A plot of

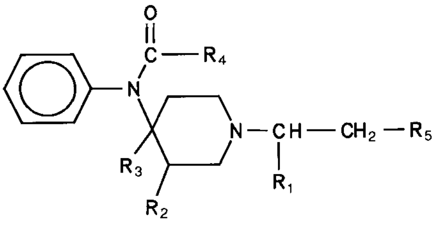


Figure 8. Fentanyl derivatives.

derivatives (structure see Fig. 8) measured at different times, data taken from Balaz et al. [87]. *t*1: *t* = 4 h; *t*9: *t* = 6 h; *t*10: *t* = 8 h. Log *P* values and values of *Max* Balaz et al. [87] of factor 2

*t*5	*t*6	*t*7	*t*8	*t*9	*t*10	log *P*	*Max*	f_1	f_2
8.58	8.24	7.61	6.87	6.07		2.74	8.74	−0.04	−0.61
8.54	8.12	7.59	6.67	6.21		2.39	8.73	−0.08	−0.69
8.97	8.78	8.36	7.34	6.55	6.04	2.54	9.09	0.90	−0.55
8.89	8.65	8.19	7.28	6.42	5.95	2.19	9.03	0.79	−0.73
9.69	8.50	8.31	7.57	6.86	6.46	2.85	8.80	0.32	0.57
8.66	8.53	8.43	8.33	8.21	7.96	2.65	8.66	0.45	1.86
8.73	8.72	8.69	8.46	8.43	8.19	2.94	8.76	0.68	2.04
8.74	8.43	7.86	6.64	6.06		2.26	8.82	0.16	−0.75
8.65	8.41	7.92	6.89	6.25		2.37	8.74	0.09	−0.34
8.27	8.06	7.39	6.39	5.95		2.79	8.15	−0.57	−0.51
8.56	8.03	7.28	6.48	6.03		2.44	8.85	−0.16	−1.05
7.14	6.86	6.29	5.56	5.03		2.35	7.48	−3.03	0.59
8.72	8.54	8.05	7.56	7.07	6.76	2.75	8.78	0.48	0.18

the factor loadings against time of measurement is presented in Fig. 9. As can be seen, the proportion of variance, accounted for by the first factor, decreases with increasing time of measurement (squares), while the variance explained by the second factor increases with time (crosses). According to Balaz et al. [87], the data are primarily dependent on pharmacokinetics, with transport being much faster than elimination, while differences with respect to receptor affinity were regarded to be very small. As the first factor reflects those effects common to all measurements, while the second factor represents differences, it is, thus, tempting to speculate that the first factor is *primarily* related to the pharmacokinetic processes (which will lose in importance as time increases) while the second factor can be attributed to receptor affinity (pharmacodynamic effect).

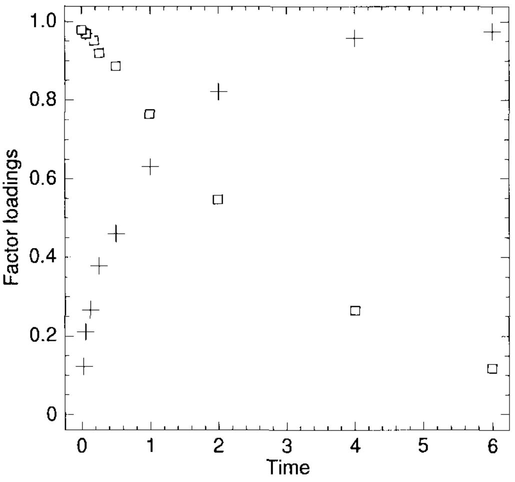


Figure 9. Plot of factor loadings (VARIMAX rotated) against time of measurement. Squares: factor 1; crosses: factor 2

If this were true, one would expect the following:

1. Scores of factor 1 should be related to hydrophobicity, as this is the dominating property governing transport.
2. With scores of factor 2 as independent variable, it should be possible to derive a QSAR for receptor affinity, if such a QSAR exists.
3. A QSAR for receptor affinity can only be obtained with the original measurement, if the pharmacokinetic effect is not a dominating feature. According to the factor loadings, this is true for $t8$ and $t9$ (see Fig. 9).

A simple relationship exists between scores of factor 1 and log P after eliminating compounds **6** and **12**, which behaved as pharmacokinetical outliers [87] and were also left out in the derivation of Eq. (30) ($FA1$ = scores of factor 1 after VARIMAX rotation):

$$FA1 = -0.63(\pm 0.36)\,(\log P)^2 + 1.76(\pm 0.95)\log P \qquad (31)$$

$$n = 11 \qquad r = 0.851 \qquad s = 0.313 \qquad F = 11.8$$

Eq. (31) is statistically highly significant. Compound **3** shows a relatively large deviation from the regression line for reasons unknown. If this compound is also removed Eq. (3) is then improved to give:

$$FA1 = -0.60(\pm 0.28)\,(\log P)^2 + 1.65(\pm 0.74)\log P \qquad (32)$$

$$n = 10 \qquad r = 0.891 \qquad s = 0.236 \qquad F = 15.4$$

A plot of the observed versus predicted values of $FA1$ is shown in Fig. 10. A parabolic expression in log P is usually considered to be a good approximation to describe pharmacokinetic processes. Non-linear regression analysis does not result in an improvement but provides a description which corresponds very well to the model in Eq. (30):

$$FA1 = -\log(0.0037P + 1) - 3.56/(0.0037P + 1) \qquad (33)$$

$$n = 10 \qquad r = 0.888 \qquad s = 0.241 \qquad F = 14.7$$

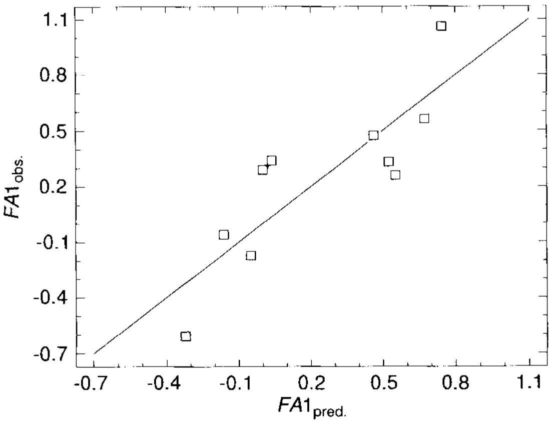


Figure 10. Plot of observed versus predicted (Eq. (32)) values of $FA1$.

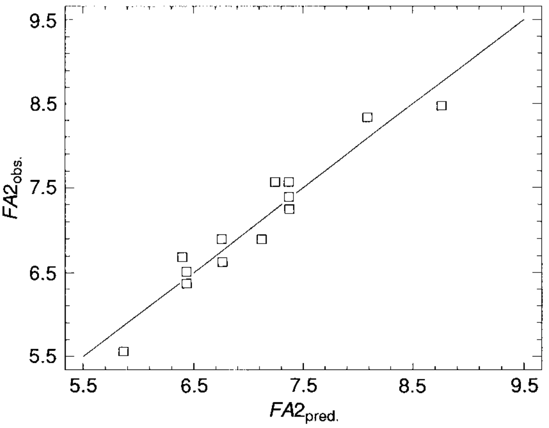


Figure 11. Plot of observed versus predicted values of *t*8 (from Free-Wilson analysis).

Eqs. (31) to (33) are, thus, in good agreement with the hypothesis that the first factor primarily represents the pharmacokinetic aspect of biological potencies.

When Free-Wilson analysis was applied to the scores of the second factor (*FA*2; VARIMAX rotated) and to scores for *t*8 and *t*9, highly significant results were obtained (see also Fig. 11), which had the following statistical characteristics ($n = 13$):

$$FA2\colon\ r = 0.963,\ s = 0.270,\ F = 12.8$$
$$t8\colon\ r = 0.965,\ s = 0.208,\ F = 13.8$$
$$t9\colon\ r = 0.963,\ s = 0.250,\ F = 12.7$$

As expected, a significant Free-Wilson solution does not exist for the other measurements ($t1 - t7$).

It was very satisfying to see how the results of factor analysis could indicate which measurements could be employed to derive a QSAR. The results obtained were fully consistent with the hypothesis that factor analysis has achieved a separation of the pharmacokinetic and pharmacodynamic effects.

4.1.3.3 Analysis of QSAR Descriptors

In statistical QSAR analyses chemical compounds (or substituents) are usually described by physico-chemical parameters and/or substituent constants as independent variables. These parameters can be divided roughly into three main groups: hydrophobic, electronic, and steric. A large number of such parameters coexists with many intercorrelations and redundancies. One way of gaining a better understanding would be to apply principal component or factor analysis to data matrices, in which descriptor variables or physico-chemical properties are listed for a representative set of substituents or compounds. This has been done with the following objectives:

1. Many parameters, as for example, hydrophobic and electronic substituent constants or log *P*, are determined by experimental measurements. Parameter values, which are obtained in this way, can be very sensitive to experimental

conditions or the type of molecules investigated. Principal component or factor analysis can be used in this case in order to understand the underlying effects and/or to create unified scales of descriptor values.

2. A grouping of descriptor variables can be obtained, leading to a better understanding of their nature and relatedness as well as to their information content, which can aid in the selection of variables for QSAR analyses.
3. Principal components, extracted from a set of descriptors or measurements, can serve as new variables in QSAR. Such variables are often referred to as "principal properties".
4. Principal component or factor analysis of descriptor variables can be used as tools in the design of training series with high information content.

These issues will be dealt with in the following sections. Only such examples will be included, which are directly related to drug design; applications outside this field (e.g. as in general chemistry) will not be considered.

Variation of Descriptors with Experimental Conditions

The most important cases, which are related to drug design are:

1. Variation of electronic and hydrophobic substituent constants of substituent X with the nature a functional group Y in aromatic molecules, with the general structure shown in Fig. 12.
2. Variation of log P, π or chromatographic hydrophobicity parameters with organic solvent.

Interactions between X and Y have led to a great number of modifications in the original Hammett equation and, thus, to many scales of σ values (see also e.g. Franke [6, 7]). Using principal component analysis, Wold and Sjöström [88–91] investigated a large number of reaction series and arrived at a unified, and later extended, σ-scale.

Franke and coworkers applied principal component analysis to aromatic π values [92, 93] (see also Ref. [7]) for *meta-* and *para-*substituents in phenyls with different functional groups, Y, and to log P values measured in different solvent/water systems [94, 95]. The principal component analysis of π values resulted in two components, with the first component reflecting "intrinsic hydrophobicity", and the second component reflecting electronic corrections. This example will be discussed in some detail in the next section. In the case of the log P values, which were determined with 18 structurally diverse solutes in 6 solvent/water systems, again two significant principal components were obtained, accounting for 71.7% and 24.5% of the data variance, respectively (total: 96.2%). The first component correlated with the cavity surface area, according to Hermann [96, 97], and to the B and C values of Cramer [98, 99] (B and C represent principal properties derived by principal component

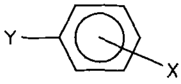


Figure 12. Disubstituted benzenes: X = substituent; Y = functional group.

analysis and will be discussed later on page 144) according to the following:

$$P_1 = 1.18(\pm 0.35)CSA - 2.61(\pm 1.90)C - 2.88(\pm 0.92) \tag{34}$$
$$n = 18 \qquad r = 0.836 \qquad s = 0.294$$

$$P_1 = 3.42(\pm 0.71)B - 5.01(\pm 1.69)C + 0.41(\pm 0.26) \tag{35}$$
$$n = 18 \qquad r = 0.910 \qquad s = 0.223$$

As *CSA* and *B* are measures of bulk effects, and *C* represents polar effects, the first component represents an "average" of the hydrophobicity common to all partitioning systems, which can be attributed to the bulk contribution and, in addition, to a polar contribution, which accounts for more specific solute-solvent interactions. The second component is highly correlated with the hydrogen bonding parameters of the solute molecules, such as, for example, Seiler's [100] I_H value:

$$P_2 = 0.28(\pm 0.03)I_H - 0.54(\pm 0.07) \tag{36}$$
$$n = 18 \qquad r = 0.964 \qquad s = 0.080$$

Table 6. Aromatic π-values of 14 substituents (*meta* and *para* positions) in 8 series of standard compounds; PhOAA = phenoxyacetic acids, PhAA = phenylacetic acids, B = benzenes, BA = benzoic acids, P = phenols, A = anilines, PAA = piperidinoacetanilides, NB = nitrobenzenes

No.	R	PhOAA	PhAA	B	BA	P	A	PAA	NB
1	H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	*m*−F	0.13	0.19	0.14	0.28	0.47	0.40	0.39	0.21
3	*m*−Cl	0.76	0.68	0.71	0.83	1.04	0.98	0.99	0.61
4	*m*−I	1.15	1.22	1.12	1.28	1.47	2.08	1.46	0.99
5	*m*−Br	0.97	0.91	0.86	0.99	1.17	1.20	1.16	0.79
6	*m*−CH_3	0.51	0.49	0.56	0.52	0.56	0.50	0.56	0.57
7	*m*−CF_3	1.07	1.16	0.66	1.07	1.49	1.54	1.36	0.87
8	*m*−OCH_3	0.12	0.04	−0.02	0.14	0.12	0.03	0.17	0.31
9	*m*−OH	−0.49	−0.52	−0.67	−0.38	−0.31	−0.73	−0.45	0.15
10	*m*−NO_2	0.11	0.01	−0.28	−0.05	0.54	0.47	0.45	−0.36
11	*m*−COOH	−0.15	−0.32	−0.28	−0.19	0.04	−0.18	−0.11	−0.02
12	*m*−CN	−0.30	−0.28	−0.57	−0.37	0.24	−0.25	−0.18	−0.68
13	*m*−$COCH_3$	−0.28	−0.36	−0.55	−0.34	−0.07	−0.27	−0.33	−0.43
14	*m*−CH_2OH	−0.82	−0.76	−1.03	−0.84	−1.02	−0.95	−0.84	−0.65
15	*p*−F	0.15	0.14	0.14	0.19	0.31	0.25	0.25	0.16
16	*p*−Cl	0.70	0.70	0.71	0.87	0.93	0.93	0.89	0.54
17	*p*−I	1.43	1.23	1.12	1.14	1.45	2.44	1.44	1.02
18	*p*−Br	1.19	0.90	0.86	0.98	1.13	1.36	1.12	0.70
19	*p*−CH_3	0.60	0.45	0.56	0.42	0.48	0.49	0.50	0.52
20	*p*−CF_3	1.13	0.87	0.66	0.83	1.24	1.72	1.30	0.80
21	*p*−OCH_3	−0.04	0.15	−0.02	0.08	−0.12	0.05	0.04	0.18
22	*p*−OH	−0.61	−0.14	−0.67	−0.30	−0.67	−0.86	−0.36	0.11
23	*p*−NO_2	0.24	−0.04	−0.28	−0.02	0.50	0.49	0.48	−0.39
24	*p*−COOH	−0.22	0.30	−0.28	−0.05	0.12	0.20	−0.02	0.03
25	*p*−CN	−0.32	0.01	−0.57	−0.31	0.14	−0.15	−0.05	−0.66
26	*p*−$COCH_3$	−0.37	−0.45	−0.55	−0.32	−0.11	−0.36	−0.26	−0.36
27	*p*−CH_2OH	−0.78	−1.32	−1.03	−0.91	−1.26	−1.30	−0.87	−0.60

The second component, thus, reflects specific differences between the partitioning systems, which are related to hydrogen bonding.

Principal component analyses of log P in disubstituted benzenes and of a set of solutes in different solvent/water systems were also performed by Dunn and coworkers, who reached similar conclusions, but with a slightly different interpretation [101 – 105].

Example: Factor Analysis of Aromatic π Values from Different Series of Standard Compounds

The principal component analysis presented by Franke et al. [92, 93] will be repeated here for the sake of illustrating factor analysis [64]. Investigations in [92, 93] started from a known physical model, which applied to principal component analysis, while, in this section, it will be assumed that no model is known, and factor analysis is then used to create a model. In addition *meta*- and *para*-substituents will be treated simultaneously, whereas previously they were analyzed separately in [92, 93].

The π values for 14 substituents at the *meta* and *para* position (including H) for 8 series of standard compounds (different Y functional group, see Fig. 12) are summarized in Table 6. Factor analysis revealed two significant factors with eigenvalues of $\lambda_1 = 7.51$ and $\lambda_2 = 0.24$, accounting for 96.2% and 3.0% of the data variance, respectively (total: 99.2%). The factor matrix, after VARIMAX rotation, is shown in Table 7, and factor scores (unrotated) are summarized in Table 8.

A plot of the factor scores is presented in Fig 13. The *meta*- and *para*-substituents are arranged in pairs in a pattern, which reflects their hydrophobicity and electronic properties: substituents are arranged according to their hydrophobicity from left to right and according to their electron withdrawing power from top to bottom. This suggests that the first component reflects basic hydrophobicity of substituents, while the second component represents electronic substituent proper-

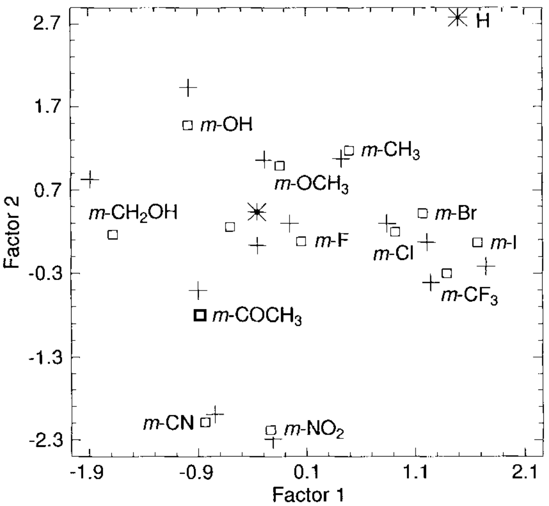


Figure 13. Score plot of substituents. Squares represent *meta*-substituents, crosses represent *para*-substituents and a star represents hydrogen. *Meta*- and *para*-substituents are arranged in pairs (labels only at the *meta*-substituents).

Table 7. Factor matrix (unrotated and after VARIMAX rotation) and communalities. PhOAA = phenoxyacetic acids, PHAA = phenylacetic acids, B = benzenes, BA = benzoic acids, P = phenoles, A = anilines, PAA = piperidinoacetanilides, NB = nitrobenzenes

Compound series	Factor 1		Factor 2		Communality
	Unrot.	VARIMAX	Unrot.	VARIMAX	
PhOAA	0.99	0.77	−0.04	0.62	0.974
PhAA	0.97	0.73	−0.00	0.64	0.946
B	0.98	0.66	0.13	0.74	0.974
BA	0.99	0.70	0.08	0.71	0.995
P	0.96	0.87	−0.22	0.46	0.969
A	0.97	0.83	−0.15	0.52	0.967
PAA	0.99	0.82	−0.12	0.56	0.990
NB	0.90	0.45	0.35	0.85	0.930

Table 8. Factor scores

No.	Substituent	Factor 1	Factor 2
1	H	−0.35	0.43
2	*m*-F	0.05	0.07
3	*m*-Cl	0.90	0.20
4	*m*-I	1.67	0.05
5	*m*-Br	1.18	0.41
6	*m*-CH_3	0.47	1.14
7	*m*-CF_3	1.39	−0.31
8	*m*-OCH_3	−0.16	1.02
9	*m*-OH	−0.98	1.46
10	*m*-NO_2	−0.23	−2.21
11	*m*-COOH	−0.59	0.24
12	*m*-CN	−0.84	−2.05
13	*m*-$COCH_3$	−0.87	−0.80
14	*m*-CH_2OH	−1.67	0.15
15	*p*-F	−0.05	0.29
16	*p*-Cl	0.84	0.29
17	*p*-I	1.75	−0.21
18	*p*-Br	1.20	0.06
19	*p*-CH_3	0.42	1.08
20	*p*-CF_3	1.24	−0.41
21	*p*-OCH_3	−0.28	1.06
22	*p*-OH	−0.98	1.93
23	*p*-NO_2	−0.21	−2.29
24	*p*-COOH	−0.35	0.03
25	*p*-CN	−0.74	−1.99
26	*p*-$COCH_3$	−0.89	−0.50
27	*p*-CH_2OH	−1.88	0.83

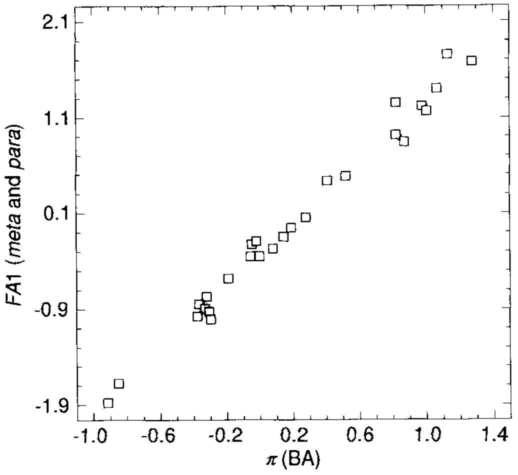


Figure 14. Plot of scores from the first factor ($FA1$) against π values from the benzoic acid system (π(BA)).

ties. This is, indeed, the case as indicated in Figs. 14 and 15. Fig. 14 shows a plot of the scores from factor 1 against π values for the benzoic acid series (π(BA) was selected, because this variable has a communality of 1.00). Fig. 15 shows the scores from factor 2 plotted against σ of the substituents. Correlating factor scores with hydrophobic and electronic substituent constants leads to Eqs. (37) and (38):

$$f_1 = 1.61(\pm 0.07)\,\pi(\text{BA}) + 0.19(\pm 0.14)\sigma_{\text{m,p}} - 0.38(\pm 0.05) \tag{37}$$

$$n = 27 \qquad r = 0.995 \qquad s = 0.099 \qquad F = 1225.0$$

$$f_2 = -3.23(\pm 0.77)\sigma_{\text{m,p}} + 0.82(\pm 0.30) \tag{38}$$

$$n = 27 \qquad r = 0.863 \qquad s = 0.557 \qquad F = 72.5$$

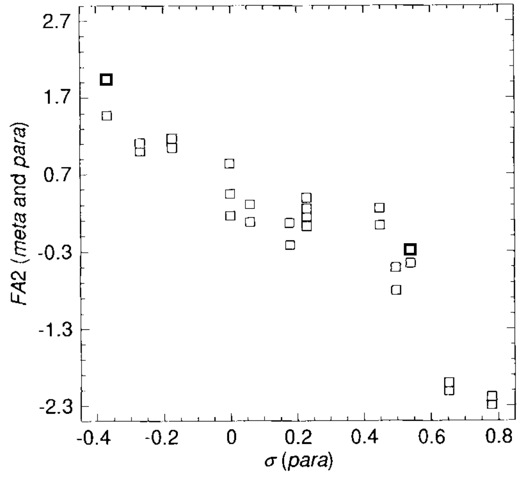


Figure 15. Plot of scores from the second factor ($FA2$) against σ.

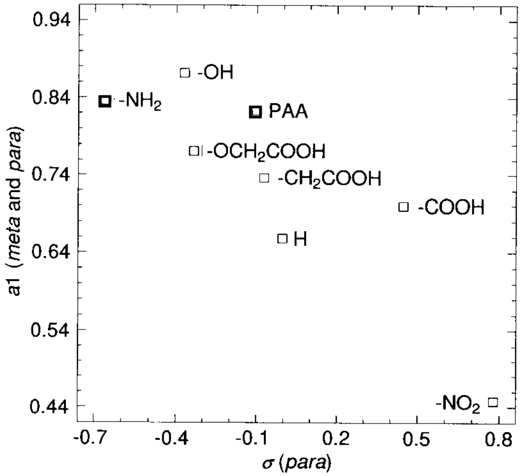


Figure 16. Plot of loadings from the first factor ($a1$) against σ of the functional groups. For abbreviations, see Table 6.

Eq. (37) shows that the first factor also contains a slight electronic correction in addition to basic hydrophobicity.

Figs. (16) and (17) show plots of the factor weights for factors 1 and 2 against the σ values of the functional groups Y (σ_p used). Obviously, the loadings are related to the electronic properties of the functional groups. The corresponding equations are:

$$a_1 = -0.26(\pm 0.14)\sigma_p(\text{Y}) + 0.72(\pm 0.06) \tag{39}$$

$$n = 8 \quad r = 0.868 \quad s = 0.067 \quad F = 18.3$$

$$a_2 = 0.24(\pm 0.14)\sigma_p(\text{Y}) + 0.64(\pm 0.05) \tag{40}$$

$$n = 8 \quad r = 0.871 \quad s = 0.063 \quad F = 18.9$$

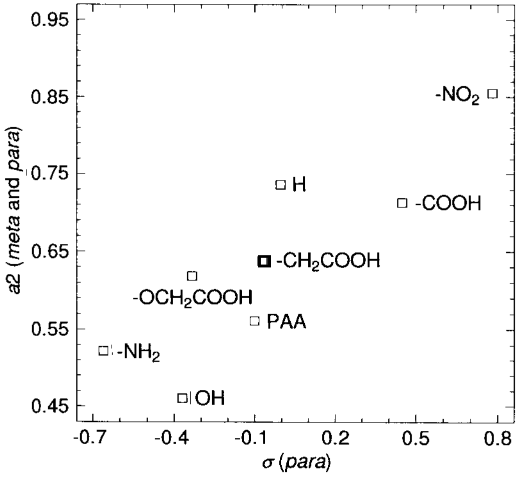


Figure 17. Plot of loadings from the second factor ($a2$) against σ of the functional groups. For abbreviations, see Table 6.

Table 9. Regression coefficients from Eq. (41) with $\pi_x(Y_j) = \sigma(BA)$

No.	Group	b_1	b_2
1	OCH_2COOH	1.03	0.17
2	CH_2COOH	0.99	0.00
3	H	1.05	−0.18
4	COOH	–	–
5	−OH	1.06	0.66
6	NH_2	1.38	0.49
7	PAA	1.07	0.42
8	NO_2	0.83	−0.50

If Eqs. (37) and (38) are back-translated into a factor model with two factors, it then follows that the following relationship must exist between π values from series of compounds with different functional groups:

$$\pi_x(Y_i) = b_0 + b_1\pi_x(Y_j) + b_2\sigma_x \tag{41}$$

In this equation $\pi_x(Y_i)$ and $\pi_x(Y_j)$ are π values for substituent X in an aromatic series with the functional groups, $Y = Y_i$ and $Y = Y_j$, respectively; σ_x is the σ value of the substituent, and b_0, b_1, and b_2 are regression coefficients. An example of Eq. (41) is the following relationship derived from the π values given in Table 6:

$$\pi(B) = 1.05(\pm 0.07)\pi(BA) - 0.18(\pm 0.15)\sigma - 0.12(\pm 0.06) \tag{42}$$
$$n = 27 \qquad r = 0.986 \qquad s = 0.105 \qquad F = 449.1$$

A further conclusion, which can be drawn from Eqs. (39) and (40), is that the coefficients b_1 and b_2 in Eq. (41) should be related to $\sigma(Y_i)$ (to $\sigma(Y_j)$), if equations with the same Y_j (the same Y_i) and different Y_is (different Y_js) are compared. The regression coefficients b_1 and b_2, obtained when relating π values of all the series of compounds considered to $\pi(BA)$ are summarized in Table 9. These coefficients do, indeed, correlate with σ of the respective functional groups according to:

$$b_1 = -0.31(\pm 0.22)\sigma_p + 1.02(\pm 0.09) \tag{43}$$
$$n = 7 \qquad r = 0.850 \qquad s = 0.087 \qquad F = 13.1$$

$$b_2 = -0.81(\pm 0.49)\sigma_p \tag{44}$$
$$n = 7 \qquad r = 0.857 \qquad s = 0.227 \qquad F = 16.6$$

The results compare very well with those obtained from principal component analysis (*meta*- and *para*-substituents treated separatel) [92, 93] and are in keeping with the bidirectional Hammett-type relationship, suggested by Fujita and coworkers [106, 107], to describe π values in disubstituted benzenes.

Relationships according to Eqs. (41) and (42) have been long known and have been determined empirically (see Franke [6, 7] and references cited therein). The present example demonstrates the capability of factor analysis as a "model generator": with no assumptions to begin with, a physically meaningful model has been obtained in an easy and straightforward way.

The Grouping of Substituents, Principal Properties, and Principal Component Regression Analysis

The appropriate selection of descriptors is essentially at the heart of QSAR work. However, this is not always an easy task, because of the great diversity of descriptors which are potentially available. In order to better understand the problems of similarity, redundancy, and information content, Van de Waterbeemd [108] performed a principal component analysis on 58 descriptors for 59 substituents, which was extended in a subsequent paper to include 74 substituents [109] (for earlier work, see e.g. Alunni et al. [110] and Tichy [111]). Five significant principal components explained 83.94% of the data variance as follows: $PC1 = 33.46\%$, $PC2 = 25.07\%$, $PC3 = 13.12\%$, $PC4 = 7.48\%$ and $PC5 = 4.82\%$. Loading plots showed a clustering of lipophilic, steric and electronic parameters. A number of less well-defined descriptors were found around the origin and random numbers were also situated there. Obviously, parameters situated close to the origin (considering all significant components) provide little information. The score plots revealed a very interesting picture: substituents were arranged in the order of increasing bulk from left to right and increasing polarity from top to bottom, forming five groups.

An even larger number of substituents, but with less parameters, was analyzed by Skagerberg et al. [112]. A hundred aromatic substituents were characterized by 9 variables (π, MR, σ_m, σ_p, and Verloop's STERIMOL parameters) and submitted to principal component analysis. Four components were extracted, accounting for 39%, 21%, 9%, and 7% of the data variance, respectively (total: 76%). The first component was related mainly to steric bulk and hydrophobicity, the second component represents electronic properties, the third component again is mainly hydrophobicity with a contribution for shape, and the last component is believed to have no real chemical interpretation and to be of minor importance, even though it is statistically significant as determined by cross-validation. The first three components were then used for selecting substituents for a training series by factorial design (see next section).

Principal components, derived from a set of descriptors or, to be more general, a set of property values for a given set of compounds, can be used as independent variables in QSAR analyses. If the components are derived from as large a (but still meaningful) selection of physical and chemical properties as possible, then they represent so-called principal properties (PPs), which can be very useful, especially for substances which are difficult to parameterize by classical QSAR descriptors (e.g. amino acids in peptides, sugars, etc.). The first contribution in this field was made by Cramer [98, 99, 113–115], who derived the so-called $BC(DEF)$ parameters, as principal components, from a data matrix containing six physico-chemical properties (activity coefficient in water, log P, MR, boiling point, MV, and heat of vaporization) for 114 structurally diverse chemical compounds. The first two components (B and C) already explain 95.7% of the data variance, while the proportion of variance explained by the subsequent components amounted to 2.8%, 0.7%, 0.5%, and 0.3%, respectively. The most important contribution to property description was, thus, contained in B and C, which were attributed to bulk and to bulk-corrected cohesiveness. The *BC(DEF)* parameters can be (roughly) estimated from chemical structure and have been shown to yield excellent descriptions and predictions of those physical properties of compounds, which are dominated by non-covalent interactions in a bulk fluid phase. In the case of non-specific biological effects, such as general anesthesia, nerve blockade, and erythrocyte stabilization, highly significant QSAR equations were also obtained. *BC(DEF)* certainly are interesting parameters for QSAR purposes, but have not found much practical application for mainly two reasons: (i) they are difficult to compute, and (ii) more specific interactions in a biological system would require that these global parameters are broken down into individual contributions for different parts of the molecules, which is not possible (or, at least, has not yet been attempted).

Much work has been done to derive PP's for amino acids with the objective of creating parameters for peptide QSARs, especially by Wold and coworkers [116–127] following the earlier studies of Sneath [128] and Kidera et al. [129] (for monosaccharides, see Eriksson et al. [130]). In the first study [116], a principal component analysis of 20 variables (molecular weight, pK_{COOH}, pK_{NH_2}, isoelectric point, van der Waals volume, 7 NMR measurements, and 8 parameters relating to hydrophobicity) for the 20 naturally occurring amino acids, yielded three components, which accounted for 58% of the data variance. Score plots revealed the relationships between the properties of the amino acids and the genetic code. The first component primarily reflected hydrophobicity, the second component reflected size, and the third component reflected electronic properties. The first application of these components, now referred to as z_1, z_2, and z_3, as variables in a peptide QSAR, has been reported by Hellberg et al. [117]. A PLS model, using these z-scales, was derived for bradykinin potentiating pentapeptides, which was shown to possess predictive capability. In later work, the z-scales were extended to include non-coded amino acids [119, 121, 123, 124]. New and extended z-scales, now designated as z'_1, z'_2, and z'_3, have now been derived for a total of 55 amino acids from a principal component analysis of the following variables: R_f values from 7 different TLC systems, van der Waals volume, molecular weight, and 3 NMR measurements. Further QSAR studies using z variables can be found in the literature [118, 120, 121, 125–127]. An alternative to the z-scales has been suggested by Norinder [131] who started from computed interaction energies of amino acids with 2016 grid points, taking into account non-bonded, charge-charge and hydrophobic interactions. Principal component analysis of the resulting data matrix resulted in 5 principal components, accounting for 81% of the data variance. When applied to a series of biologically active pentapeptides, a good PLS model was obtained. This work was repeated and extended by Cocchi and Johansson [132], who used interaction energies from 6 different probes and obtained 7 significant principal components (t-scales), accounting for 72% of the total data variance. A PLS modeling of the z-scales by the t-scales gave only poor results, which was attributed to the fact that the z-scales are based on experimental measurements of intact amino acids, while the t-scales are only representative of the side-chains. When applied to the ACE inhibitory potency of 58 dipeptides, however, both scales led to compatible results. It is also possible to mix the two scales.

Hemken and Lehmann [133] computed steric and electrostatic parameters (size, shape, and *MEP* properties at the van der Waals surface) by scanning a grid placed around aromatic substituents. Principal components derived from these parameters correlated well with conventional steric and electronic substituent constants and could even be used to replace the latter in a few QSAR examples. If such parameters are not available, dispensing with tabulated data could make this procedure attractive in certain instances. This is, of course, also true for the electronic, steric and hydrophobic parameters, derived by Kim and Martin [134 – 138] directly from 3D structures, using the CoMFA method [139] in a number of extensive studies. Again, part of this analysis involves principal component analysis, as implemented in the PLS part of CoMFA. Aromaticity scales derived by principal component analysis from other variables have been suggested as principal properties in QSAR work and series design for aromatic and heteroaromatic compounds [140 – 142].

An interesting application of principal component analysis, as an aid to identifying a pharmacophore in amnesia-reversal compounds, has been presented by Cosentino et al. [143]. The result led to the identification of three interatomic distances, which were able to provide all the information necessary to describe the relative spatial position of two key centers for interaction.

Principal components, extracted from a matrix of x variables (descriptor matrix, X), can be used as independent variables in a regression model (principal component regression analysis; PCRA). In this context, these principal components are called latent variables. The model of PCRA (e.g. as in [144]) is as follows,

$$y = b_0 + \sum_k b_k P_k(X) + \varepsilon \tag{45}$$

where y is a single biological potency (measurements from one biological test), $P_k(X)$ is the k-th principal component, derived from the descriptor matrix X, b_k are regression coefficients, obtained from correlating y as dependent variable with the $P_k(X)$. $P_\mathrm{k}(X)$ will then be independent variables, and ε is the residual term.

Moulton and Schultz [145] used principal component regression to investigated structure – activity relationships for inhibiting the growth of the ciliate *Tetrahymena pyriformis* by 20 *para*-substituted pyridines. They started form eight substituent constants, including π, two indicator variables to characterize H-acceptor and H-donor ability, MR, σ_p, Swain-Lupton's F and R constants, and the single bond fragment molecular connectivity index, ${}^1\chi^\mathrm{V}_\mathrm{sub}$, and extracted four principal components accounting for 95% of the variance. The first component expressed steric properties, the second component was related to hydrophobicity (including hydrogen bonding), the third component reflected resonance electronic effects, and, finally, the last component reflected electronic field effects. The first two components (after VARIMAX rotation) afforded a significant regression equation:

$$\log BR = 0.45P_1 - 0.25P_2 - 0.59 \tag{46}$$

$$n = 17 \qquad r = 0.831$$

The authors stated that the results were in good agreement with earlier investigations, performed with the original variables using multiple regression analysis. If this is true, then question arises as to why the investigation was repeated with principal components, which are more difficult to interpret. Turner et al. [146] considered the toxicity of two groups of metal ions in mice. With the variables "ionic radius", "ionization potential", "atomic weight", "William's softness parameter", and "electronegativity", two components were extracted for each group of metal ions, accounting for 96% and 94% of variance, respectively, which are related to toxicity. The statistical quality of the resulting equations, however, is only poor. A series of 32 8-azasteroids was investigated by Sokolov et al. [147]. Two principal components derived from electronic densities at various atoms led to significant regression equations for the hemolytic and cathepsin as well phosphatase inhibiting potency. Domine et al. [148] performed a principal component analysis on 5 physicochemical properties of 64 pesticides. Two components explained 91.7% of the total variance; a score plot showed an overall structure with respect to the membership of compounds in the families "herbicides" and "insecticides". Investigating the neurotoxicity of pyrethroids Ford and Livingstone [149] extracted 8 significant components from a large variety of molecular descriptors. Their use as independent variables in a subsequent regression step led to highly significant relationships. The following equation for the neurotoxicity was obtained:

$$NT = -0.72P_1 - 0.49P_2 + 1.37P_4 + 1.04P_6 - 0.51P_7 + 0.87 \qquad (47)$$
$$n = 19 \qquad r = 0.970 \qquad F = 21.4$$

Some additional applications of PCRA (see also [5]) are given in the literature [150–153].

In comparison with multiple regression analysis, PCRA has the advantage that collinearities between x variables are not a disturbing factor, and that the number of x variables included in the analysis, may exceed the number of observations. In comparison with automated stepwise regression procedures (only to be recommended in exceptional cases), the danger of chance correlations [154] is further reduced. However, the principal component approach solves the collinearity problem only from a purely mathematical point of view. Nothing has been gained from the perhaps even more important chemical point of view. If, for a given case, certain parameters, as for example, π and MR, are correlated, they will remain so also after principal component analysis has been performed. Thus, a conclusion whether steric or hydrophobic effects are operating, is still impossible. What is really necessary in such cases, is to introduce some carefully selected additional compounds in order to break the collinearity. In the case of PCRA and similar methods the danger exists that, in an uncritical way, too many variables will, inadvertently, be included in the principal component analysis step, which (within certain limits) may be acceptable mathematically, but which will render chemical interpretation increasingly difficult. Benigni and Giuliani [155] stated that "an analysis becomes fruitful when the correlation found can be explained within the context of physical-chemical and biological theories, or when it leads to formulating new hypotheses". This aspect is as important as the mathematical soundness and robustness of a result and must not be lost by overemphasizing the mathematical aspect alone. Only if, in particular cases, no reasonable assumptions about the selection of descriptors are possible, or if the compounds in question are difficult to parameterize in a straightforward way, then the principal component analysis step is very useful (e.g. use of z-scales for peptides). In such cases, however, PLS (see Chap. 4.4) is the preferred method over principal component regression.

The principal problem of QSAR is to understand which properties affect biological potency and why they do so. Even with the great number of x variables available, which may be included into PCRA or PLS, this problem is by no means solved, but is only shifted to a different level. If a single y variable is to be analyzed, multiple regression analysis (MRA) is not performed as automatic procedure, and descriptors are selected and screened according to the state-of-the-art methodology, then there is no reason to replace MRA by PLS or PCRA.

Series Design

A better alternative to PCRA in many cases is the application of principal components or factors for the preselection of parameters in a given QSAR problem or for the design of a training series.

A good training series should provide maximum information with a minimum of compounds. For this reason, the following conditions must be fulfilled:

1. All important properties must be varied over a sufficiently large range (sufficient variance in descriptor variables).
2. The parameter space must be covered systematically in order to avoid situations, such as is shown in Fig. 18. In such a situation there is too much redundancy in the two point clusters (one analog in each cluster would have provided the same information), and there is no way of determining what is occurring in the range between the two clusters. Moreover if, for some reason, it is decided that a straight line should be fitted to the data, the number of degrees of freedom will be overestimated, as the two point clouds can be regarded as two superpoints. Statistical tests would then provide a much higher level of significance than would be justified by the structure of the data.
3. Different properties (electronic, hydrophobic, and steric) must be varied independently of each other, since, otherwise, a mechanistic interpretation of later derived QSARs would become impossible (no collinearities between descriptor variables).

In order to fulfil these conditions simultaneously is no easy task and requires special mathematical methods of series design (see, e.g. references [6, 7, 156–159]; the advantage is a pronounced increase of information per compound synthesized (see

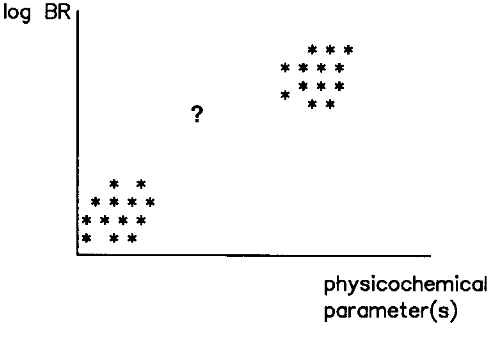


Figure 18. Series with low information content.

e.g. Martin [157], and Unger et al. [160]). Among a variety of different approaches [158], principal component analysis has also been applied for this purpose.

Franke and coworkers introduced the PCMM and the TMIC method [161 – 165] (see also [6, 7]). PCMM is a combination of **p**rincipal **c**omponent analysis and the **m**ultidimensional **m**apping technique of Wootton et al. [166]. The multidimensional mapping technique starts from a presentation of all possible substituents in parameter space. In a stepwise procedure a substituent, which is closest to the center of gravity of all hitherto selected points in parameter space but further apart than a predefined minimal Euclidean distance, D_{Min}, is selected in each step. In this way the variance of variables is maximized and a set of well-spread derivatives is selected. Multicollinearities, however, are not dealt with by this approach and are, thus, not necessarily eliminated. If collinearities exist, a hyperplane can be fitted to substituent points in parameter space and collinearities are mainly due to those points, which are close to the hyperplane. First substituents are divided into two sets, such that Set 1 contains the objects close to the hyperplane and Set 2 contains the substituents which are distant from the hyperplane, as judged by the Euclidean distances between substituents and hyperplane. These distances are computed with the help of principal component analysis. The multidimensional mapping technique of Wootton et al. is now applied separately to each of the two sets, in such a way that a higher percentage of substituents is selected from Set 2. Since the hyperplane changes its position during the selection procedure, this procedure is performed in a stepwise iterative manner, where the hyperplane position is adjusted after each step. This technique yields series with high data variance and minimized collinearities.

The TMIC method was devised for less than 50 substituents and is somewhat closer to the applications of principal component analysis, which have been discussed so far in this chapter. It is based on a score plot of substituents derived from an intraclass correlation matrix (two-dimensional **m**apping of **i**ntraclass **c**orrelation matrices). The intraclass correlation coefficient, r_{I}, is related to Euclidean distances and can be used to characterize the interrelatedness of two substituents X and Y with respect to m (standardized) descriptor variables, x_i ($i = 1, \ldots, m$):

$$r_{\mathrm{I(X,Y)}} = 2 \sum_i (x_{i\mathrm{Y}} = x_.)\,(x_{i\mathrm{X}} - x_.) / \sum_i (x_{i\mathrm{Y}} - x_.)^2 + (x_{i\mathrm{X}} - x_.)^2 \qquad (48)$$

with

$$x_. = (1/2m) \sum_i (x_{i\mathrm{Y}} + x_{i\mathrm{X}})$$

and

$$-1 \leq r_{\mathrm{I}} \leq 1$$

A series with low collinearities is characterized by low values of $|r_{\mathrm{I}}|$ for all possible pairs of substituents. In a score plot of principal components, derived from an intraclass correlation matrix, substituents interrelated with respect to the considered x_i will be close together (high positive values of $r_{\mathrm{I(X,Y)}}$), or in positions which are symmetrical with respect to the origin (high negative values of $r_{\mathrm{I(X,Y)}}$); substituents, which are placed close to the origin bear little information. If the first two components extract a sufficient amount of information, the plot is two-dimensional and a good training series with high x_i variances and low collinearities can now be obtained by

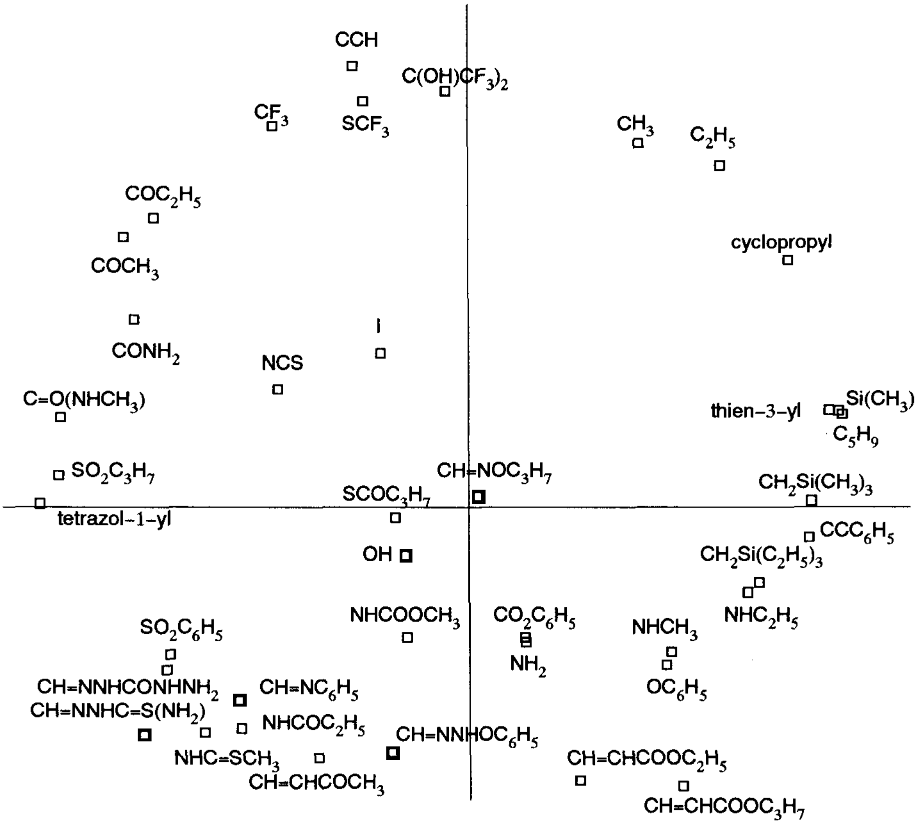


Figure 19. TMIC plot for 40 substituents (variables: π, F, R, and MR).

simple visual inspection. Substituents which are distant from each other, are selected in such a way that the whole space is systematically covered, while not including points which have a symmetrical position with respect to the origin. A TMIC map for 40 substituents (parameter space: π, F, R, MR) is shown in Fig. 19. A distinct clustering of substituents, similar with respect to the properties considered, becomes evident and is very reasonable from a chemical viewpoint.

In the last few years, 2^n factorial design with principal properties (see above) has frequently been used in series design [112, 120, 125, 126, 167–173] with the aim of performing PLS analyses on the data of the training series. Factorial design for the purpose of series selection in the QSAR field was introduced by Austel [174–178] using design variables based on substituent descriptors. This technique is discussed in more detail in Chap. 3.1. PPs are highly suitable for factorial design as they are independent, orthogonal and represent a reduction of dimensionality. Thus, this is the method of choice, if the data are to be analyzed by PLS. If, however, a multiple regression analysis (Hansch analysis) is planned, series designed in this way are not necessarily optimal. For example, the following substituents were selected by Skagerberg et al. [112] by a factorial design based on principal components, which was derived from a data table of 100 aromatic substituents (variables: π, MR, σ_m, σ_p, STERIMOL parameters): H, $-CH_3$, Br, $-NO_2$, $-C_6H_5$, $-OC_3H_7$, $-COC_6H_5$, $-CO_2C_6H_5$. For multiple regression analysis, this would not be a very good selection for the following reasons:

- Collinearities are not eliminated. There is a significant correlation between π and MR ($r = 0.735$) and a multicollinearity between π, MR, and σ (σ-term significant at 94%):

$$\pi = 0.058(\pm 0.020)MR - 0.88(\pm 0.93)\sigma_p \qquad (49)$$

$$n = 8 \qquad r = 0.942 \qquad s = 0.336 \qquad F = 23.8$$

- variances and ranges covered by, especially, σ_m and σ_p are not optimal.

Preselection of Variables for Regression Analysis

In practice, a drug designer is frequently confronted with data of series, which have not been designed according to the principles outlined in the previous section. If then a Hansch analysis is attempted, two problems have to be solved:

1. Selection of variables connected with biological potency from a, sometimes, very large pool of potential descriptors [108].
2. Investigation of (multiple) collinearities in order to understand which effects cannot be separated (important for interpretation), and in order to avoid (multi)collinear variables in the same equation (necessary for statistical reasons).

Factor analysis can serve as a data preprocessing step for both these objectives [6, 7, 82, 157, 179–182]. If a factor analysis is performed on a data matrix, containing the variable log BR (BR = biological response) and all descriptor variables which are to be considered, the resulting factor pattern (after VARIMAX rotation) will yield the following information:

1. Only those factors are connected with biological potency (variable log BR), where log BR has a loading, which is not equal to 0. The number of terms to be expected in a regression equation should, therefore, be equal to the number of factors with non-zero loadings for log BR.
2. Variables with a high loading in the same factor are interrelated (the higher the loadings the higher the correlation). Variables with non-zero loadings in different factors only are unrelated.
3. As follows from 1. and 2., only variables with non-zero loadings in those factors, where log BR also has non-zero loadings, are important for the description of log BR.
4. Another consequence of 2. is that only variables with non-zero loadings in different factors may be combined in one regression equation, in order to avoid collinearities.
5. The results of factor analysis indicate whether or not a satisfactory description of log BR can be achieved in the parameter space considered. If not, one can immediately choose a different variable space, thus preventing the calculation of useless regression equations.

One example is given by Franke [7], where data concerning the inhibition of the NADH oxidase system from ETP for 17 ring-substituted phenoxyacetic acids were analyzed. Factor analysis provided three significant factors for the variable space considered with non-zero loadings for the biological potency (pI_{50}) in factors one and three. The descriptor variables had non-zero loadings as follows:

Factor 1: σ, σ^2, σ^-, σ^{-2}, S
Factor 2: π, π^2, P
Factor 3: E_s, E_s^2.

It, thus, follows that biological potency depends on electronic and steric effects, while hydrophobicity has no significant role. In regression analysis, one of the electronic substituent constants from factor 1 should be combined with E_s (factor 3); there is no reason at this stage to preferably use squared variable terms. Such combinations do, indeed, lead to a satisfactory description as follows, for example, from Eq. (50):

$$pI_{50} = 0.75\sigma^- - 0.23E_s + 3.34 \qquad (50)$$
$$n = 17 \qquad r = 0.923 \qquad s = 0.249$$

The result was checked by screening all conceivable combinations of descriptor variables using regression analysis (a strategy, which is frequently employed in the Hansch analysis). In comparison with this strategy the use of FA as a preprocessing step, saves on more than 90% of computations made and gives a clear picture of the steps being undertaken.

The general strategy of applying factor analysis as a preprocessing step in regression analysis, is similar to that in principal component regression analysis (PCRA). As in PCRA, relationships of biological y variables and "factors" ("patterns") inherent in the x variables are investigated. The difference is, that in PCRA all descriptors are assumed to be important, while the aim of factor analysis is to find out the relevant descriptors. With FA as preprocessing step before regression analysis, some of the drawbacks of "latent variable" models (low interpretability) and of "pure" regression models (disturbing effects of collinearities) can be avoided. In addition, the probability of obtaining chance correlations is reduced. Unfortunately, the factor analysis approach, as described above, has been seldom used in QSAR work. A variation of factor analysis, which is more frequently applied (see literature for examples [149, 183–185]), is to subject only descriptor variables to FA and to then correlate biological activity with one highly loaded variable from each factor. In this way, collinearities are avoided, but one cannot ascertain how many terms the final regression equations should have, and which of the molecular parameters are connected with biological potency.

References

[1] Lewi, P. J., Computer Technology in Drug Design. In: Drug Design, Vol. **VII**, Ariens, E. J., ed., Academic Press, New York, 1976, p. 209–278

[2] Lewi, P. J., Multivariate Data Analysis in Structure-Activity Relationships. In: *Drug Design*, Vol. **X**, Ariens, E. J., ed., Academic Press, New York (1980) p. 307–342

[3] Lewi, P. J., *Multivariate Data Analysis in Industrial Practice*, Research Studies Press, Chichester, 1982

[4] Malinowski, E. R., and Howery, D. G., *Factor Analysis in Chemistry*, Wiley, New York, 1980

[5] Mager, P. P., *Multidimensional Pharmacochemistry: Design for Safer Drugs*, Academic Press, New York, 1983

[6] Franke, R., *Optimierungsmethoden in der Wirkstofforschung*, Akademie-Verlag, Berlin, 1980

[7] Franke, R., *Theoretical Drug Design Methods*. Elsevier, Amsterdam, 1984, Akademie-Verlag, Berlin, 1984

[8] Wold, S., and Sjöström, M., SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In: *Chemometrics: Theory and Applications* (ACS Symposium Series **52**), Kowalski, R., ed. (1977) 243–282

[9] Wold, S., Albano, C., Dunn III, W. J., Edlund, U., Esbensen, D., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., and Sjöström, M., Multivariate Data Analysis in Chemistry. In: *Chemometrics. Mathematics and Statistics in Chemistry*, Kowalski, B. R., ed., D. Reidel Publ., Dordrecht (1984) p. 17–91

[10] Wold, S., Esbensen, K., and Geladi, P., *Chemom. Intell. Lab. Syst.* **2**, 37–52 (1987)

[11] Livingstone, D. J., *Methods in Enzymology* **203**, 613–638 (1991)

[12] Escofier, B., and Pages, J., Presentation of Correspondence Analysis and Multiple Correspondence Analysis with the Help of Examples. In: *Applied Multivariate Analysis in SAR and Environmental Studies*. Devillers, J., and Karcher, W., eds., Kluwer, Academic Publishers, Dordrecht (1991) p. 1–32

[13] Lebreton, J. B., Sabatier, R., Banco, G., and Bancou, A. M., Principal component and Correspondent Analyses with Respect to Instrumental Variables: An Overview of Their Role in Studies of Structure-Activity and Species-Environment Relationships. In: *Applied Multivariate Analysis in SAR and Environmental Studies*. Devillers, J., and Karcher, W., eds., Kluwer Academic Publishers, Dordrecht (1991) p. 85–114

[14] Rummel, R. J., *Applied Factor Analysis*, Northwestern University Press, Evanston, 1970

[15] Harman, H. H., *Modern Factor Analysis*, 3rd Ed., University of Chicago Press, Chicago, 1967

[16] Joliffe, I. T., *Principal Component Analysis*, Springer-Verlag, New York, 1986

[17] Seal, H., *Multivariate Analysis for Biologists*, Methuen, London, 1968

[18] Dillon, W. R., and Goldstein, M., *Multivariate Analysis Methods & Applications*, Wiley, New York, 1984

[19] Glen, W. G., Dunn III, W. J., and Scott, D. R., *Tetrahedron Comput. Meth.* **2**, 349–376 (1989)

[20] Greenacre, M. J., *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984

[21] Anton, H., *Elementary Linear Algebra*, 2nd Ed., Wiley, New York, 1977

[22] Glen, G. W., Sarker, M., Dunn III, W. J., and Scott, D. R., *Tetrahedron Comput. Methodol.* **2**, 377–396 (1989)

[23] Lukovits, I., *J. Med. Chem.* **26**, 1104–1107 (1983)

[24] Livingstone, D. J., Evans, D. A., and Saunders, M. R., *J. Chem. Soc. Perkin Trans.* 1545–1550 (1992)

[25] Cattell, R. B., *Multivariate Behav. Res.* **1**, 245–276 (1966)

[26] Exner, O., *Coll. Czechoslov. Chem. Commun.* **37**, 3222–3228 (1966)

[27] Malinowski, E. R., *Anal. Chem.* **49**, 606–612 (1977)

[28] Malinowski, E. R., *Anal. Chem.* **49**, 612–618 (1977)

[29] Wold, S., *Technometrics* **20**, 379–406 (1978)

[30] Dove, S., Kühne, R., and Franke, R., *Pharmacochemistry Library* **8**, 277–292 (1985)

[31] Lukovits, I., and Lopata, A., *J. Med. Chem.* **23**, 449 (1980)

[32] Darvas, F., Meszaros, Z., Kovaszs, L., Hermecz, I., Balogh, M., and Kordos, J., *Arzneimittel-Forsch.* **29**, 1334–1340 (1979)

[33] Davies, R. H., and Morris, T. R., *Int. J. Quant. Chem.* **23**, 1385–1389 (1983)

[34] Coats, E. A., Cordes, H.-P., Kulkarni, V. M., Richter, M., Schaper, K.-J., Wiese, M., and Seydel, J. K., *Quant. Struct.-Act. Relat.* **4**, 99–109 (1985)

[35] Dove, S., *Zur Anwendung einiger multivariater Verfahren auf spezielle Probleme der quantitativen Struktur-Wirkungs-Analyse*, Ph. D. Thesis, Martin-Luther-Universität, Halle/Wittenberg, Halle, 1978

[36] Wiese, M., Seydel, J. K., Pieper, H., Krüger, G., Noll, K. R., and Keck, J., *Quant. Struct.-Act. Relat.* **6**, 164–172 (1987)

[37] Esaki, T., *Chem. Pharm. Bull.* **35**, 3105–3111 (1987)

[38] Seydel, J. K., Cordes, H.-P., Wiese, M., Chi, H., Croes, N., Hanpft, R., Lüllmann, H., Mohr, K., Patten, M., Padberg, Y., Lüllmann-Rauch, R., Vellguth, S., Meindl, W. R., and Schönenberger, H., *Quant. Struct.-Act. Relat.* **8**, 266–278 (1989)

[39] Leibovici, L., Wysenbeek, A. J., Konisberger, H., Samra, Z., Pitlik, S. D., and Drucker, M., *Eur. J. Clin. Microbiol. Infect. Dis.* **11**, 782–788 (1992)

[40] Baraldi, P. G., Brigidi, P., Casorali, A., Manfredini, S., Periotto, V., Recanatini, M., Roberti, M., and Rossi, M., *Arzneim. Forsch./Drug Res.* II **39**, 1406–1410 (1989)

[41] Codarin, M., Linda, P., Ebert, C., Lassiani, L., Rubessa, F., Alunni, S., Clementi, S., Sjöström, M., Wold, S., and Dunn III, W. J., Principal Component Analysis (PCA) of Substituent Effects in Biological Activity: Ortho-, Meta- and Para-Substituted Phenyls. In: *QSAR in Design of Bioactive Compounds*, Kuchar, M., ed., J. R. Prous Publishers, Barcelona (1984) p. 347–358

[42] Franke, R., Barth, A., Dove, S., and Laass, W., *Pharmazie* **35**, 181–182 (1980)

[43a] Szigeti, Z., Cserhati, T., and Bordas, B., *Gen. Physiol. Biophys.* **4**, 321–330 (1985)

[43b] Hecht, P., Vyplel, H., Nussbaumer, P., and Berner, H., *Quant. Struct.-Act. Relat.* **11**, 339–347 (1992)

[44] Ebert, C., Sassiani, L., Linda, P., Nisi, C., Alunni, S., and Clementi, S., *Quant. Struct.-Act. Relat.* **3**, 143–147 (1984)

[45] Moret, E. E., and Janssen, L. H. M., *Pharmacochemistry Library* **16**, 381–384 (1991)

[46] Benigni, R., and Giuliani, A., *Mutat. Res.* **205**, 227–236 (1988)

[47] Benigni, R., and Giuliani, A., Multivariate Analysis in Genetic Toxicology. In: *Applied Multivariate Analysis in SAR and Environmental Studies*, Devillers, J., and Karcher, W., eds., Kluwer, Dordrecht (1971) p. 347–376

[48] Benigni, R., and Giuliani, A., *Environ. Health Perspectives* **96**, 81–84 (1991)

[49] Benigni, R., *Mutagenesis* **7**, 335–341 (1992)

[50] Nendza, M., and Seydel, J. K., *Chemosphere* **17**, 1575–1584 (1988)

[51] Nendza, M., and Seydel, J. K., *Quant. Struct.-Act. Relat.* **7**, 165–174 (1988)

[52] Eriksson, L., Sandström, B. E., Sjöström, M., Tysklind, M., and Wold, S., *Quant. Struct.-Act. Relat.* **12**, 124–131 (1993)

[53] Kaiser, K. L. E., and Esterby, S. R., *Sci. Total Environm.* **109/110**, 499–514 (1991)

[54] Onfelt, A., Hellberg, S., and Wold, S., *Mutat. Res.* **174**, 109–113 (1986)

[55] Dove, S., Coats, E. A., Scharfenberg, P., and Franke, R., *J. Med. Chem.* **28**, 447–451 (1985)

[56] Weiner, M. L., and Weiner, P. H., *J. Med. Chem.* **16**, 655–660 (1973)

[57] Franke, R., Schwarz, G., Dove, S., Dietrich, A., and Klinger, W., *Zbl. Pharm.* **122**, 459–464 (1983)

[58] Schulz, E., Sprung, W.-D., Kröning, G., and Vietinghof, G., *Pharmazie* **33**, 749–753 (1978)

[59] Schulz, E., Sprung, W. D., Kröning, G., and Lange, P., *Agents Act. Suppl.* **10**, 119–124 (1982)

[60] Schulz, E., and Sprung, W. D., personal communication

[61] Seydel, J. K., Cordes, H.-P., Wiese, M., Chi, H., Croes, N., Hanpft, R., Lüllmann, H., Mohr, K., Patten, M., Padberg, Y., Lüllmann-Rauch, R., Vellguth, S., Meindl, W. R., and Schönenberger, H., *Quant. Struct.-Act. Relat.* **8**, 266–278 (1989)

[62] Cserhati, T., and Magyar, K. J., *Pharmac. Biomed. Anal.* **10**, 1033–1039 (1992)

[63] Bouzard, D., Di Cesare, P., Jacquet, J. P., Ledoussal, B., Remuzon, P., Kessler, R. E., and Fung-Tome, J., *J. Med. Chem.* **35**, 518–525 (1992)

[63a] Franke, R., Gruska, A., and Presber, W., *Pharmazie* **49**, 600–605 (1994)

[64] *STATGRAPHICS 5.0*, STSC Inc., Rockville, MD, USA

[65] Free, S. M., and Wilson, J. W., *J. Med. Chem.* **7**, 395–399 (1964)

[66] Fujita, T., and Ban, T., *J. Med. Chem.* **14**, 148–152 (1971)

[67] Dove, S., Franke, R., and Oehme, P., QSAR in DBH-Blocking Hypotensive Fusaric Acid Derivatives. In: *QSAR in Design of Bioactive compounds*. Kuchar, M., ed., J. R. Prous Publishers, Barcelona (1984) p. 117–134, 449–450

[68] Oehlke, J., Schrötter, E., Piesche, L., Dove, S., Schick, H., and Niedrich, H., *Pharmazie* **38**, 624–630 (1983)

[69] Piesche, L., Hilse, H., Oehlke, J., Schrötter, E., and Oehme, P., *Pharmazie* **38**, 335–338 (1983)

[70] Roux, M., Basic Procedures in Hierarchical Cluster Analysis. In: *Applied Multivariate Analysis in SAR and Environmental Studies*, Devillers, J., and Karcher, W., eds., Kluwer Academic Publishers, Dordrecht (1991) p. 115–135

[71] Grossy, G., Tenlade, J.-C., Chapat, J. P., Simeon de Buochberg, M., and Attisso, M., *Eur. J. Med. Chem.* **17**, 109–115 (1982)

[72] Grossy, G., Rival, Y., Tenlade, J.-C., Chapet, J.-P., Simeon de Buochberg, M., and Attisso, M., *Eur. J. Med. Chem.* **20**, 199–204 (1985)

[73] Bawden, D., Gymer, G. E., Marriott, M. S., and Tute, M. S., *Eur. J. Med. Chem.* **18**, 91–96 (1983)

[74] Chen, B.-K., Horvath C., and Bertino, J. R., *J. Med. Chem.* **22**, 483–486 (1979)

[75] Rose, V. S., Wood, J., and MacFie, H. J. H., *Quant. Struct.-Act. Relat.* **10**, 359–368 (1991)

[76] Cammarata, A., and Menon, G. K., *J. Med. Chem.* **19**, 739–747 (1976)

[77] Menon, G. K., and Cammarata, A., *J. Pharm. Sci.* **66**, 304–314 (1977)

[78] Schaper, K.-J., and Seydel, J. K., Multivariate Methods in Quantitative Structure-Pharmacokinetics Relationship Analysis. In: *QSAR and Strategies in the Design of Bioactive Compounds*, Seydel, J. K., ed., VCH, Weinheim (1985) p. 173–189

[79] Szydlo, R. M., Ford, M. G., Greenwood, R., and Salt, D. W., The Use of Multivariate Data Sets in the Study of Structure-Activity Relationships of Synthetic Pyrethroid Insecticides: The relationship between physicochemical and pharmacokinetic properties, Part I. In: *QSAR and Strategies in the Design of Bioactive Compounds*, Seydel, J. K., ed., VCH, Weinheim (1985) p. 219–228

[80] Szydlo, R. M., Ford, M. G., Greenwood, R., and Salt, D. W., The Use of Multivariate Data Sets in the Study of Structure-Activity Relationships of Synthetic Pyrethroid Insecticides: The relationship between pharmacokinetics and toxicity, Part II. In: *QSAR and Strategies in the Design of Bioactive Compounds*, Seydel, J. K., ed., VCH, Weinheim (1985) p. 229–237

[81] Dove, S., and Franke, R., *Wiss. Beiträge Martin-Luther Univ. Halle-Wittenberg* **42(R34)**, 216–222 (1977)

[82] Franke, R., *Il Farmaco – Ed. Sc.* **34**, 545–570 (1979)

[83] Bekemeyer, H., private communication

[84] Kubinyi, H., *Arzneim.-Forsch.* **26**, 1991–1999 (1976)

[85] Penniston, J. T., Beckett, L., Bentley, D. L., and Hansch, C., *Mol. Pharmacol.* **5**, 333–339 (1969)

[86] van Bever, W. F. N., Niemegeers, C. J. E., Schellekens, K. H. L., and Janssen, P. A. J., *Drug Res.* **26**, 1548–1551 (1976)

[87] Balaz, S., Wiese, M., and Seydel, J. K., *J. Pharm. Sci.* **81**, 849–857 (1992)

[87a] Franke, R., and Gruska, A., *Quant. Struct.-Act. Relat.* **13**, 148–151 (1994)

[88] Wold, S., and Sjöström, M., *Chemica Scripta* **2**, 49–55 (1972)

[89] Sjöström, M., and Wold, S., *Chemica Scripta* **6**, 114–120 (1974)

[90] Sjöström, M., and Wold, S., *Chemica Scripta* **9**, 200–207 (1976)

[91] Sjöström, M., and Wold, S., *Acta Chem. Scand.* **B 35**, 537–546 (1981)

[92] Franke, R., *Pharmacochem. Library* **2**, 251–267 (1977)

[93] Franke, R., Dove, S., and Kühne, R., *Eur. J. Med. Chem.* **14**, 363–374 (1979)

[94] Dove, S., Franke, R., and Kühne, R., Principal Component Analysis of Partition Coefficients in Different Solvent Systems. In: *Chemical Structure-Biological Activity Relationships*. Knoll, J., and Darvas, F., eds., Akademiai Kiado, Budapest (1980) p. 247–253

[95] Franke, R., Kühne, R., and Dove, S., *Pharmacochem. Library* **6**, 15–32 (1983)

[96] Hermann, R. B., *J. Phys. Chem.* **76**, 2754–2759 (1972)

[97] Hermann, R. B., *Proc. Natl. Acad. Sci. USA* **74**, 4144–4145 (1974)

[98] Cramer III, R. D., *J. Am. Chem. Soc.* **102**, 1837–1849 (1980)

[99] Cramer III, R. D., *J. Am. Chem. Soc.* **102**, 1849–1859 (1980)

[100] Seiler, P., *Eur. J. Med. Chem.* **9**, 473–479 (1974)

[101] Dunn III, W. J., and Wold, S., *Acta Chem. Scand.* **B 32**, 536–542 (1978)

[102] Dunn III, W. J., *Quant. Struct.-Act. Relat.* **2**, 156–163 (1983)

[103] Dunn III, W. J., Koehler, M. G., and Grigoras, S., *J. Med. Chem.* **30**, 1121–1126 (1987)

[104] Koehler, M. G., Grigoras, S., and Dunn III, W. J., *Quant. Struct.-Act. Relat.* **7**, 150–159 (1988)

[105] Dunn III, W. J., Nagy, P. I., Collantes, E. R., Glen, W. G., Alagona, G., and Ghio, C., *Pharmacochem. Library* **16**, 59–65 (1991)

[106] Fujita, T., *Progr. Phys. Org. Chem.* **14**, 75–113 (1983)

[107] Yamagami, C., Takao, N., and Fujita, T., *Quant. Struct.-Act. Relat.* **9**, 33–320 (1990)

[108] van de Waterbeemd, H., and Testa, B., *Adv. Drug Res.* **16**, 87–225 (1987)

[109] van de Waterbeemd, H., El Tayar, N., Carrupt, P.-A., and Testa, B., *J. Computer-Aided Mol. Design* **3**, 111–132 (1989)

[110] Alunni, S., Clementi, S., Edlund, U., Johnels, B., Hellberg, S., Sjöström, S., and Wold, S., *Acta Chem. Scand.* **B37**, 47–53 (1983)

[111] Tichy, M., *Int. J. Quant. Chem.* **16**, 509–515 (1979)

[112] Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G., and Ebert, C., *Quant. Struct.-Act. Relat.* **8**, 32–38 (1989)

[113] Cramer III, R. D., *Quant. Struct.-Act. Relat.* **2**, 7–12 (1983)

[114] Cramer III, R. D., *Quant. Struct.-Act. Relat.* **2**, 13–19 (1983)

[115] Yunger, L. M., and Cramer III, R. D., *Quant. Struct.-Act. Relat.* **2**, 149–156 (1983)

[116] Sjöström, M., and Wold, S., *J. Mol. Evol.* **22**, 272–277 (1985)

[117] Hellberg, S., Sjöström, M., and Wold, S., *Acta Chem. Scand.* **40B**, 135–140 (1986)

[118] Hellberg, S., Sjöström, M., Skagerberg, B., and Wold, S., *J. Med. Chem.* **30**, 1126–1135 (1987)

[119] Skagerberg, B., Sjöström, M., and Wold, S., *Quant. Struct.-Act. Relat.* **6**, 158–164 (1987)

[120] Hellberg, S., Sjöström, M., Skagerberg, B., Wilkström, C., and Wold, S., *Acta Pharm. Jugosl.* **37**, 53–65 (1987)

[121] Wold, S., Eriksson, L., Hellberg, S., Jonsson, J., Sjöström, M., Skagerberg, B., and Wikström, C., *Can. J. Chem.* **65**, 1814–1819 (1986)

[122] Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S., *Quant. Struct.-Act. Relat.* **7**, 144–150 (1988)

[123] Jonsson, J., Eriksson, L., Hellberg, S., Sjöström, M., and Wold, S., *Quant. Struct.-Act. Relat.* **8**, 204–209 (1989)

[124] Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S., *Progr. Clin. Biol. Res.* **291**, 131–134 (1989)

[125] Sjöström, M., Eriksson, L., Hellberg, S., Jonsson, J., Skagerberg, B., and Wold, S., *Progr. Clin. Biol. Res.* **291**, 313–317 (1989)

[126] Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Skagerberg, B., Sjöström, M., and Wold, S., *Acta Chem. Scand.* **44**, 50–56 (1990)

[127] Hellberg, S., and Kem, W., *Int. J. Peptide Protein Res.* **36**, 440–444 (1990)

[128] Sneath, P. H., *J. Theor. Biol.* **12**, 157–195 (1966)

[129] Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A., *J. Protein Chem.* **4**, 23–55 (1985)

[130] Eriksson, L., Jonsson, J., Sjöström, M., Wikström, C., and Wold, S., *Acta Chem. Scand.* **42B**, 504–514 (1988)

[131] Norinder, U., *Peptides* **12**, 1223–1237 (1991)

[132] Cocchi, M., and Johansson, E., *Quant. Struct.-Act. Relat.* **12**, 1–8 (1993)

[133] Hemken, H. G., and Lehmann, P. A., *Quant. Struct.-Act. Relat.* **11**, 332–338 (1992)

[134] Kim, K. H., and Martin, Y. C., *J. Org. Chem.* **56**, 2723–2729 (1991)

[135] Kim, K. H., and Martin, Y. C., *J. Med. Chem.* **34**, 2056–2060 (1991)

[136] Kim, K. H., *Quant. Struct.-Act. Relat.* **11**, 127–134 (1992)

[137] Kim, K. H., *Quant. Struct.-Act. Relat.* **11**, 453–460 (1992)

[138] Kim, K. H., *Quant. Struct.-Act. Relat.* **12**, 232–238 (1993)

[139] Cramer III, R. D., Patterson, D. E., and Bunce, J. D., *J. Am. Chem. Soc.* **110**, 5959–5967 (1988)

[140] Ebert, C., Katritzky, A. R., and Musumarra, G., *Quant. Struct.-Act. Relat.* **10**, 101–106 (1991)

[141] Ebert, C., Katritzky, A. R., and Musumarra, G., *Quant. Struct.-Act. Relat.* **10**, 107–109 (1991)

[142] Caruso, L., Musumarra, G., and Katritzky, A. R., *Quant. Struct.-Act. Relat.* **12**, 146–151 (1993)

[143] Cosentino, U., Moro, G., Pitea, D., Todeschini, R., Brossa, S., Gualandi, F., Scolastico, C., and Diannessi, F., *Quant. Struct.-Act. Relat.* **9**, 195–201 (1990)

[144] Draper, N. R., and Smith, H., *Applied Regression Analysis*, Wiley, New York, 1981

[145] Moulton, M. P., and Schultz, T. W., *Chemosphere* **15**, 59–67 (1986)

[146] Turner, J. E., Williams, M. W., Hingerty, B. E., and Hayden, T. L., Multiparameter Correlations Between Properties of Metal Ions and their Acute Toxicity in Mice. In: *QSAR in Environmental Toxicology – II.*, Kaiser, K. L. E., ed., D. Reidel Publishers, Dordrecht (1987) p. 375–383

[147] Sokolov, Y. A., Golubovich, V. P., Gurskii, S. G., and Ahrem, A. A., *Dokl. Akad. Nauk BSSR* **31**, 662–664 (1987)

[148] Domine, D., Devillers, J., Chastrette, M., and Karcher, W., *Pestic. Sci.* **35**, 73–82 (1992)

[149] Ford, M. G., and Livingstone, D. J., *Quant. Struct.-Act. Relat.* **9**, 107–114 (1990)

[150] Mager, P. P., *Tox. Letters* **11**, 67–71 (1982)

[151] Lukovits, I., *J. Med. Chem.* **26**, 1104–1107 (1983)

[152] Kubota, T., Hanamura, J., Kano, K., and Uno, B., *Chem. Pharm. Bull.* **33**, 1488–1492 (1985)

[153] Mardia, K. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis*, Academic Press, New York, 1979

[154] Topliss, J. G., and Edwards, R. P., *J. Med. Chem.* **22**, 1238–1242 (1979)

[155] Benigni, R., and Giuliani, A., *Quant. Struct.-Act. Relat.* **10**, 99–100 (1991)

[156] Hansch, C., and Leo, A., *Substituent Constants for Correlation Analysis in Chemistry and Biology*, Wiley, New York, 1979

[157] Martin, Y. C., and Panas, H. N., *J. Med. Chem.* **22**, 784–792 (1979)

[158] Unger, S. H., Consequences of the Hansch Paradigm for the Pharmaceutical Industry. In: *Drug Design*, Vol. **IX**, Ariens, E. J., ed., Academic Press, New York (1980) p. 48–119

[159] Pleiss, M. A., and Unger, S. H., The Design of Test Series and the Significance of Structure-Activity Relationships. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Hansch, C., Sammes, P. G., Taylor, J. B., Emmett, J. C., Kennewell, P. D., and Ramsden, C. A., eds., Pergamon Press, Oxford (1990) p. 561–587

[160] Martin, Y. C., Kim, K.-H., and Bures, M. G., Computer-Assisted Drug Design in the 21st Century. In: *Medicinal Chemistry of the 21st Century*, Wermuth, C. G., Koga, N., König, H., and Metcalf, B. W., eds., Blackwell Scientific Publications (1992) p. 295–315

[161] Franke, R., *Med. Chem.* **6**, 237–245 (1979)

[162] Franke, R., Streich, W. J., and Dove, S., Sample Selection Methods. In: *Chemical Structure-Biological Activity Relationships*, Knoll, J., and Darvas, F., eds., Akademiai Kiado, Budapest (1980) p. 153–163

[163] Streich, W. J., Dove, S., and Franke, R., *J. Med. Chem.* **23**, 1452–1455 (1980)

[164] Dove, S., Streich, W. J., and Franke, R., *J. Med. Chem.* **23**, 1456–1459 (1980)

[165] Krause, G., Klepel, M., and Franke, R., *Pharmacochem. Library* **6**, 233–234 (1983)

[166] Wootton, R., Cranfield, R., Sheppey, G. C., and Goodford, P. J., *J. Med. Chem.* **18**, 607–612 (1975)

[167] Jonsson, J., Eriksson, L., Sjöström, M., Wold, S., and Tosato, M. L., *Chemometrics Intelligent Lab. Syst.* **5**, 169–186 (1988)

[168] Skagerberg, B., Clementi, S., Sjöström, M., Tosato, M. L., and Wold, S., *Progr. Clin. Biol. Res.* **291**, 127–130 (1989)

[169] Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S., *Chemometrics Intelligent Lab. Syst.* **7**, 131–141 (1989)

[170] Tosato, M. L., Marchini, S., Passerini, L., Pino, A., Eriksson, L., Lindgren, F., Hellberg, S., Jonsson, J., Sjöström, M., Skagerberg, B., and Wold, S., *Environ. Toxicol. Chem.* **9**, 265–277 (1990)

[171] Eriksson, L., Jonsson, J., Hellberg, S., Lindgren, F., Skagerberg, B., Sjöström, M., Wold, S., and Berglind, R., *Environ. Toxicol. Chem.* **9**, 1339–1351 (1990)

[172] Bonelli, D., Cechetti, V., Clementi, S., Cruciani, G., Fravolini, A., and Savino, A. F., *Quant. Struct.-Act. Relat.* **10**, 333–343 (1991)

[173] Norinder, U., *J. Appl. Toxicol.* **12**, 143–147 (1992)

[174] Austel, V., *Eur. J. Med. Chem.* **17**, 9–15 (1982)

[175] Austel, V., *Eur. J. Med. Chem.* **17**, 339–342 (1982)

[176] Austel, V., *Quant. Struct.-Act. Relat.* **2**, 59–63 (1983)

[177] Austel, V., *Pharmacochem. Library* **6**, 223–229 (1983)

[178] Austel, V., Manual Design of Test Series for Free-Wilson Analysis. In: *QSAR and Strategies in the Design of Bioactive Compounds*. Seydel, J. K., ed. (VCH) Weinheim, 1985, p. 247–250

[179] Martin, Y. C., *Drug Design Methods: A Critical Introduction*, Marcel Dekker, New York, 1978

[180] Franke, R., *Pharmacochem. Library* **4**, 355–377 (1982)

[181] Laass, W., Eichler, G., Dove, S., Vogt, W.-E., Franke, R., and Vahle, H., Erste Erfahrungen mit der Faktoranalyse zur Auswahl von Variablen für Regressionsansätze in der quantitativen Struktur-Wirkungs-Analyse. In: *Quantitative Structure-Activity Analysis*. Franke, R., and Oehme, P., eds., Akademie-Verlag, Berlin (1978) p. 267–271

[182] Szydlo, R. M., Ford, M. G., Greenwood, R., and Salt, D. W., The use of multivariate Techniques for the Prediction of Biological Activity. In: *QSAR in Design of Bioactive Compounds*, Kuchar, M., ed., J. R. Prous International Publishers, Barcelona (1984) p. 301–320

[183] Livingstone, D. J., *Pestic. Sci.* **27**, 287–304 (1990)

[184] Henrie II, R. N., Plummer, M. J., Smith, S. E., Yeager, W. H., and Witkowski, D. A., *Quant. Struct.-Act. Relat.* **12**, 27–37 (1993)

[185] Miyashita, Y., Takahashi, Y., Daiba, S., Abe, H., and Sasaki, S., *Anal. Chim. Acta* **143**, 35–42 (1982)

[186] van der Waterbeemd, H., ed., *Advanced Computer-Assisted Techniques in Drug Discovery*, Methods and Principles in Medicinal Chemistry, Vol. 3, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1995

4.2 Graphical Analysis as an Aid in Medicinal Chemistry

James Devillers and Daniel Chessel

"The greatest value of a picture is when it forces us to notice what we never expected to see"

J. W. Tukey [1]

Abbreviations and Symbols

CFA	Correspondence factor analysis
log ***P***	*n*-octanol/water partition coefficient
MEM	Minimum essential medium
PBS	Phosphate buffered saline
PCA	Principal components analysis
*PC*s	Principal components
PLS	Partial least squares
QSAR	Quantitative structure-activity relationship
SAR	Structure-activity relationship

4.2.1 Introduction

Despite the fact that graphical techniques in connection with official statistics can be traced back more than two centuries [2, 3], until the mid-1970's, routine large-scale use of graphs in data analysis was not feasible since computer graphics facilities were not available at a reasonable cost. Since this period, graphs have provided very powerful tools, both for analyzing scientific data and for communicating qualitative and quantitative information [4, 5]. Indeed, graphical methods enable the data analyst to explore data thoroughly, to look for patterns and relationships, to confirm or disprove hypotheses, to discover new phenomena, to serve as a mnemonic device for remembering major conclusions, and to communicate these conclusions to others [6, 7]. Therefore, in most cases, graphs enhance the different numerical methods used in classical data analysis. This fact can be easily illustrated in medicinal chemistry, where regression models are widely used to describe how a response variable (i.e. biological activity) is related to, or can be explained by one or more explanatory variables (i.e. physico-chemical descriptors or topological indices). Indeed, it is now well accepted that in regression analysis, graphs provide powerful diagnostic tools for conveying properties of fitted regressions, assessing the adequacy of the fits, detecting outliers, and suggesting improvements [6, 8, 9]. Conversely, even if chemometric methods are now well established in medicinal chemistry for the reduction of the dimensionality of data matrices or for classification problems, the usefulness of graphical methods for optimizing the use of these approaches is rarely emphasized [10]. Under these circumstances, the scope of this paper is first to briefly review some of the basic principles of graphics, and then to illustrate them from a case study, dealing with the co-inertia analysis [11, 12], which can be particularly suitable in medicinal chemistry to detect the co-structure between two data tables (e.g. biological activities and molecular descriptors).

4.2.2 Graphical Displays

In this paragraph, our intention is to formulate some of the basic principles, which allow graphs to be employed more incisively in medicinal chemistry. Indeed, the study of the theory of data graphics is beyond the scope of this paper and can be found in numerous reference textbooks [e.g. 2, 13, 14].

4.2.2.1 Overall Strategy

When a graph is drawn up, quantitative and categorical information is encoded chiefly through the combined use of points, lines, numbers, symbols, words, scales, and/or colors. This graph should deliver true messages without artifacts linked to the display technique itself (e.g. the Rorschach effect [6]). It should be able to reveal the data at several levels of detail with precision, lack of distortion, compactness and comprehensiveness [15]. Lastly, it is important to note that the most valuable graphical approaches are flexible enough to be applied to a wide variety of data [6].

4.2.2.2 Techniques in Graphical Design

Elementary Principles

Numerous publications deal with techniques of plot constructions [e.g. 2, 13, 14]. On a more basic level, some elementary, but important suggestions for the design of efficient graphs have to be followed. The amount of uninformative detail (e.g. the logo of a laboratory) and clutter (e.g. grid lines) in a plot must be minimized. Conversely, explanations, which highlight the richness of the data, must be encouraged, since they make graphical displays more attractive. Thus, it is always useful to write short messages on the plot to explain the data, characterize the outliers or some interesting data points, write QSAR equations and/or display molecular formulae on the graph itself, and to integrate the caption and legend into the design, so that the reader does not have to dart back and forth between the text and the graph. However, the combined use of words and graphs requires the adoption of some typographical conventions. Thus, for example, words used in a graph must not be abbreviated and elaborate coding avoided. Due to the usual reading direction in western languages, words must run from left to right. The typeface must be clear and precise, using upper-case and lower-case letters with serifs [2].

Proportions and Scales

A common task in constructing graphs is rescaling the data. This fundamental step has been widely discussed by Tukey [11]. Thus, for example, it is generally well accepted that choosing scales to reduce curved configurations of points in a graph is highly recommended. However, we have to mention that the choice of a scaling procedure basically depends on the type of data to be represented and the aim of the study. Tufte [2] emphasized that graphs should tend toward the horizontal rather than the vertical and mentioned rules of thumb (p. 189) for drawing up ideal rectangles for graphical purposes. One of them these rectangles, the Golden Rectangle, finds its origin five centuries B.C. and has a length/width ratio of 1.618 (i.e. $(1 + \sqrt{5})/2$).

Display of Supplementary Information

The display of supplementary information on a graph is crucial for successful interpretation. It is also a convenient tool for easily communicating results and for adding a new dimension to the graph. To reach this goal, we can, for example:

- replace a point on a graph by a word, a symbol or a shape encoding qualitative and/or quantitative information [16–18],
- use lines of different weights [2] or various lengths, which emanate from a point in different directions [3],
- employ isometric plots [3],
- add colors (especially blue which can be distinguished from other colors by most color-deficient people who represent 5 to 10% of the population [2]),
- use stereographic and cinematographic techniques [3].

Numerous illustrative examples have been given by Tufte [2], Bertin [14], and Gnanadesikan [19].

4.2.2.3 Visual Perception

We can consider a graphical method to be successful, only if it can be effectively decoded visually [5]. Even if our eye-brain system is a particularly sophisticated device, the study of how we perceive graphs shows that there are some limitations in our perception of graphical displays. Thus, for example, our visual system is able to perceive:

- angles more easily than slopes and straight lines more clearly than curved lines,
- simple patterns more easily than complex ones,

- large or dark objects (or clusters) with greater impact than small, light, or isolated ones,
- a $\log_2$ scale with less difficulty than a $\log_{10}$ scale [5, 6, 20].

It is obvious that these aspects must be taken into account when constructing a graph.

4.2.2.4 Software Availability

Recent advances in computer graphics technology have now made graphical software available on many microcomputers and workstations. Besides 2D and 3D systems, which are only available for multivariate exploratory graphics [21], some statistical packages now contain numerous multivariate analyses and sophisticated graphical tools to facilitate presentation and interpretation of data [22–24]. Thus, for example, the statistical analyses and graphical displays presented in this paper have been undertaken with ADE [22] running on a Macintosh®.

4.2.3 The Key Role of Graphics in Co-Structure Analysis

4.2.3.1 Background

Expressed in statistical terms, SAR or QSAR studies consist in finding qualitative or quantitative relationships between two data tables, the former being constituted of the biological activities (one or more columns) and the latter of the molecular descriptors (i.e. physico-chemical properties and/or topological indices). If the partial least squares (PLS) regression method [25] can be considered as the method of choice to obtain quantitative models, the co-inertia analysis [11, 12] appears to be the most suitable for emphasizing qualitative information from graphical displays.

The co-inertia analysis allows the determination of the co-structure between two data tables [11, 12]. The mathematical presentation of the co-inertia analysis is beyond the scope of this paper (for more details see Refs. 11 and 12), but is summarized in Fig. 1 using the classical formalism related to SAR and QSAR studies. In this context, a co-inertia analysis consists of the separate and matched analyses of a matrix of biological activities (Y) and a matrix of molecular descriptors (X). It can be viewed as a general method allowing to relate any kind of data set, using any standard multivariate analysis (e.g. PCA, CFA). Thus, for example, it is interesting to note that the Tucker's inter-battery analysis [26] is actually the co-inertia analysis of two standardized PCA. The canonical analysis on categorical variables [27] is actually the co-inertia analysis of two multiple correspondence analyses. The PLS method [25] consists mainly of using the axes derived from co-inertia analysis in a regression analysis procedure, in order to obtain QSAR models for predictive purposes.

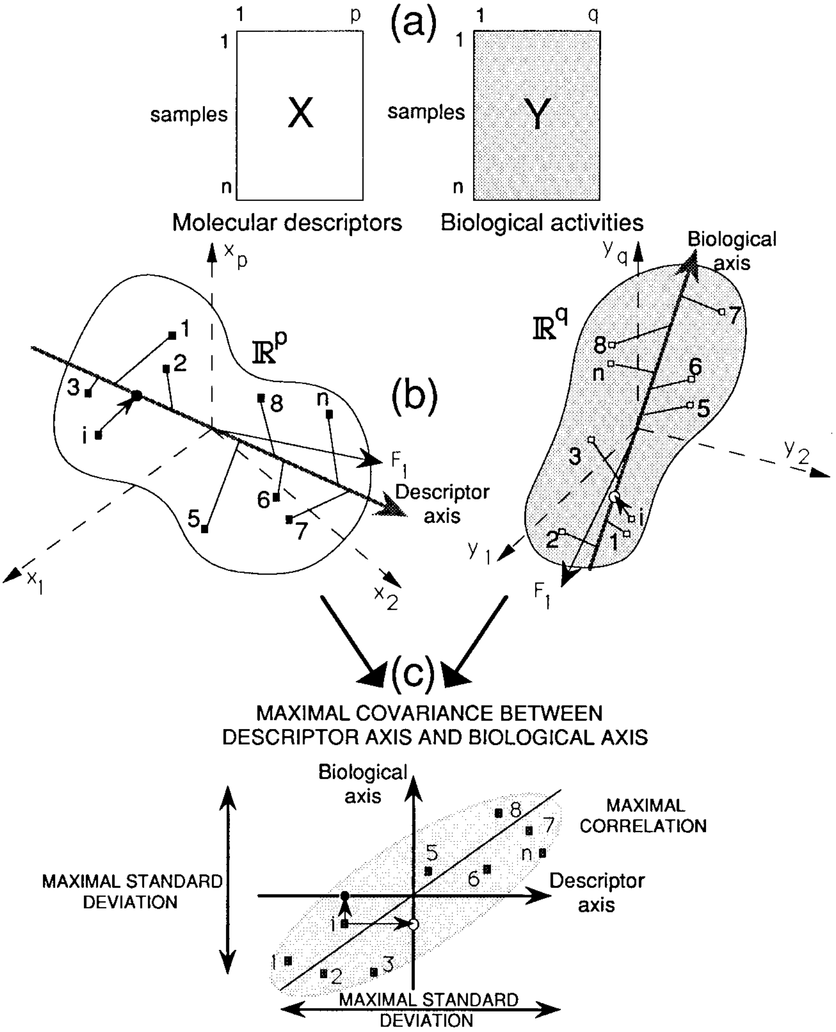
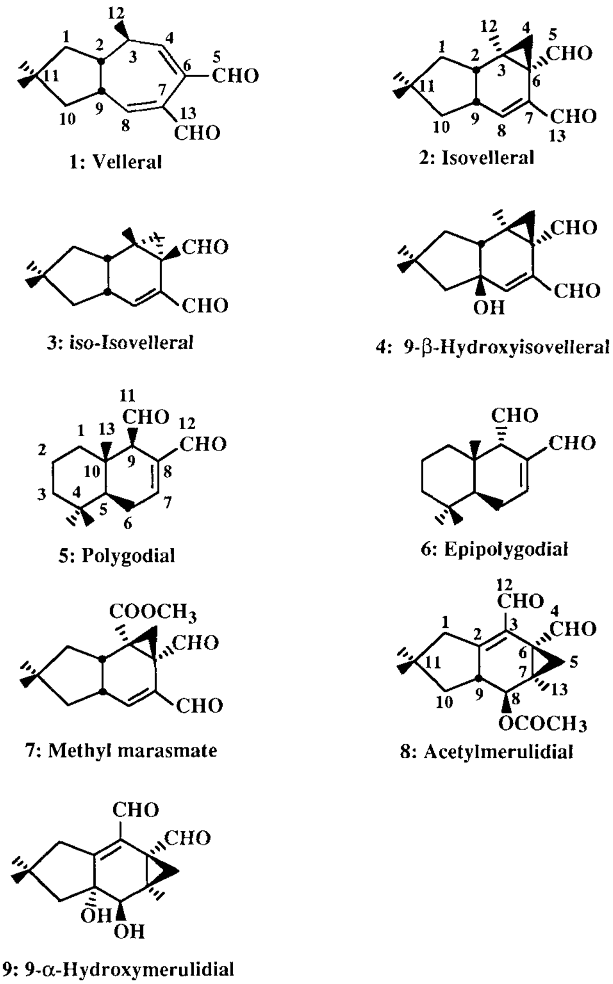


Figure 1. Flow diagram of co-inertia analysis in the context of SAR and QSAR studies.

4.2.3.2 Example: Structure-Reactivity Relationships for Unsaturated Dialdehydes

Chemical Stability of Sesquiterpenoid Unsaturated Dialdehydes

A large number of terpenoids with an unsaturated dialdehyde functionality group have been isolated from various organisms, which occupy different trophic levels in the environment (e.g. the Basidiomycete, *Lactarius vellereus* [28], and the Nudibranch, *Dendrodoris grandiflora* [29]). Their potent activity as antifeedants, antibiotics, mutagens, and so on [30–32], has stimulated both biological and



1: Velleral

2: Isovelleral

3: iso-Isovelleral

4: 9-β-Hydroxyisovelleral

5: Polygodial

6: Epipolygodial

7: Methyl marasmate

8: Acetylmerulidial

9: 9-α-Hydroxymerulidial

Figure 2. Structure of the nine sesquiterpenoid unsaturated dialdehydes under study.

chemical investigations. These studies have shown that small structural variations in the molecules can considerably change their activity [32, 33]. The have also stressed numerous contradictory results, certainly in relation with the instability of some of these chemicals in assay media [34]. In order to confirm this hypothesis, we have tried to find the co-structure between a data matrix (similar to Table *Y* in Fig. 1), describing the stability of nine sesquiterpenoid unsaturated dialdehydes (Fig. 2) in three different *in vitro* assay media [34] and another data matrix (similar to Table *X* in Fig. 1), characterizing these molecules by means of the five following molecular descriptors [35]:

- angle: the dihedral angle (°) between the two aldehyde groups,
- distance: the distance (Å) between the two aldehyde carbons,

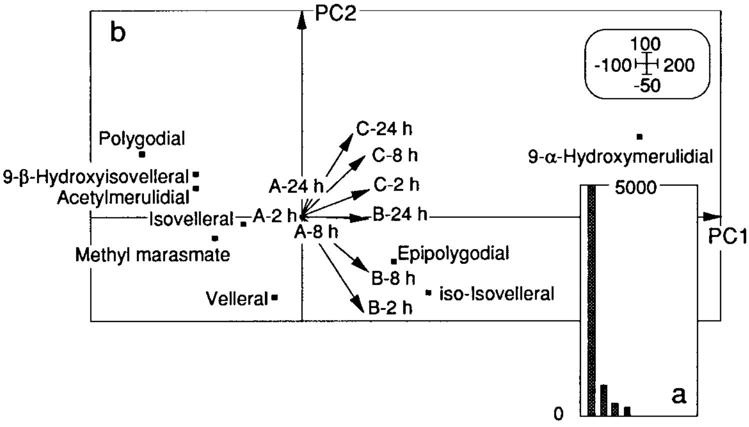


Figure 3. Centered PCA of the chemical stability data table. a) Eigenvalues. b) Biplot. A, B, and C stand for medium A (Dulbecco's phosphate buffered saline (PBS), pH 7.3, without Ca^{2+}, Mg^{2+} or bicarbonate), medium B (Eagle's minimum essential medium (MEM) with L-glutamine and Ham's F12), and medium C (medium B supplemented with 10% fetal calf serum), respectively.

– dipole *X*: the dipole moment in debyes (obtained by CNDO calculations) along the C_8-C_7 double bond in compounds 1–4, and 7, the C_7-C_8 double bond in compounds **5** and **6**, and the C_2-C_3 double bond in compounds **8** and **9**,
– dipole: the dipole moment in debyes (obtained by CNDO calculations),
– log *P*: the *n*-octanol/water partition coefficient.

Co-Structure Analysis

Separate Analyses

Table *Y* (Fig. 1), containing the amounts (in %) of the nine chemicals under study (Fig. 2) remaining after 2, 8, and 24 hours of incubation in the three media [34], was analyzed by means of a centered PCA. The graphical display of the eigenvalues (Fig. 3a) shows that the main information is carried by the first principal component (*PC*1), but that *PC*2 can be also considered for the interpretation of the data. Thus, the biplot [36] displayed in Fig. 3b reveals that only the two cell culture media B and C participate in the analysis. It also clearly underlines the effects of the media on the chemical stability of the molecules. Compounds located on the right-hand side of the map are more stable than those located on the left-hand side. Thus, 9-α-hydroxymerulidial appears as an outlier on the right-hand side since it is not reactive in media A, B, and C. Conversely, polygodial is very unstable, especially in media B and C. It is also interesting to note that small structural changes can considerably affect the reactivity of these chemicals. Thus, for example, isovelleral is distant from its stereoisomer (*iso*-isovelleral), and polygodial is opposed to its epimer (epipolygodial).

The different degrees of correlation between the molecular descriptors [35] are shown on Fig. 4 which can be easily interpreted due to the combined use of numerical

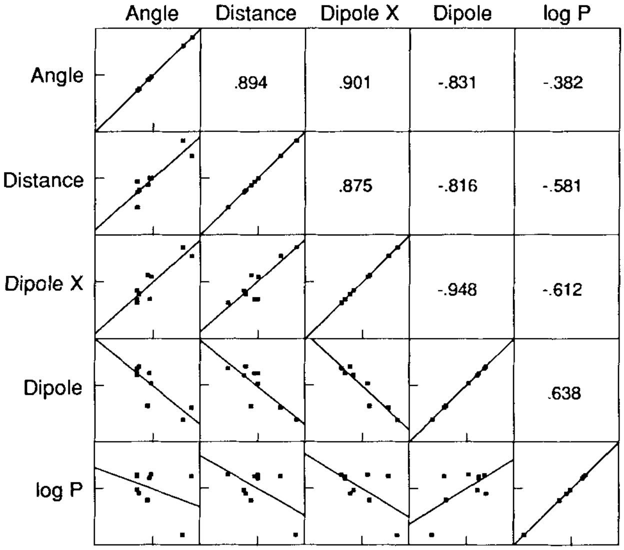


Figure 4. Correlation matrix of the five physico-chemical variables under study.

data and scatter plots. The data matrix of molecular descriptors (Table *X*, Fig. 1) was processed by a standardized PCA. According to the eigenvalues (Fig. 5a), *PC*1 and *PC*2 enable a graphical interpretation of the data. Fig. 5b shows that *PC*1 is mainly explained by the angle, distance and dipole *X* variables, which are negatively correlated to the dipole variable, and *PC*2 is principally dependent on log *P*. They principally govern the distribution of the compounds (Fig. 5c) and show

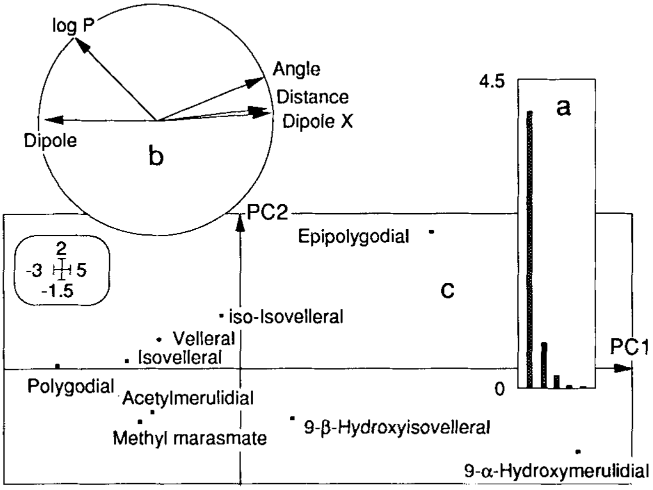


Figure 5. Standardized PCA of the physico-chemical data table. a) Eigenvalues. b) Correlation circle. c) Factorial plane (*PC*1 – *PC*2) of the compounds.

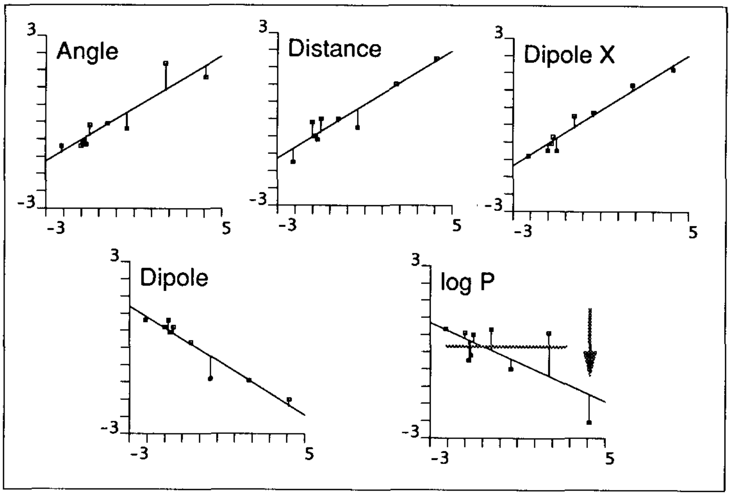


Figure 6. Plots of standardized physico-chemical variables vs *PC*1 coordinates.

that 9-α-hydroxymerulidial is an outlier. These relationships are illustrated in Fig. 6, which clearly emphasizes the atypical log *P* value of 9-α-hydroxymerulidial.

Matched Analysis

The projection of the chemicals, as defined by their physico-chemical properties on the first factorial plane resulting from the co-inertia analysis, shows (Fig. 7a) that their distribution is similar to that obtained from the separate analysis (Fig. 5), disregarding a rotation (Fig. 7b). This is confirmed by the comparison of Fig. 5c and 7a. Fig. 7c represents the weights of the physico-chemical parameters in the equations of scores of the chemicals on the first co-inertia plane. Fig. 7d gives the correlations between the physico-chemical parameters and the scores of the chemicals on the co-inertia axes. They emphasize the key role of distance, angle, and dipole *X* variables on the first axis and that of log *P* on the second axis. Furthermore, as previously mentioned, a "distortion" is introduced by 9-α-hydroxymerulidial, which suggests that the role of distance and angle variables is more important. In the same way, the projection of the nine sesquiterpenoid unsaturated dialdehydes, as defined by their chemical stability data on the factorial plane resulting from the co-inertia analysis, shows (Fig. 8) that the distribution of the chemicals is similar to that obtained from the separate analysis (Fig. 3). To compare the graphical displays in Figs. 7a and 8a, it is also possible to plot their respective scores after normalization and to link the two positions of a given chemical by an arrow (Fig. 9). This procedure emphasizes the correlations between the scores on the axes of the co-inertia analysis (designated as r_1 and r_2 on Fig. 9).

Figs. 7 to 9 clearly demonstrate the existence of a co-structure between the chemical stability of the nine compounds studied and the selected molecular descriptors. This is not surprising, if we consider that the main difference between the three media is that medium A is inorganic, while media B and C contain amino acids. Under

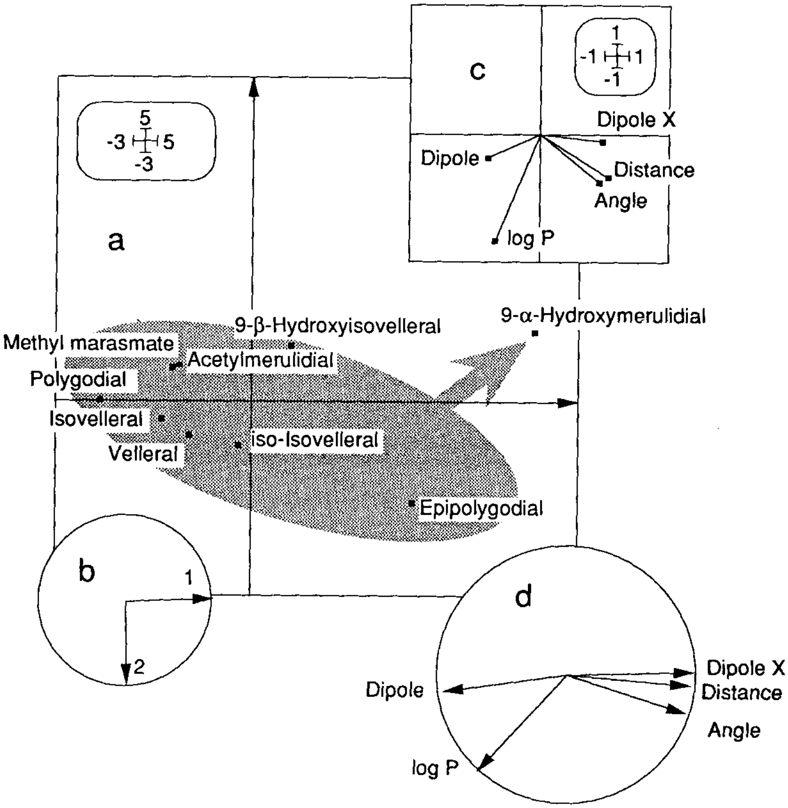


Figure 7. a) Representation of the compounds, as defined by their physico-chemical data on the first co-inertia plane. b) Projection of the PCs obtained from the separate analysis (Fig. 5) on the co-inertia axes of the physico-chemical data table. c) Weights of the physico-chemical parameters in the equations of scores of the chemicals on the first co-inertia plane. d) Correlations between the physico-chemical parameters and the scores of the compounds on the co-inertia axes.

these conditions, as already mentioned in the literature [37, 38], we can advance that some sesquiterpenoid unsaturated dialdehydes can react with amines contained in media B and C to form pyrrole derivatives. As this type of reaction obviously depends on the distance and angle between the two aldehyde groups, it is not surprising to observe a co-structure between the two data tables and to find that the two above molecular descriptors play a key role in the analysis.

4.2.4 Conclusion

The aim of this study was not to give a catalogue raisonné of all the graphical methods, which can be used in medicinal chemistry to enhance the statistical results produced by SAR and QSAR studies. Indeed, our intention was only to present

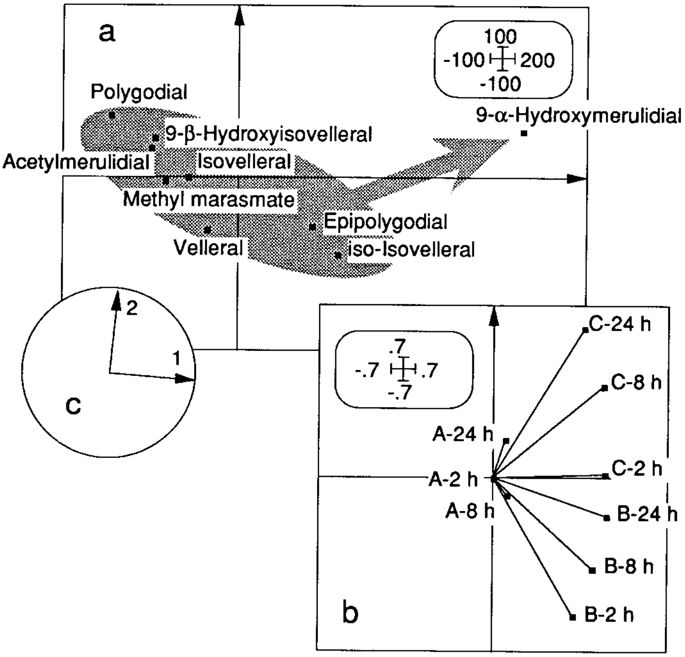


Figure 8. a) Representation of the compounds, as defined by their chemical stability on the first co-inertia plane. b) Weights of the chemical stability parameters in the equations of scores of the chemicals on the first co-inertia plane. c) Projection of the PCs obtained from the separate analysis (Fig. 3) on the co-inertia axes of the compound stability data table.

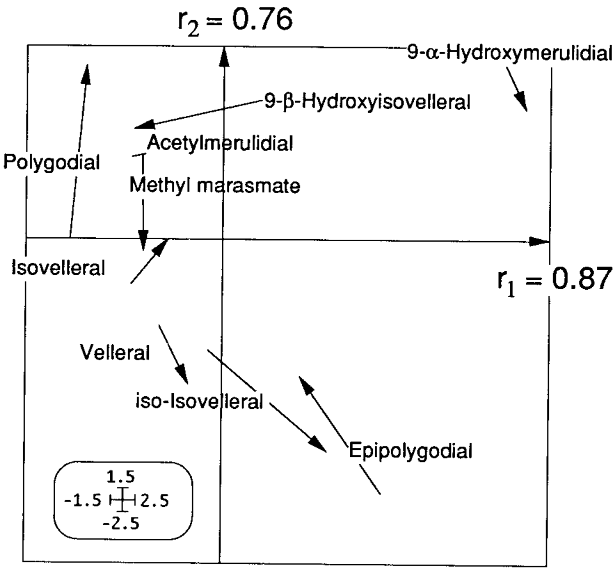


Figure 9. Matched display of Figs. 7a and 8a after normalization of their respective scores. The correlation coefficients between the scores on the axes of the co-inertia analysis are r_1 and r_2.

some basic principles of graphics and to try to illustrate them in the particular case dealing with the study of the co-structure between two data tables. Furthermore, it is obvious that these principles must be only considered as guides and not as rigid laws. Indeed, as mentioned by Tufte [2], with regards to graphical analysis, *"The principles should not be applied rigidly or in a peevish spirit; they are not logically or mathematically certain; and it is better to violate any principle than to place graceless or inelegant marks on paper"*.

References

[1] Tukey, J. W., *Exploratory Data Analysis*, Addison-Wesley Publishing Company, Reading, 1977

[2] Tufte, E. R., *The Visual Display of Quantitative Information*, Graphic Press, Cheshire 1983

[3] Cox, D. R., *Appl. Statist.* **27**, 4–9 (1978)

[4] Wang, P. C. C., and Lake, G. E., Application of Graphical Multivariate Techniques in Policy Sciences. In: *Graphical Representation of Multivariate Data*, Wang, P. C. C., ed., Academic Press, New York (1978) p. 13–58

[5] Cleveland, W. S., and McGill, R., *Science* **229**, 828–833 (1985)

[6] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A., *Graphical Methods for Data Analysis*, Wadsworth International Group, Belmont, 1983

[7] Everitt, B. S., *Graphical Techniques for Multivariate Data*, Heinemann Educational Books, London, 1978

[8] Bintein, S., Devillers, J., and Karcher, W., *SAR QSAR Environ. Res.* **1**, 29–39 (1993)

[9] Weihs, C., *J. Chemometrics* **7**, 305–340 (1993)

[10] Thioulouse, J., Devillers, J., Chessel, D., and Auda, Y., Graphical Techniques for Multidimensional Data Analysis. In: *Applied Multivariate Analysis in SAR and Environmental Studies*. Devillers, J., and Karcher, W., eds., Kluwer Academic Publishers, Dordrecht (1991) p. 153–205

[11] Chessel, D., and Mercier, P., Couplage de Triplets Statistiques et Liaisons Espèces-Environnement. In: *Biométrie et Environnement*, Lebreton, J. D., and Asselain, B., eds., Masson, Paris, 1993, p. 15–43

[12] Dolédec, S., and Chessel, D., *Freshwater Biol.* **31**, 277–294 (1994)

[13] Schmid, C. F., *Handbook of Graphic Presentations*, Ronald Press, New York, 1954

[14] Bertin, J., *Sémiologie Graphique*, Gauthier-Villars, Paris 1967

[15] Chernoff, H., Graphical Representations as a Discipline. In: *Graphical Representation of Multivariate Data*, Wang, P. C. C., ed., Academic Press, New York (1978) p. 1–11

[16] Cleveland, W. S., Harris, C. S., and McGill, R., *J. Am. Stat. Assoc.* **77**, 541–547 (1982)

[17] Domine, D., Devillers, J., Chastrette, M., and Karcher, W., *Pestic. Sci.* **35**, 73–82 (1992)

[18] Devillers, J., Thioulouse, J., and Karcher, W., *Ecotoxicol. Environ. Safety* **26**, 333–345 (1993)

[19] Gnanadesikan, R., *Methods for Statistical Data Analysis of Multivariate Observations*, John Wiley & Sons, New York, 1977

[20] Julesz, B., *Nature* **290**, 91–97 (1981)

[21] Pollak, P. T., *Eur. J. Clin. Pharmacol.* **39**, 525–532 (1990)

[22] Chessel, D., and Dolédec, S., *ADE Version 3.6, HyperCard® Stacks and Programme Library for the Analysis of Environmental Data*. User's Manual (in French). URA CNRS 1451, Université Lyon 1, E-mail: chessel@biomserv.univ-lyon1.fr.

[23] Thioulouse, J., *CABIOS* **5**, 287–292 (1989)

[24] Thioulouse, J., *Comput. Geosci.* **16**, 1235–1240 (1990)

[25] Geladi, P., and Kowalski, B. R., *Anal. Chim. Acta* **185**, 1–17 (1986)

[26] Tucker, L. R., *Psychometrika* **23**, 111–136 (1958)

[27] Cazes, P., *Cahiers de l'Analyse des Données* **5**, 145–161 (1980)

[28] Sterner, O., Bergman, R., Kihlberg, J., and Wickberg, B., *J. Natural Products* **48**, 279–288 (1985)

[29] Cimino, G., de Rosa, S., de Stefano, S., Morrone, R., and Sodano, G., *Tetrahedron* **41**, 1093–1100 (1985)

[30] Giannetti, B. M., Steffan, B., Steglich, W., Quack, W., and Anke, T., *Tetrahedron* **42**, 3579–3586 (1986)

[31] Caprioli, V., Cimino, G., Colle, R., Gavagnin, M., Sodano, G., and Spinella, A., *J. Natural Products* **50**, 146–151 (1987)

[32] Forsby, A., Andersson, M., Lewan, L., and Sterner, O., *Toxic. in Vitro* **5**, 9–14 (1991)

[33] Bergquist, J., Strandberg, C., Andersson, M., Sterner, O., Pesando, D., and Girard, J. P., *Toxic. in Vitro* **7**, 205–212 (1993)

[34] Sterner, O., Andersson, M., Forsby, A., and Morales, P., *ATLA* **19**, 171–177 (1991)

[35] Andersson, M., Bocchio, F., Sterner, O., Forsby, A., and Lewan, L., *Toxic. in Vitro* **7**, 1–6 (1993)

[36] Gabriel, K. R., Biplot Display of Multivariate Matrices for Inspection of Data and Diagnosis. In: *Interpreting Multivariate Data*, Barnett, V., ed., John Wiley, Chichester (1981) p. 147–173

[37] D'Ischia, M., Prota, G., and Sodano, G., *Tetrahedron Lett.* **23**, 3295–3298 (1982)

[38] Cimino, G., Sodano, G., and Spinella, A., *Tetrahedron* **43**, 5401–5410 (1987)

4.3 SIMCA Pattern Recognition and Classification

William J. Dunn III and Svante Wold

Abbreviations

As given in Chapter 4.4 on PLS and:

CNDO/2	Complete neglect of differential overlap
ECI	Electronic charge index
ISA	Isotropic surface area
MMFF	Molecular mechanics force field

Symbols

A	The number of latent variables in principal components or PLS models
$c_{a,j}$	*a*th PLS loading for biological activity, *j*
d	Orthogonal projection distance of a compound to the class model
d^*	Degrees of freedom corrected distance
d_{IMS}	Inside model space distance
d_n	Revised SIMCA classification distance
d_{OMS}	Outside model space distance
$e_{i,k}$	Residual for compound, *i*, variable, *k*
K	Number of physico-chemical descriptors or independent variables in a data set
M	Number of biological activities or dependent variables in a data set
N	Number of compounds in a data set
$p_{a,k}$	*a*th principal component or PLS loading for variable, *k*
$\boldsymbol{R_k^2}$	Cross-validated R^2 (often referred to as $\boldsymbol{Q}^2$)
s_{eps}	Residual standard deviation
$t_{i,a}$	*a*th principal component or PLS score for compound, *i*
$\hat{u}$	Estimated PLS loading from the inner relation
$u_{\mathrm{i,a}}$	*a*th PLS loading for compound, *i*
$w_{\mathrm{a,k}}$	*a*th PLS weight for compound, *k*
$\boldsymbol{X}$	Physico-chemical descriptor or feature matrix, independent variable data matrix
$x_{i,k}$	Physico-chemical descriptor or feature *k* for compound, *i*
$\boldsymbol{Y}$	Biological activity data or dependent variable data matrix
$y_{i,j}$	Biological activity, *j*, for compound, *i*
z_1, z_2, z_3	Principal properties of the amino acids

4.3.1 Introduction

The SIMCA method of pattern recognition and classification (hence, abbreviated **PARC**) was first described in 1976 [1] and as a tool in drug design, it was last reviewed by the authors in 1990 [2]. Initially, SIMCA was an acronym for SIMple Classification Analysis, but was soon reinterpreted by Dave Duewer as Soft Independent Modeling of Chemical Analogy.

While other methods of PARC have been applied to drug design problems, SIMCA remains the method of choice, and a number of recent quantitative structure-activity relationship (QSAR) studies have been published using the SIMCA method. Rather than focusing on these, the subject of this chapter is recently reported improvements and extensions of existing methods. Before discussing these, some ideas about PARC will be presented.

The use of PARC to solve QSAR problems was stimulated by the fact that the traditional Hansch [3] analysis could not solve the active vs inactive problem. It could only deal with structure-activity relationships of active compounds. A method was necessary to estimate the probability that a compound may be inactive and PARC was well suited for this problem. The objective of PARC is classification, making it ideal for application to the active vs inactive case. Analogous to traditional QSAR methods, features or variables are used to describe objects (compounds) quantitatively. The resulting data (the training set) are used to derive structure-based models, which can then be used to classify new objects of unknown class (the test set). Here, since we are discussing QSAR, the objects are compounds and the features or variables are generally measured physico-chemical descriptors or other variables, which can be computed from the structures of the compounds. In most cases, continuous variables are used but in some exceptional cases, discrete variables are used. The variables must be relevant to the investigated activity. If the design of new compounds is the objective, then variables must be two-way predictive. This means that (i) they must be derivable from the structure of a compound without actually synthesizing it, and (ii) one must be able to derive a compound structure from a profile of structure descriptor variables, which correspond to promising activity levels, as indicated by the model.

The information obtained from a PARC study is categorized in to what is now known as "the three levels of PARC" [4]. Hence, it is important to use a PARC method that corresponds to the information required from the analysis.

At the lowest level, level I, the objective is just to classify an unknown into one of several specified classes. The limitation of working at this level is that the unknowns are assumed to be members of these specified classes. However this is seldom the case. Considering, as an example, the problem of classifying chemical pollutants as carcinogens vs non-carcinogens, this is equivalent to assuming that all of the mechanisms by which a compound can be a carcinogen are known.

At level II, the assumption above is not strictly adhered to and the result "none of the specified classes" is allowed. Here it is possible to predict that a compound could be a non-carcinogen, but might alternatively be a member of a new class of carcinogens. This is the lowest level at which SIMCA works. At levels I and II, the structure-activity relationships are qualitative, providing only classification.

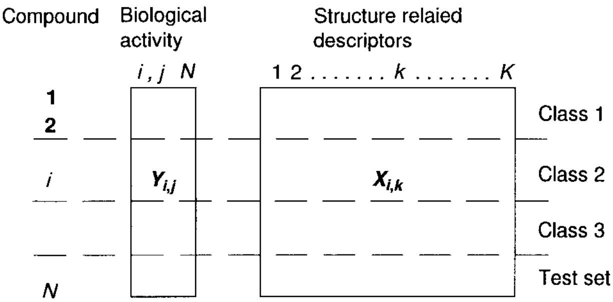


Figure 1. Standard pattern recognition data matrixes.

At level III, in addition to classification, the level of activity in one or more assays of a compound is estimated. This is similar to Hansch analysis [3] combined with discriminant analysis, but SIMCA gives a more robust and stable solution.

4.3.2 SIMCA Pattern Recognition

All classification studies begin with a data set as shown in Fig. 1. The Y-block contains the biological activities and the X-block contains the descriptors. At levels I and II, the analysis is performed only on $X_{i,\,k}$, and a separate Y-block often does not exist. At level III a predictive relationship between the X- and Y-blocks is derived.

Principal components analysis is used at the first two levels to derive a separate model for each "proper" class (see the asymmetric case below). Before the analysis, the data are scaled, usually to unit variance (autoscaling) within each class. The principal components model is given in Eq. (1):

$$x_{i,k} = \overline{x_k} + \sum_{a=1}^{A} t_{i,a} p_{a,k} + e_{i,k} \qquad (1)$$

Here $\overline{x_k}$ is the mean of column k, $t_{i,a}$ is the ath score for compound i, $p_{a,k}$ is the ath loading for variable k, and $e_{i,k}$ is the residual. The A components are calculated to make the es as small as possible in the least squares sense. The number of components is determined by cross validation [6].

SIMCA works by deriving a model for each class. Thereafter, classification is accomplished by projecting the data of the test compounds onto each of the training sets via the class models in the descriptor space, and classification is determined from the magnitude of the resulting residual standard deviation of the es. This is shown graphically in 3-dimensions in Fig. 2.

Since similar compounds cluster in the same regions of descriptor space, compounds with similar biological activities will also have similar score values, $t_{i,a}$, in Eq. (1). However, the principal component scores, or $t_{i,a}$s, are not optimal for estimation of dependent variables at level III. Instead, PLS is used for classification and prediction at level III (see also the chapter on PLS, Chap. 4.4).

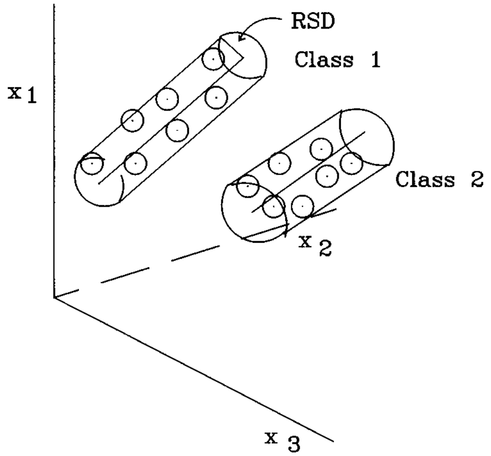


Figure 2. Graphical description of SIMCA classification rules.

The form of the PLS model is given in Eqs. 2—4, below:

$$x_{i,k} = \overline{x_k} + \sum_{a=1}^{A} t_{i,a} p_{a,k} + e_{i,k} \tag{2}$$

$$y_{i,j} = \overline{y_j} + \sum_{a=1}^{A} u_{i,a} c_{a,j} + f_{i,j} \tag{3}$$

$$\hat{u} = b \cdot t \tag{4}$$

The variables, *t*s and *u*s, are latent variables calculated along the axes of greatest variation in *X* and *Y*. The latent variables make the *x*-residuals and *y*-residuals as small as possible and are maximally correlated. They are related through the inner relation, expressed by Eq. (4). The PLS model is shown graphically in Fig. 3.

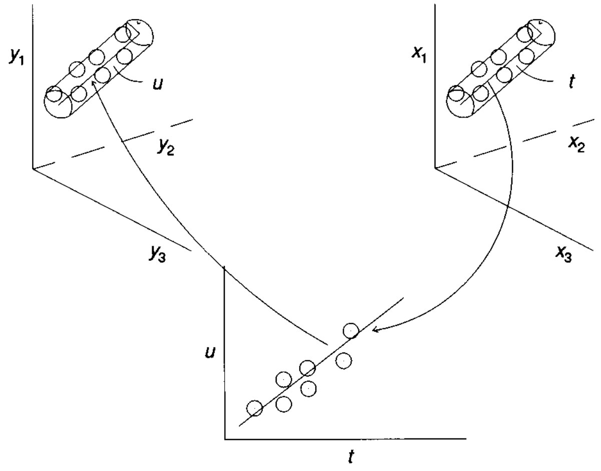


Figure 3. Graphical representation of the PLS model.

4.3.3 Steps in a Pattern Recognition Study

PARC studies are carried out in defined steps which are: 1) selecting the training set compounds and developing the training data set (Fig. 1), 2) data preprocessing (transformation, scaling, centering), 3) developing, optimizing and validating the classification models, and 4) classification of the test set compounds.

Step 1, ideally, should involve experimental design if the training set is to span the descriptor space. This topic is discussed by others in this volume. Data preprocessing, Step 2, is data set dependent and will not be discussed here. SIMCA is unique in that it derives *separate* class models in Step 3 making it work at level II, if classification is the sole objective. Step 4 is a matter of fitting data for the unknown or untested compounds to the class models from Step 3. A number of recent developments have been made in the areas above, especially in Steps 1 and 3.

4.3.4 Establishing the Training Sets

The training set refers to the set of compounds, whose relevant descriptors or features are to be used in the learning phase. Historically, training sets are designed from a lead compound. The result is a group of compounds, which are "similar" to the lead compound and, for the most part, those that should be most easily synthesized. As mentioned above, training sets should be established from experimental design methods, but this is seldom the case. Even though far from ideal, such data can be, and have been shown to be, very useful.

The most difficult aspect of a QSAR study is finding the relevant descriptors. Traditionally, the Hansch method uses linear free energy related parameters. These are what are termed macroscopic variables or properties of the system, in that they are Boltzmann averages of the properties of the many states of the system. Such data are log P, pKa, etc., and may have minimal information about the active state of the system if multiple states are possible. With the advent of molecular modeling, it is now possible to generate descriptors for compounds in discrete states, thus, adding additional dimensions to the QSAR problem. An example of such methods is the CoMFA method [7] which generates descriptors according to a user-specified conformation and alignment.

4.3.4.1 Consideration of Conformation and Alignment of Flexible Compounds

An important, unsolved problem in modeling the changes in biological activity with differences in chemical structure within a series of flexible molecules, e.g. peptides, is finding the optimal conformation and alignment for the series (if it exists). **Conformations** of a molecule, as defined by Eliel et al. [8] are the non-identical arrangements of the atoms in a molecule obtained by rotation about one or more

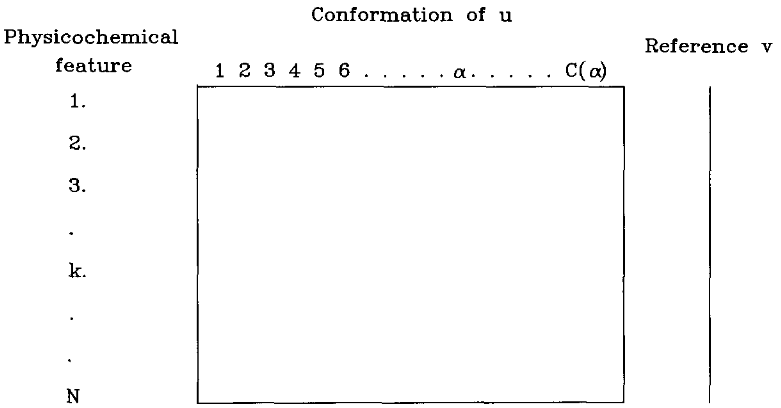


Figure 4. MATRIX analysed by PLS to find a conformation similar to a reference, v.

single bonds. A common conformation for a series of compounds would be one, in which a common set of atom positions or torsion angles is specified. An **alignment** of conformations is the arrangement of two or more molecules, in which a common set of atoms, substructures or features is superimposed. Currently, the active conformation and alignment must be known a priori. Recently, a general solution for finding the conformation/alignment responsible for biological activity of flexible compounds has been proposed [9] and applied to the structure-activity data for a series of twenty-one flexible tricyclic pyridodenzodiazepinone (I) inhibitors of the muscarinic receptors [9], M2 and M3. In this case, the alignment was known and each analog could exist in as few as 9, but also in as many as 706 conformations with energies 6 kcal/mol or less. Each conformer was represented by 29 variables, most of which were conformationally dependent [10]. It was assumed that the lowest energy conformation of the most active compound was the active conformer. To find the conformation of each analog most similar to that of the reference compound, a PLS analysis of the matrix in Fig. 4 was carried out.

The X-block is the physico-chemical data for all low energy conformers of compound, u. In this case, $N_D = 29$ variables, most of which are conformationally dependent. The dependent variable was the vector of variables for the reference compound, v, in the active conformation. The conformation in X was selected that was most "similar" to the reference compound in the PLS sense. A scoring system was devised [9] to score each conformer for each compound. Then the features of this conformer for each compound were used to construct a regular data matrix, as shown in Fig. 1, for QSAR development. The result was a predictively significant 3D QSAR.

4.3.4.2 Novel Descriptors for Peptide QSAR

There have been recent developments in the QSAR of peptides using newly developed structure-based descriptors and PLS [11]. Hellberg and coworkers [12] were the

Table 1. Descriptors for bitter tasting dipeptides, $H_2N-CH(R_1)-C(=O)NH-CH(R_2)COOH$

No.	Peptide	Isotropic surface area		Electronic charge index	
		R_1	R_2	R_1	R_2
1	GV	19.93	120.91	0.02	0.07
2	GL	19.93	154.35	0.02	0.10
3	GI	19.93	149.77	0.02	0.09
4	GP	19.93	122.35	0.02	0.16
5	GF	19.93	189.42	0.02	0.14
6	GW	19.93	179.16	0.02	1.08
7	GY	19.93	132.16	0.02	0.72
8	AV	62.90	120.91	0.05	0.07
9	AL	62.9	154.35	0.05	0.10
10	AF	62.90	189.42	0.05	0.14
11	VG	120.91	19.93	0.07	0.02
12	VA	120.91	62.90	0.07	0.05
13	VV	120.91	120.91	0.07	0.07
14	VL	120.91	154.35	0.07	0.10
15	LG	154.35	19.93	0.10	0.02
16	LA	154.35	62.90	0.10	0.05
17	LL	154.35	154.35	0.10	0.10
18	LF	154.35	189.42	0.10	0.14
19	LW	154.35	179.16	0.10	1.0
20	LY	154.35	132.16	0.10	0.72
21	IG	149.77	19.93	0.09	0.02
22	IA	149.77	62.90	0.09	0.05
23	IV	149.77	120.91	0.09	0.07
24	IL	149.77	154.35	0.09	0.10
25	II	149.77	149.77	0.09	0.09
26	IP	149.77	122.35	0.09	0.16
27	IW	149.77	17.87	0.09	1.08
28	IN	149.77	17.87	0.09	1.31
29	ID	149.77	18.46	0.09	1.25
30	IQ	149.77	19.53	0.09	1.36
31	IE	149.77	30.19	0.09	1.31
32	IK	149.77	102.78	0.09	0.53
33	IS	149.77	19.75	0.09	0.56
34	IT	149.77	59.44	0.09	0.65
35	PA	122.35	62.9	0.16	0.05
36	PL	122.35	154.35	0.16	0.10
37	PI	122.35	149.77	0.16	0.09
38	PY	122.35	132.16	0.16	0.72
39	PF	122.35	189.42	0.16	0.14
40	FG	189.42	154.35	0.14	0.10
41	FL	189.42	154.35	0.14	0.16
42	FP	189.42	122.35	0.14	0.16
43	FF	189.42	189.42	0.14	0.14
44	FY	189.42	132.16	0.14	0.72
45	WE	179.16	30.19	1.08	1.31
46	WW	179.16	179.16	1.08	1.08
47	YL	132.16	154.35	0.72	0.10
48	SL	19.75	154.35	0.56	0.10

first to successfully develop a strategy for deriving QSAR for these important compounds. By tabulating a large number of measured and theoretical properties of amino acids and their derivatives, principal components analysis was used to derive three principal properties, z_1, z_2, and z_3, for each amino acid. The principal properties are linear combinations of the primary data and were proposed to model the hydrophilic, bulk and electronic nature of the side-chain substituent, respectively. In QSAR studies, the principle properties, $z_1 - z_3$, are used as substituent constants for each amino acid as it appears in the peptide sequence, and PLS is applied to the resulting data matrix to derive the QSAR. This approach has been applied to a number of peptide structure-activity studies, but has been criticized for being difficult to interpret, because of the linear combination problem and for not considering the conformational state of the peptides in the derivation of the principle properties, $z_1 - z_3$. This interpretation question can be resolved by attempting to correlate z_1, z_2 and z_3 with variables which can be computed from the structure. If the primary variables can be identified, and if conformation of the peptide is considered in the computation of the variable, the drug design process will be much more straightforward.

In order to identify the underlying primary variables of z_1, z_2 and z_3, features related to the two most significant, z_1 and z_2, were explored. These are hydrophilic in character, or inversely hydrophobic in character, and electronic in character, respectively. It has been shown that the isotropic surface area of a solute is highly correlated with hydrophobicity [13, 14]. This parameter, defined as the solvent-accessible surface area associated with the nonpolar portion of the supermolecule solute structure, was found to be highly correlated with z_1. The isotropic surface area is computed on the free amino acid structure which results from its optimization with the AMBER force field with Molecular Mechanics Force Field, MMFF.

In order to model the electronic nature of the α-carbon of the amino acid, the sum of the absolute values CNDO/2 charges of the atoms in the substituents of the α-carbon were used. This variable, which models the charge separation, is termed

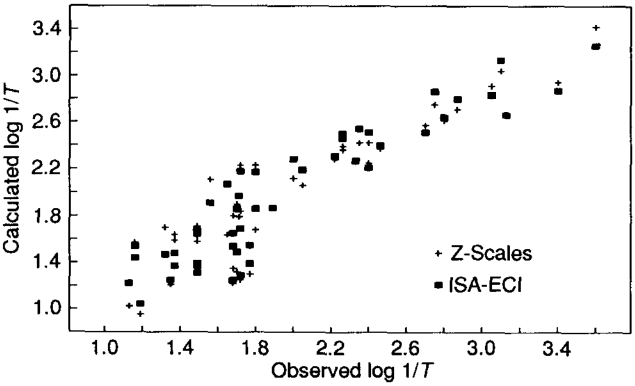


Figure 5. A comparison of the predicted activity of bitter dipeptides using the z-scales and isotropic surface area, ISA, and the electronic charge index, ECI.

the electronic charge index, or ECI. In both cases the variables are conformationally dependent. Data for bitter dipeptides, analyzed by Hellberg et al. [15], are given in Table 1.

When used as side-chain or substituent constants for the α-carbon for amino acids in peptides, these two variables worked as well as z_1-z_3, as determined by R^2 or the variance explained. Fig. 5 is a plot of the observed and predicted activities for bitter dipeptides, analyzed previously by Hellberg et al. [15]. The same number of PLS components was used in both cases, but only two variables were used per amino acid residue in the peptide.

4.3.5 Symmetric and Asymmetric Data Structures

The problem of predicting that some compounds will be biologically active and others will be inactive stimulated much of the early applications of PARC to structure-activity data. Such applications lead to our proposal that QSAR problems lead to two types of data structure: 1) symmetric data and 2) asymmetric data [4, 16]. Fig. 2 is an example of symmetric data structure. Two or more classes form well-defined clusters in descriptor space. This results in classification studies of antagoniste vs. agonists, substrate vs inhibitor, etc. This contrasts with another notation for asymmetric data structure, namely embedded structure, which is discussed in recent article by Rose et al. [17].

However, in studies of active vs inactive, carcinogen vs non-carcinogen, for example, the data structures are often asymmetric. Here, only one of the classes, usually the one with *active* compounds has a data structure that can be modeled; only this class contains compounds that are *similar* to each other (biologically and structurally). The other class, usually the one with *inactives*, is not a **proper class**, and thus, has no inherent similarity and cannot be modeled. This is because a compound can be inactive for many different reasons, but activity needs a

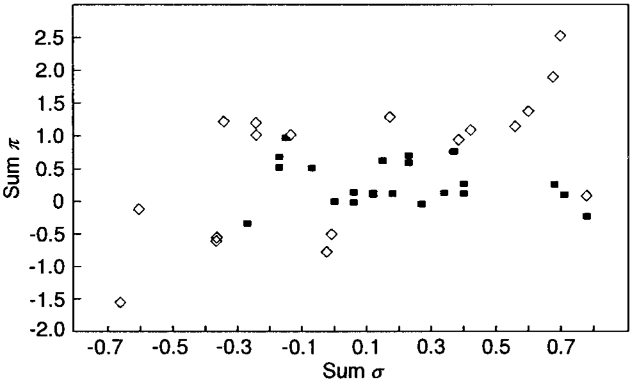
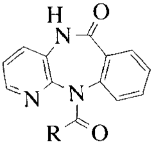


Figure 6. Plot of asymmetric data for carcinogenic (■) and noncarcinogenic (◇) dimethylaminoazobenzenes.

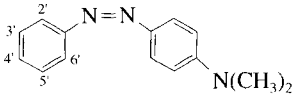
well-defined structure. This is analogous to control theory, where a process under control occupies a small regular part of the multivariate space, while the process can be anywhere in this space when it is out of control.

An example of asymmetric data structure is given in Fig. 6. The data are sum pi, which is the sum of the Hansch π constants for the substituents on the x-axis, and sum sigma, which is the sum of the Hammett σ constants for the substituents on the y-axis for substituents in the substituent (′) ring of dimenthylaminoazobenzenes (II). They were analyzed with SIMCA by Miyashita et al. [18] using the original data of Hansch [19].

The asymmetric nature of the data in Fig. 6 is striking and illustrates the power of the SIMCA method. It is the only method, which is routinely used in QSAR studies, that can handle this type of data structure.



(I)



(II)

4.3.6 Variable Selection

As discussed previously, SIMCA is the method of choice for classification problems that require results at level III. It must be realized, however, that each data set is unique, and to obtain the best results from a method, care must be taken to insure that the prediction rules are optimal for that data set. One aspect of optimizing the classification rules is variable selection. In SIMCA and PLS one can select X-variables on the basis of their residuals, i.e. $\boldsymbol{R}_k^2$ (or the similar MPOW of early SIMCA papers), their discriminating power (importance for distinguishing between classes), and, on level III, their importance for predicting Y. Here, a number of methods have been recently developed by Clementi (see the chapter on GOLPE in Vol. 3), Marsili [20], and others. These are based on cross-validation and are computationally extensive. The *VIP* statistic of Wold et al. (see the chapter on PLS), which is a measure based on the weighted PLS coefficients ($w_{u,k}$) in significant model components *VIP*, seems to form a reasonable basis for variable selection, and has the advantage of not demanding additional computations beyond the model estimation. There is a clear need to evaluate these measures of variable relevance before any strong recommendations can be made.

4.3.7 Determining the Model Complexity

An important point to stress is the difference in prediction error and fitting error [18]. Fitting error is based on predicting the training objects and decreases with model complexity (adding components). Prediction error is based on estimation of compounds not included in model development. It decreases, goes through a minimum and then generally increases with model complexity. Selection of components based on cross-validation [6] gives models with optimal prediction capability. Indeed, predictive capability is identical to the cross-validated R^2 statistic (often denoted as Q^2) used by Cramer et al. [7] for selecting of the optimal PLS model in their CoMFA® method.

4.3.8 Developing, Optimizing and Validating Classification Rules

SIMCA classification rules are geometric structures in descriptor space. They are (for 3 or more variables) a sphere or hypersphere for $A = 0$ (A is the number of components or product terms in Eq. 1), a cylinder or hypercylinder for $A = 2$, and a parallelepiped or hyperparallelepiped for $A = 3$ or more. Attempts to improve classification by adjusting the SIMCA classification rules have been limited in number. An early report by Forina and Lanteri [21] suggested that SIMCA models be modified to hyperellipsoids to classify Italian wines according to their region of origin. There seemed to be little improvement in classification results with this variation, however.

A more recent variation of the SIMCA models was more successful [22]. Even though the method was developed for application to mass spectral data, it is a general approach which can be applied to any type of data to improve classification results with SIMCA. In an effort to develop an automatic scheme for identification of members of a target list of five classes of airborne environmental pollutants, based on their mass spectral data, it was observed that SIMCA worked well. Also, an important aspect of environmental analysis is the detection of non-target compounds, as these may become important later. At present, no effort for this undertaking has been made. Thus, one objetive of the study was to obtain a class assignment for an unknown mass spectrum, if it was not a member of the target list. The mass spectra were converted in to their autocorrelated transformed spectra. SIMCA rules were derived and variables were deleted using modeling power, MPOW [1]. Variable selection was done so that the analysis of each class was performed on the same subset of variables. Using this strategy, classification results were 99% when applied to the training set data and the results were verified by visual interpretation. Classification accuracies were considerably diminished, however, when the rules were applied to true unknown spectra.

The SIMCA classification rule, shown in Fig. 7, determines class membership by the orthogonal projection distance, d, of the unknown to the class models. In the case where the unknown is beyond the class window, as determined by the extreme principal component scores for the training data, the distance, d^*, is calculated from the unknown to the edge of the classes. The distances, d and d^*, when corrected for differences in degrees of freedom, can be directly compared with the class residual standard deviation, s_{eps}, as defined in Eq. (5). The e_{ik} are those given by Eq. (1):

$$s_{\text{eps}} = \left[\frac{1}{(N - A)(M - A - 1)} \sum_{i=1} \sum_{k=1} e_{ik}^2\right]^{1/2} \tag{5}$$

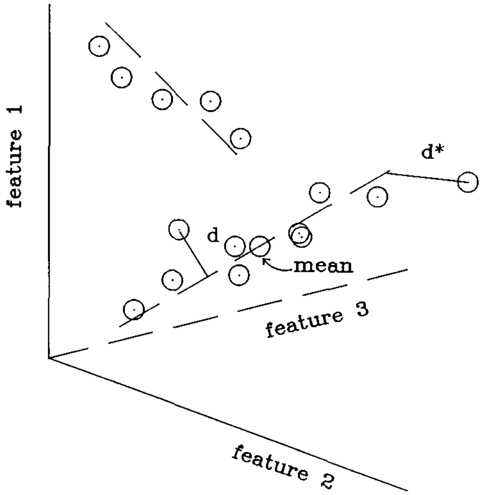


Figure 7. Geometric interpretation of the SIMCA classification rule.

If the unknown is "similar" to the class of training compounds, d or d^* will be approximately equal to s_{eps}. An approximate F-statistic can be calculated to determine the level of significance of similarity and, therefore, of the classification result. This similarity rule has been discussed elsewhere [23] and a variation of it has been proposed [24].

The original SIMCA classification rule gives equal weight to objects with projections near the extremes of the class and to those near the geometrical mean or centroid of the class. Principal component scores, $t_{i,a}$s, in Eq. (1), are the positions of the compounds in the models. In Fig. 7, the distances of two objects to the same class model are compared and under the usual classification rule, they are equidistant from the model. SIMCA, thus, gives them the same classification result.

However, the ts also contain information about class assignment. In order to have the ts considered in the classification rule, a variation of the usual SIMCA classification rule was proposed and used in this study.

PARC or feature space can be divided into two subspaces [24]. The subspace defined by the ps, the loading vectors in Eq. (2), is the inside model space, or IMS. The remaining axes are referred to as the outside model space axes, or OMS axes. The root-mean-square variance, s_a, along each p_a vector is given in Eq. (6). Here, $t_{i,a}$ is the principal component score for a compound (***i***) in component a, which measures the distance from the center of the class model to the point of projection of the object onto the class model. N is the number of compounds in the class. Thus, s_a is the standard deviation of the of the ts along axis, ***a***:

$$s_a = \left[\frac{1}{(N-a)}\sum_i t_{ia}^2\right]^{1/2} \tag{6}$$

The remaining root-mean-square variance, of OMS distance, is s_{eps} from Eq. (5). The summation is taken over both compounds and variables.

In the revised SIMCA classification study, class models are derived, and unknowns are then fitted to the various class models. An unknown compound, when fitted to an A-component model, will have scores, $t_{i1}, t_{i2}, \ldots t_{i,A}$ and an OMS distance, d_{OMS} of,

$$d_{\mathrm{OMS}} = \left[\frac{1}{(N-A)} \sum e_{ik}^2\right]^{1/2} \tag{7}$$

and an IMS distance, d_{IMS} of,

$$d_{\mathrm{IMS}} = \left[\frac{1}{A} \sum_a t_{ia}^2 \left(\frac{s_{\mathrm{eps}}}{s_a}\right)^{\beta}\right]^{1/2} \tag{8}$$

and a total distance:

$$d = \alpha d_{\mathrm{OMS}} + (1-\alpha)\, d_{\mathrm{IMS}} \tag{9}$$

The OMS distance is calculated as in SIMCA. If the unknown actually belongs to the class, whose model it is being fitted to, then $d_{\mathrm{OMS}} \cong s_{\mathrm{eps}}$. Otherwise, $d_{\mathrm{OMS}} \gg s_{\mathrm{eps}}$. The residuals are smallest when the unknown is fitted to its correct class. The IMS distance, d_{IMS}, is different from that calculated by the ordinary SIMCA. With ordinary SIMCA, a class window is defined by the class model. If the projection of the unknown spectrum lies within this window, then its d_{IMS} is set to zero. If its projection lies outside the class window, its s_{ia} is equal to the edge of the window. In practice, the projections of nearly all of the unknowns lie within the SIMCA class windows, hence, the $d_{\mathrm{IMS}} = 0$. Therefore, the ordinary d_{IMS} provides less than optimal class discrimination in SIMCA.

The modified d_{IMS} as given in Eq. 8, with $\alpha = 0.75$ and $\beta = 2$, has been found to be useful in improving class discrimination, particularly by reducing the number of false positive classifications [24]. In the classification step, a new distance d_n, for a compound is computed for each of the class models. This is shown in Fig. 8. The unknown is then assigned to the nearest class.

The revised SIMCA rule was an improvement, with a classification accuracy of unknowns of 221/230 (96%) compared to 209/230 (90%) for the regular SIMCA model. It actually was a poorer classifier of the training compounds, but gave only 4 false positives.

This revised SIMCA rule can be further adjusted with the parameters α and β. In this way SIMCA can be based on Mahalanobis distances ($\alpha = 0.5$, $\beta = 0$) and other variants. The "standard" SIMCA has $\alpha \approx 1$ and $\beta = 2$. A value of β less than 2.0, say 1.0, seems reasonable, since it gives more weight to the initial more important components in the classification rule. A value for α of between 0.5 and 1, say 0.75, also seems reasonable. Again, more experience must be obtained before any generalizations can be made.

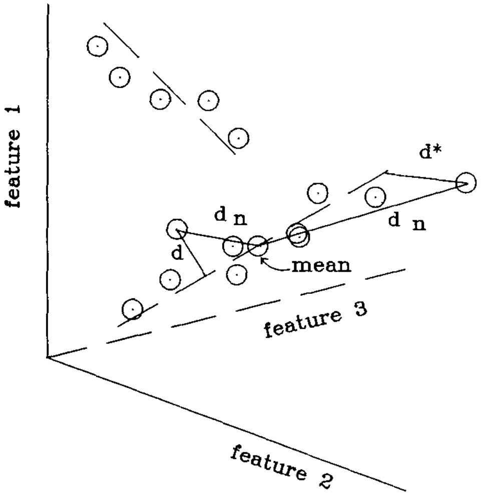


Figure 8. Geometric interpretation of the revised SIMCA classification rule.

4.3.9 Discussion

In QSAR, classification is a common problem due to the strong non-linearity of the interaction between chemical compounds and biological systems-receptors, membranes, enzymes, etc. Since most QSAR models are approximately linear, separating the compounds into distinct classes, each with a fairly linear behavior, is the best approach.

Among all available classification methods, e.g., linear discriminant analysis, quadratic discriminant analysis, ALLOC, UNEQ, K-nearest neighbors, etc., SIMCA is unique, in that it gives *models* of the classes. These models improve our understanding of the structural requirements for activity, etc., and are best interpreted graphically by score plots (plotting $t_{i,1}$ against $t_{i,2}$, etc.) for each class, loading plots of $p_{a,k}$, and so on (see, e.g. the PLS chapter).

The score plots give an indication of the data homogeneity in each class. If there are strong clusters in one of the score plots, this indicates that such a class should be further divided into subclasses.

The fact that SIMCA is based on principal components (PC) or PLS models makes it applicable also when the number of structural descriptor variables (K) is large compared to the number of compounds (N). With the masses of variables derived from quantitative molecular modeling, this becomes an important asset. Also, these PC and PLS models tolerate moderate amounts of missing data, which is often important in practice.

As in any modeling, SIMCA results must be *validated* before they are used for interpretation or prediction. Cross-validation, randomized training sets, and external prediction sets are available approaches, as discussed further in the chapter concerning validation in this volume.

References

[1] Wold, S., *Pattern Recognition* **8**, 127–136 (1976)

[2] Dunn III, W. J. and Wold, S., Pattern Recognition Techniques in Drug Design. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Ramsden, C. A., ed., Pergamon Press, Oxford (1990)

[3] Hansch, C., *Acc. Chem. Res.* **2**, 232–239 (1969)

[4] Albano, C., Dunn III, W. J., Edlund, U., Johansson, E., Norden, B., Sjöström, M., and Wold, S., *Anal. Chim. Acta* **103**, 429–441 (1978)

[5] Wold, S., Ruhe, A., Wold, H., and Dunn III, W. J., *SIAM J. Sci. Stat. Comput.* **5**, 735–743 (1984)

[6] Wold, S., *Technometrics* **20**, 397–405 (1978)

[7] Cramer III, R. D., Patterson, D. E., and Bunce, J. D., *J. Am. Chem. Soc.* **110**, 5959–5967 (1989)

[8] Eliel, E. L., Allinger, N. L., Angyal, S. J., and Morrison, G. A., *Conformational Analysis*, American Chemical Society, Washington, D.C., 1981

[9] Burke, B. A., *Developments in Molecular Shape Analysis to Establish Spatial Similarity among Flexible Molecules*, Ph.D. Thesis, University of Illinois at Chicago, Chicago, 1993

[10] Hopfinger, A. J., *J. Med. Chem.* **26**, 990–998 (1983)

[11] Collantes, E. R., *Novel Structure-Based Descriptors for Peptide Quantitative Structure-Activity Relationships*, Ph.D. Thesis, University of Illinois at Chicago, Chicago, Ill., USA, 1994

[12] Hellberg, S., Sjöström, M., Skagerberg, B., and Wold, S., *J. Med. Chem.* **30**, 1126–1135 (1987)

[13] Dunn III, W. J., Koehler, M. G., and Grigoras, S., *J. Med. Chem.* **30**, 1121–1126 (1987)

[14] Koehler, M. G., Grigoras, S., and Dunn III, W. J., *QSAR* **7**, 150–159 (1988)

[15] Hellberg, S., Ericsson, L., Jonsson, J., Lindgren, F., Sjöström, M., Skagerberg, B., Wold, S., and Andrews, P., *Int. J. Peptide Protein Res.* **37**, 414–424 (1991)

[16] Dunn III, W. J., and Wold, W., *J. Med. Chem.* **23**, 595–597 (1980)

[17] Rose, V. S., Wood, J., and MacFie, H. J. H., *QSAR* **11**, 492–504 (1992)

[18] Miyashsita, Y., Li, Z., and Sasaki, S., *Trends Anal. Chem.* **12**, 50–60 (1993)

[19] Hansch, C., and Fujita, T., *J. Am. Chem. Soc.* **86**, 1616–1623 (1964)

[20] Marsili, M., *Tetrahedron Comput. Method.* **1**, 71–79 (1988)

[21] Forina, M., and Lanteri, S., Data Analysis in Food Chemistry. In: *Chemometrics: Mathematics and Statistics in Chemistry*, Kowalski, B., ed., D. Reidel, Dortrecht, 305–350 (1983)

[22] Dunn III, W. J., Emery, S. L., Glen, W. G., Scott, D. R., *Environ. Sci. Technol.* **23**, 1499–1505 (1989)

[23] van der Voet, H., and Doornbos, D. A., *Anal. Chim. Acta* **161**, 115–123 (1984)

[24] Dunn III, W. J., Koehler, M. G., Emery, S. L., and Scott, D. R., *J. Chemom. Intell. Lab. Syst.* **1**, 321–329 (1987)

4.4 PLS for Multivariate Linear Modeling

Svante Wold

Abbreviations

AA	Amino acid
CV	Cross-validation
DMod$X$	Distance to model in X space
LV	Latent variable
MLR	Multiple linear regression
NN	Neural networks
PCA	Principal components analysis
PCR	Principal components regression
PLS	Partial Least Squares Projections to Latent Structures
PRESS	Predictive residual sum of squares
QMM	Quantitative Molecular Modeling
QSAR	Quantitative structure-activity relationship
RSD	Residual *SD*
SD	Standard deviation
SS	Sum of squares
VIP	Variable influence on projection
*	Multiplication
a	Index of components (model dimensions); ($a = 1, 2, \ldots, A$)
i	Index of objects (molecules); ($i = 1, 2, \ldots, N$)
k	Index of X variables ($k = 1, 2, \ldots, K$)
m	Index of Y variables ($m = 1, 2, \ldots, M$)
$\boldsymbol{X}$	Matrix of structure descriptors, size ($N * K$)
$\boldsymbol{Y}$	Matrix of activity variables, size ($N * M$)
$\boldsymbol{Z}'$	The transpose of a matrix $\boldsymbol{Z}$
$\boldsymbol{b}_m$	Regression coefficient vector of the mth y. Size ($K * 1$)
$\boldsymbol{B}$	Matrix of regression coefficients of all Ys. Size ($K * M$)
c_a	PLS Y weights of component $\boldsymbol{a}$
$\boldsymbol{C}$	The ($M * A$) Y-weights matrix; c_a are columns in this matrix
$\boldsymbol{E}$	The ($N * K$) matrix of X residuals
$\boldsymbol{f}_m$	Residuals of mth y variable; ($N * 1$) vector
$\boldsymbol{F}$	The ($N * M$) matrix of Y residuals
$\boldsymbol{p}_a$	PLS X loadings vector of component $\boldsymbol{a}$
$\boldsymbol{P}$	Loadings matrix; $\boldsymbol{p}_a$ are columns of $\boldsymbol{P}$
R^2	Multiple correlation coeofficient; the amount of Y "explained"
Q^2	Cross-validated R^2; the amount of Y "predicted"

t_a	X scores of component $\boldsymbol{a}$
$\boldsymbol{T}$	Score matrix ($N * A$), where the columns are $\boldsymbol{t_a}$
u_a	Y scores of component $\boldsymbol{a}$
$\boldsymbol{U}$	Score matrix ($N * A$), where the columns are $\boldsymbol{u_a}$
w_a	PLS weights of component $\boldsymbol{a}$
$\boldsymbol{W}$	The ($K * A$) X weights matrix; w_a are columns in this matrix
w_a^*	PLS weights transformed to be independent between components
$\boldsymbol{W}^*$	($K * A$) matrix of transformed PLS weights; w_a^* are columns in $\boldsymbol{W}^*$

Notation

Vectors are denoted by lower case characters and are column vectors, unless otherwise shown transposed (e.g. $\boldsymbol{v}'$). Matrices are denoted by upper case characters, e.g. X.

4.4.1 Introduction

QSAR is an approach to understanding how structural variation affects the biological activity of a set or structural class of compounds. This approach is also useful for studying properties of chemical compounds other than their biological activity, e.g. solubility, retention times in various chromatographic systems or catalytic properties. Such applications are often called Quantitative structure-*property* relationships) (QSPR). As has already been set out in this volume, QSAR can be roughly divided in to the following steps:

1. Problem formulation, i.e. selection of the biological activities of interest, choice of structural domain (structural class) and the choice of structural features to be varied,
2. quantitative description of the structural variation,
3. choice of model for the QSAR, i.e. either a linear, quadratic polynomial, hyperbolic or exponential model, etc.,
4. selection of compounds (series design),
5. synthesis and biological testing,
6. data analysis, and validation,
7. interpretation of results,
8. proposal of new compounds.

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

Although QSAR can not really be separated into several distinct steps, we shall nevertheless adhere to this breakdown and be concerned here mainly with Steps 6 to 8, i.e. data analysis, validation, interpretation, and use of the results obtained. Some of the consequences of the newer methods of data analysis in Steps 1 to 5 will also be discussed, however.

PLS (Partial Least Squares projections to latent structures) is a generalization of regression, which has been recently developed [1–6]. PLS is of particular interest in QSAR because, unlike Multiple Linear Regression (MLR), data with strongly correlated (collinear) and/or noisy or numerous X variables (structural descriptors) can be analyzed, and several activitiy variables, Y, i.e. *profiles* of activity, can be modeled simultaneously.

Being a generalization of MLR, PLS contains MLR as a special case when a MLR solution exists, i.e. when the number of X and Y variables is fairly small. This will be shown in the example below, where it will be seen that in such cases PLS gives a "reduced" solution, which is statistically more robust than the MLR solution, and hence, gives better predictions than MLR. PLS gives results analogous to MLR, such as PLS regression coefficients, Y residuals, R^2, and cross-validated R^2 (denoted here as Q^2). PLS, in addition, gives a set of plots (scores and loadings) that provide information about the correlation structures of the variables and the structural similarities/dissimilarities between the compounds. These plots are most useful for interpreting the model.

A recent development in QSAR is "Quantification of Molecular Modeling" (QMM) with methods such as CoMFA [7] and (GRID) [8]. With QMM, the number of X variables is large, often exceeding 10000, while the number of compounds is still moderate, for instance, between 10 and 100. PLS is a suitable tool for data analysis in QMM as discussed in the next volume of this series. Thus, being able to handle numerous collinear X variables, and activity profiles (Y), PLS allows us to investigate more complex and interesting structure-activity problems than previously, and to analyze the available data in a more realistic way. However, PLS still warrants some caution and we are still far from a good understanding of how molecular structure influences biological activity. Multivariate analysis methods such as PLS, principal component analysis (PCA), correspondence factor analysis (CFA), linear discriminant analysis (LDA) and neural networks (NN) are still in their in fancy, particularly in applications where there are many variables and few observations (in this case compounds).

4.4.2 Objectives and Data Homogeneity

Data analysis is very much like chemical analysis: one must know what one is looking for in order to select an appropriate analytical method, and a given problem can be solved by a variety of methods. Moreover, for a given problem, not all of the data is of interest, just as a chemical sample contains constituents of little interest. We shall refer to the uninteresting parts of the data as *noise*, and the data of interest as *information.* In this context, we must remember that noise is only partly random, but also systematic due to inadequacies in X (the structure descriptors) as well as to deficiencies in the model. To illustrate this, QSAR models are often linear, while the real world always is non-linear but in a way that is usually not well known.

In order to analyze data, one must have a specific *objective* or *aim*, which can be rather vague, such as obtaining an "overview" of a data set, or more specific, such as finding the relationship between given sets of variables, X and Y, which must be then elaborated. The objective is then translated in to a *model*, taking into account the expected relationships between variables, and the type of noise. The first part of the data analysis then consists of using the data to determine values of *parameters* in the model so that the model fits the data well.

Data analysis, as in any scientific investigation, is based on an assumption of *homogeneity*. In the present context of QSAR, this means a similarity in the biological mechanism with all the investigated compounds, which in turn, corresponds to having some limits on structural variability and diversity. These limits may be wide ranging if the biological activity is not specific such as anaesthetic activity, or the limits may be narrow, if the biological activity involves binding to a structurally well-defined receptor.

Since the results of the analysis depend on that, among other things, those critical assumptions concerning model shape and data homogeneity are fulfilled, it is essential that the analysis provides *diagnostics* about how well these assumptions indeed are, fulfilled. Much of the recent progress in applied statistics has concerned diagnostics [9], and many of these diagnostics can be also used in PLS modeling, as discussed below. PLS also provides additional diagnostics outside of regression-like methods, particularly those which are based on modeling X (score and loading plots and X residuals).

In the example below, the first PLS analysis indeed indicates that the data set analyzed is inhomogeneous: three aromatic amino acids (AAs) exert a different *type* of effect on the modeled activity in comparison to the other amino acids. This type of information is difficult to obtain in ordinary regression modeling, or indeed in most data analysis methods used in the QSAR field.

In fact, PLS can be used for classification (pattern recognition, discriminant analysis), similar to the Soft Independent Modeling Class Analogy method (SIMCA) which is based on disjoint principle component (PC) models of each class. If response data (y) also exist within each class, a PLS model instead of a PC model can be used. An example is given in the chapter on SIMCA by Dunn and Wold (Chap. 4.3).

An Example

In order to illustrate PLS modeling and the interpretation of the results, we shall use a small example form the literature with one Y variable and seven X variables. The example chosen is simple, and yet illustrative of most aspects of PLS and regression modeling. It must be emphasized however, that the present analysis is in no way a criticism of the work by El Tayar et al. [10], who carried out another type of analysis with the aim of obtaining a more detailed molecular interpretation.

Table 1. Raw data for the AA example. The lower half of the table shows the pairwise correlation coefficients of the data. *PIE* and *PIF* are the lipophilicity constant of the AA side chain according to El Tayar [10], and Fauchère and Pliska, respectively, *DGR* is the free energy of transfer of an AA side chain from the protein interior into water according to Radzicka and Woldenden. *SAC* is the water-accessible surface area of AAs calculated by *MOLSV*, *MR* is the molecular refractivity (Daylight data base), *Lam* is a polarity parameter according to El Tayar [10]. *Vol* is the molecular volume of the AAs calculated by *MOSLV*. All the data, except *MR*, were taken from the data reported by El Tayar et al. [10]

	PIE	*PIF*	*DGR*	*SAC*	*MR*	*Lam*	*Vol*	*DDGTS*
1 Ala	0.23	0.31	−0.55	254.2	2.126	−0.02	82.2	8.5
2 Asn	−0.48	−0.60	0.51	303.6	2.994	−1.24	112.3	8.2
3 Asp	−0.61	−0.77	1.20	287.9	2.994	−1.08	103.7	8.5
4 Cys	0.45	1.54	−1.40	282.9	2.933	−0.11	99.1	11.0
5 Gln	−0.11	−0.22	0.29	335.0	3.458	−1.19	127.5	6.3
6 Glu	−0.51	−0.64	0.76	311.6	3.243	−1.43	120.5	8.8
7 Gly	0.00	0.00	0.00	224.9	1.662	0.03	65.0	7.1
8 His	0.15	0.13	−0.25	337.2	3.856	−1.06	140.6	10.1
9 Ile	1.20	1.80	−2.10	322.6	3.350	0.04	131.7	16.8
10 Leu	1.28	1.70	−2.00	324.0	3.518	0.12	131.5	15.0
11 Lys	−0.77	−0.99	0.78	336.6	2.933	−2.26	144.3	7.9
12 Met	0.90	1.23	−1.60	336.3	3.860	−0.33	132.3	13.3
13 Phe	1.56	1.79	−2.60	366.1	4.638	−0.05	155.8	11.2
14 Pro	0.38	0.49	−1.50	288.5	2.876	−0.31	106.7	8.2
15 Ser	0.00	−0.04	0.09	266.7	2.279	−0.40	88.5	7.4
16 Thr	0.17	0.26	−0.58	283.9	2.743	−0.53	105.3	8.8
17 Trp	1.85	2.25	−2.70	401.8	5.755	−0.31	185.9	9.9
18 Tyr	0.89	0.96	−1.70	377.8	4.791	−0.84	162.7	8.8
19 Val	0.71	1.22	−1.60	295.1	3.054	−0.13	115.6	12.0
PIE	1.000	0.967	−0.970	0.518	0.650	0.704	0.533	0.645
PIF	0.967	1.000	−0.968	0.416	0.555	0.750	0.433	0.711
DGR	−0.970	−0.968	1.000	−0.463	−0.582	−0.704	−0.484	−0.648
SAC	0.518	0.416	−0.463	1.000	0.955	−0.230	0.991	0.268
MR	0.650	0.555	−0.582	0.955	1.000	−0.027	0.945	0.290
Lam	0.704	0.750	−0.704	−0.230	−0.027	1.000	−0.221	0.499
Vol	0.533	0.433	−0.484	0.991	0.945	−0.221	1.000	0.300
DDGT	0.645	0.711	−0.648	0.268	0.290	0.499	0.300	1.000

The data in question concerns modeling the energy for unfolding a protein (tryptophane synthase *alpha* unit of bacteriophage T4 lysozome), where each of the 19 coded amino acids (AAS), except arginine (Arg), had been introduced into position 49 [10]. The AAs are described in terms of $x_1 = PIE$ and $x_2 = PIF$ (two measures of side chain lipophilicity), x_3 = DGR = ΔG of transfer from protein interior to water, $x_4 = SAC$ = surface area, $x_5 = MR$ = molecular refractivity, $x_6 = LAM$ = side chain polaritiy, and $x_7 = Vol$ = molecular volume. Computational and other details are given by El Tayar et al. [10]. The X data are highly correlated, with $\boldsymbol{r^2(x1, x2, x3) > 0.964}$, and $\boldsymbol{r^2(x4, x5, x7) > 0.945}$. The raw data are given in Table 1 together with their correlation coefficients. In summary, the data comprise one activity variable (y) and seven correlated structure descriptors, $\boldsymbol{x_1}$ to $\boldsymbol{x_7}$, for 19 coded AAs. The individual correlations between the response (*DDGTS*) and each of the X variables were between 0.268 ($x_4 = SAC$) and 0.711 ($x_2 = PIF$).

4.4.3 The QSAR model

Any scientific model consists of several stages, starting with the philosophical viewpoint, conceptualization to the execution. All aspects are essential in order to comprehend the model and its underlying concepts.

4.4.3.1 The Conceptual Model

Our way of thinking in chemistry consists of translating the influence of structure on activity in terms of "*effects*", such as lipophilic, steric, polar, hydrogen bonding, and possibly other "effects". Some of these can be "localized" to a part of a molecule, for instance, a part that fits into a "lipophilic pocket" of a receptor. They may also be "global", such as global lipophilicity that may be related to the transport of the compound across lipophilic/polar boundaries. Much of the efforts in QSAR involves the translation of structural variation into reasonable scales, corresponding to these effects, both for localized parts of the molecules as well as for "global" whole molecules.

Although this formulation of how chemical structure influences biological and other properties of our molecules is of no immediate concern as regards to the technicalities of PLS, it is still of interest in that PLS modeling is consistent with seeing structural influences, mediated by "effects". The concept of *latent variables* in PLS may be seen as directly corresponding to these effects in QSAR-PLS models. In order to be able to estimate the influence, structure → "effects" → activity, each effect must be parametrized by at least one X variable (structural descriptor), preferably several. In simple situations, such as in the present example, with compounds having the same structural "backbone" and just changing substituents at specific "sites", the X variables are few- one X variable (substituent scale) for each "effect" and "site". In CoMFA and GRID paramerizations of more complicated sets of molecules, the X variables are numerous and the derivation of the "effects" is done as an integral part of the modeling and data analysis.

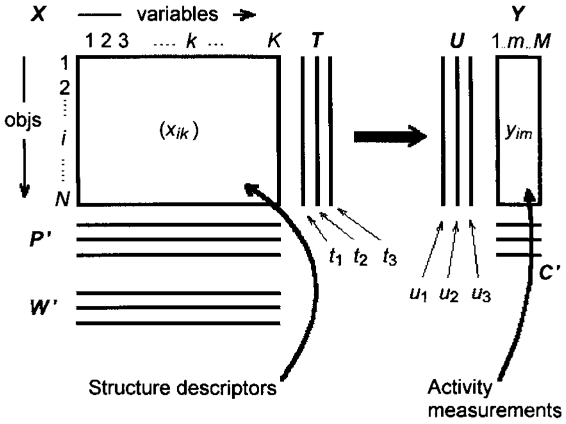


Figure 1. Data of a QSAR model can be arranged into two tables, matrices, X and Y. Note that the raw data may be transformed (e.g. logarithmically), and are usually centered and scaled before the analysis.

In the example, the side chains of the AAs are modeled using three lipophilicity parameters (*PIE*, *PIF*, and *DGR*), three steric parameters (*SAC*, *MR*, and *Vol*), and one polar parameter (*Lam*). Some are highly correlated (see Table 1).

Having translated the structural variation of N compounds in to a number of structure descriptor variables, as denoted by x_k ($k = 1, \ldots, K$), and measured the biological activity of these compounds by a number of variables, y_m ($m = 1, 2, \ldots, M$), we can collect the data into two matrices X and Y, of dimensions ($N \times K$) and ($N \times M$), respectively, as shown in Fig. 1. In this example $N = 19$, $K = 7$, and $M = 1$.

4.4.3.2 Transformation, Scaling and Centering

Before the analysis, the X and Y variables are often transformed, so that their distribution is consistent with chemical and biological theory. Thus, activity variables, with a range covering more than one order of magnitude of ten, are often logarithmically transformed, and the same applies to structure descriptor variables. If the variable has zero value, the fourth root transformation is a good alternative to the logarithm. The response variable in the example has alread been logarithmically transformed, i.e. expressed in thermodynamic units.

The results of projection methods, such as PLS, depend on the *scaling* of the data. This is an advantage, because with appropriate scaling, one can focus on the more important Y variables in the model, and use one's experience to increase the weights of informative X variables.

In the absence of knowing the relative importance of the variables, the standard PLS procedure consists in scaling each variable to unit variance, the so-called autoscaling. The software calculates the standard deviation (*SD*) of each colum of the data, and thereafter divides each column by the corresponding *SD*. This corresponds to assigning each variable with the same weight, and, thus, importance prior to the analysis.

In the example given, the autoscaled weights of the three lipophilicity variables and the three steric variables have been divided by 1.5, so that the single polarity variable does not become masked (so-called blockscaling).

In CoMFA and GRID-QSAR, however, autoscaling is often not the best method of scaling X, but non-scaled X data or some X-data, which has been subjected to same form of intermediate scaling between autoscaling and non-scaling, may still be appropriate. This has been discussed in detail recently [6].

For ease of interpretation and numerical stability, it is recommended that the data are *centered* before the analysis. This is done – either before or after scaling – by subtracting the column averages from all data in the X and Y columns. Hence, the analysis concerns the deviations from the means, and how these deviations are correlated. This centering does not lead to changes in the coefficients or weights of variables, and hence does not alter the interpretation of the results.

4.4.3.3 The PLS Model

The linear PLS model finds "new" variables, A **latent variables**, also called X scores and which are also denoted by $\boldsymbol{t}_a$ ($a = 1, 2, \ldots, A$). These scores are linear combinations of the original variables, x_k with "weights" of the coefficients, $\boldsymbol{w}^*_{ka}$ ($a = 1, 2, \ldots, A$).

$$t_{ia} = \Sigma_k w^*_{ka} x_{ik} \tag{1}$$

These X scores (t_as) have the following properties:

(a) They are good predictors of Y, so that

$$y_{im} = \Sigma_a c_{ma} t_{ia} + f_{im} \tag{2}$$

$$\boldsymbol{Y} = \boldsymbol{TC}' + \boldsymbol{F} \tag{2a}$$

In Eq. (2a) the model is expressed in matrix form. The residuals, $\boldsymbol{f}_{im}$ express the deviations between the observed and modeled data, and comprise the elements of the $\boldsymbol{Y}$ residual matrix, $\boldsymbol{F}$ in Eq. (2a). The index $\boldsymbol{i}$ is used for compounds, i.e., $i = 1, 2, \ldots, N$.

Because of the relationships expressed in Eqs. (1) and (2), the latter can be rewritten in the form of a regression model:

$$y_{im} = \Sigma_a c_{ma} \, \Sigma_k w^*_{ka} x_{ik} + f_{im} = \Sigma_k b_{mk} x_{ik} + f_{im} \tag{3}$$

The "PLS regression coefficients", $\boldsymbol{b}_{mk}$, can be written as:

$$b_{mk} = \Sigma_a c_{ma} w^*_{ka} \tag{4}$$

(b) They are few in number (A) and orthogonal; the summations in Eqs. (2) and (5) are made over the component index, $\boldsymbol{a}$ ($a = 1, 2, \ldots, A$).

(c) They are good "summaries" of X, so that the residuals, $\boldsymbol{e}_{ik}$, in Eq. (5) are "small":

$$x_{ik} = \Sigma_a t_{ia} p_{ak} + e_{ik} \tag{5}$$

$$\boldsymbol{X} = \boldsymbol{TP}' + \boldsymbol{E} \tag{5a}$$

Eq. (5a) is the X-model in matrix form.

With multivariate Y (when $M > 1$), the corresponding "Y scores" ($\boldsymbol{u}_a$) are good "summaries" of Y, so that the residuals, $\boldsymbol{g}_{im}$, in Eq. (6) are "small":

$$y_{im} = \Sigma_a u_{ia} c_{am} + g_{im} \tag{6}$$

$$\boldsymbol{Y} = \boldsymbol{UC}' + \boldsymbol{G} \tag{6a}$$

Eq. (6a) is the $\boldsymbol{Y}$-model in matrix form.

After each dimension, $\boldsymbol{a}$, the $\boldsymbol{X}$ matrix is "peeled off" by subtracting $\boldsymbol{t}_{ia} * \boldsymbol{p}_{ka}$ from the element $\boldsymbol{x}_{ik}$. This allows the PLS model to be expressed in weights, $\boldsymbol{w}_a$, with reference to the residuals according to the previous dimension, $\boldsymbol{E}_{a-1}$, instead of relating to the X variables themselves. Thus, instead of Eq. (1), we have Eq. (7):

$$t_{ia} = \Sigma_k w_{ka} e_{ik,a-1} \tag{7}$$

$$e_{ik,a-2} = e_{ik,a-1} - t_{i,a-1} p_{a-1,k}$$

$$e_{ik,0} = x_{ik}$$

However, the weights, $\boldsymbol{w}$, can be transformed in to $\boldsymbol{w}^*$, which directly relate to X, giving Eq. (1) above. The relationship between w and w^* is given by [6]:

$$\boldsymbol{W}^* = \boldsymbol{W}(\boldsymbol{P}'\boldsymbol{W})^{-1} \tag{8}$$

4.4.3.4 Interpretation of the PLS Model

One way of looking at PLS is that it forms "new x variables", $\boldsymbol{t}_a$, as linear combinations of the old ones, and then uses these new $\boldsymbol{t}$s as predictors of Y. Only as many new $\boldsymbol{t}$s are formed as are required to be predictively significant, and this is discussed below.

The parameters, $\boldsymbol{t}$, $\boldsymbol{u}$, $\boldsymbol{w}$ (and w^*), $\boldsymbol{p}$, and $\boldsymbol{c}$ (see Fig. 1), are determined by a PLS algorithm as described below. For the *interpretation* of the PLS model, the scores, $\boldsymbol{t}$ and $\boldsymbol{u}$, contain information about the compounds and their similarities/dissimilarities with respect to the given problem and the model.

The weights $\boldsymbol{w}$, (see below), and $\boldsymbol{c}$, provide information about how the variables can be combined to form a quantitative relation between X and Y. Hence, these weights are essential for understanding which X variables are important (numerically large $\boldsymbol{w}$ values), which X variables provide the same information (similar profiles of $\boldsymbol{w}_a$ values), the interpretation of the scores, $\boldsymbol{t}$, etc.

The part of data that are not explained by the model, that is, the **residuals**, are of diagnostic interest. Large residuals of Y indicate that the model is poor, and a normal probability plot of the residuals of a single Y variable are useful for identifying outliers, just as in MLR. In PLS residuals for X, the data not used in the modeling of Y, are also obtained. These X residuals are useful for identifying outliers in X-space, i.e. molecules with structures that are not consistent with the model.

Geometric Interpretation

PLS is a projection method and, thus, has a simple geometric interpretation with a projection of the $\boldsymbol{X}$ matrix (a swarm of points in a K-dimensional space) on to an A-dimensional hyperplane, in such a way that the coordinates of the projection ($\boldsymbol{t}_a$, $a = 1, 2, \ldots, A$) are good predictors of Y. This is indicated in Fig. 2, and the scores, $\boldsymbol{t}_a$ are explained below.

The direction of the plane is expressed in terms of the slope, $\boldsymbol{p}_{ak}$, of each principal direction of the plane (each component), with respect to each coordinate axis, $\boldsymbol{x}_k$. This slope is the cosine of the angle between the principal direction and the coordinate axis.

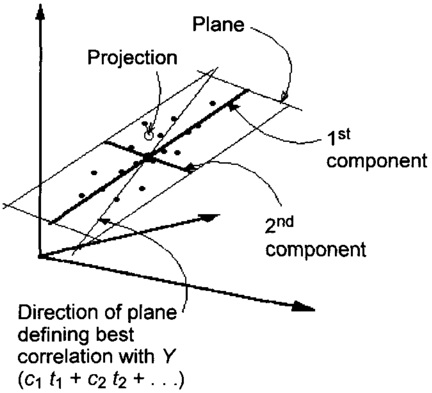

Plane
Projection
1st component
2nd component
Direction of plane defining best correlation with Y
($c_1 t_1 + c_2 t_2 + \ldots$)

Figure 2. The geometric representation of PLS. The X (structural descriptor) matrix can be represented as N points in K-dimensional space where each column of $X(x_k)$ defines a coordinate axis. The PLS model defines an A-dimensional hyperplane, which in turn, is defined by a line in one direction per component. The direction coefficient of the line is given by p_{ak}. The coordinates of each compound, i, when its structural descriptor data (row i in X) are projected down on to this plane, are given by t_{ia}. The corresponding positions are related to Y, so that projections of all points onto a line in this plane correlate with the values of Y.

Thus, PLS develops an A-dimensional hyperplane in X-space such that this plane is a good approximation of X (N points, row vectors of X), and at the same time, the positions of the projected data points onto this plane, described by the scores t_{ia}, are closely related to the values of the responses: activities, Y_{im} (see Fig. 2).

Incomplete X and Y Matrices (Missing Data)

Projection methods such as PLS tolerate certain amounts of missing data both in X (structural descriptors) and in Y (activities). In order to have missing Y data, the Y data must be multivariate, i.e. have at least two columns. The larger the matrices X and Y are, the higher the proportion of missing data can be tolerated. For the normal sizes of QSAR data, with around 20 compounds, 10 to 20% missing data can be tolerated, provided that they are not missing as the result of some systematic procedure.

The PLS algorithm, in principle, automatically accounts for the missing values by iteratively substituting the missing values with predictions given by the model. This corresponds to assigning the missing data with values that have zero residuals and, thus, have no influence on the model parameters.

One Y Variable at a Time, or all Y Variables in the Same Model?

PLS can model and analyze several Ys simultaneously, which has the advantage of providing a much simpler picture than if a separate model were employed for each Y. In general, when the Ys variables are correlated, it can be recommended that they be analyzed simultaneously. If, however, the Y variables really measure different activities and are fairly independent, one gains very little by analyzing them with in the same model. On the contrary, with uncorrelated Y

variables, the PLS model tends to have many components and may be difficult to interpret. Modeling the Y variables separately, thus, gives a set of simpler models with fewer dimensions, are hence, easier to interpret.

In order to ascertain whether the Y variables are correlated or not, it is recommended that the analysis is started with a separate PCA of the $\boldsymbol{Y}$ matrix. This will give information about the rank of $\boldsymbol{Y}$ in practice, i.e. the number of components in the PC model, A. If A is small compared to the number of Y variables ($\boldsymbol{M}$), and if we can interpret the resulting components, we can conclude that the Ys are correlated, and that a PLS model of all the Ys together in combination is warranted. Often, however, one finds from the PCA that the Ys cluster in to two or three groups according to the nature of the activity being measured. Thus, this is an indication that a separate PLS model for each group of Ys is warranted.

4.4.3.5 The Number of PLS Components, A

With any empirical modeling, it is essential to determine the correct complexity of the model. In the case of correlated X variables, there is a substantial risk of "overfitting", i.e. obtaining a well-fitted model, with little or no predictive capability. Hence, a strict test for the significance of each consecutive PLS component is necessary, and then stopping when components are non-significant.

Cross-validation (CV) is a practical and reliable method for testing this significance [2–6], which has become the standard in PLS analysis, and is incorporated in one form or another in all available PLS software. A good discussion of cross-validation was given recently by Wakeling and Morris [11], and Clark and Cramer [12]. In principle, CV is performed by dividing the data in to a number of groups, say, between five and nine, and then developing a number of parallel models from the reduced data with one of the groups omitted. It should be noted that having the number of CV groups equal to N, i.e., the so-called "leave-one-out" approach, is not recommended [13].

After developing a model, the omitted data is used as a test set, and differences between actual and predicted Y values are calculated for the test set. The sum of squares of these differences are computed and assembled from all the parallel models to form ***PRESS*** (Predictive Residual Sum of Squares), which is a measure of the prepdictive capability of the model.

When CV is used in the **sequential mode**, $PRESS_a/SS_{a-1}$ is evaluated for each component, and a component is considered to be significant if this ratio is smaller than around 0.9 for at least one of the y variables (sharper bonds can be obtained from the results of Wakeling and Morris [11]). Here SS_{a-1} denotes the (fitted) residual sum of squares *before* the current component (index a). The calculations continue until a component is found to be non-significant.

Alternatively, one can calculate *PRESS* for each component for up to say 10 or 15, for instance, and use the model which gives the lowest $PRESS/(N - A - 1)$. This "total" approach is computationally much more demanding, and is, therefore, used less often. Although it may in theory be preferable to the sequential approach, in practice, the difference seems to be small.

With both the sequential mode and the "total" mode, a *PRESS* is calculated for *the final model with the estimated number of significant components*. This is often expressed as Q^2 ("cross-validated R^2") which is $(1 - PRESS/SS)$ where *SS* in the sum of squares of *Y* corrected for the mean. This can be compared with $R^2 = (1 - RSS/SS)$, where *RSS* is the residual sum of squares. In models with several *Y* variables one also obtains R_m^2 and Q_m^2 for each *Y*-variable, y_m.

These measures, of course, can be also expressed as *RSD*s (Residual *SD*s) and *PRESD*s (Predictive Residual *SD*s). The latter is often called *SDEP* (standard Error of Prediction). If any knowledge of the noise in the system under investigation can be obtained, for example ±0.3 units for log $(1/C)$, these *SD*s should, of course, be similar in size to the noise.

4.4.3.6 Model Validation

Any model needs to be validated before it can be seriously used to "comprehend" or predict biological activity. It would seem that there are two reasonable principles of validation: validation based on predictions, and validation based on fitting to random numbers. The best validation method for a model is, of course, that which precisely predicts the activity of new compounds consistently. An independent validation set of several compounds (at least 4 or 5 with varying activity) is, however, a rare luxury.

In the absence of a real validation set, two interesting and powerful ways of model validation are available: cross-validation (CV) simulates how well the model predicts new data, and data randomization estimates the chances (probability) to of obtaining a good fit with randomly reorganized response data. CV has been described above, and will be discussed more in conjunction with randomization methods in the chapter on model validation (Ch. 5.1).

4.4.3.7 PLS Algorithms

The algorithm for calculating the PLS model is mainly of technical interest here, and we would just like to point out that several variants have been developed for different shapes of data [13, 14]. Most of these algorithms calculate one component at a time with cross-validation, testing the significance of each component. The calculations stop when a component is found to be insignificant. Most of these algorithms also allow for moderate amounts of missing data.

4.4.4 The first PLS Analysis of the AA Data

The first PLS analysis of autoscaled and centered AA data gives one significant component accounting for 43% of the *Y* variance ($R^2 = 0.435$, $Q^2 = 0.299$), with the second component being far from significant ($Q^2 = -0.130$). In contrast, MLR

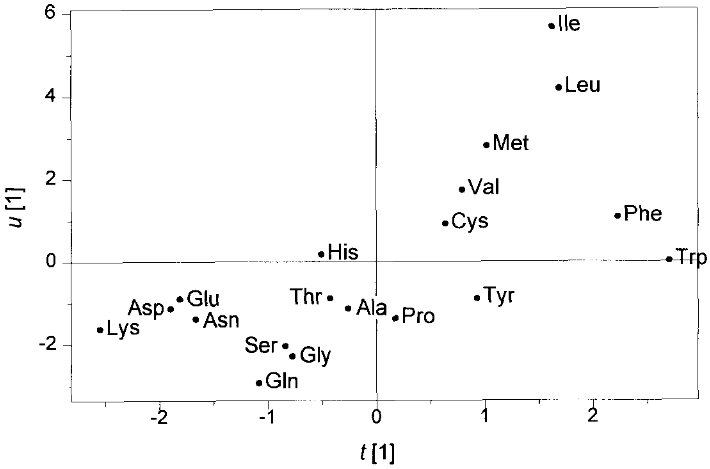


Figure 3. The PLS scores, u_1 and t_1, for the $N = 19$ AA example (first analysis).

gives an R^2 of 0.788, which apparently is a much better solution. This is equivalent to the PLS solution with $A = 7$ components. The full MLR solution, however, has a Q^2 of -0.215, indicating that the model is of low quality and cannot predict much better than chance. A MLR model with just x_2 and x_4 gives $R^2 = 0.507$ and $Q^2 = 0.248$ which is comparable with the two-component PLS model.

4.4.4.1 Score Plots

With just one PLS component, the only meaningful score plot is one of y (or of the equivalent, u_1) against t. This plot shows the correlation between $X(t)$ and Y, and is given in Fig. 3. We can see that the aromatic AAs, Trp, Phe and Tyr, show a much worse fit than the other amino acids. This is a clear indication of the lack of homogeneity in the data, which has a detrimental effect on the model.

4.4.4.2 PLS Weights *w* and *c*

For the interpretation of PLS models, the standard procedure is to plot the PLS weights, w, of one model dimension against another. Alternatively, one can plot the w^*s to give similar results and a similar interpretation.

The plot of w_1 versus w_2 values for the AA example is shown in Fig. 4. We see that the first dimension consists mainly of *PIF*, *PIE*, and *Lam* at the positive end, and *DGR* at the negative end. Considering the correlations between *PIF*, *PIE*, and *Lam*, this is not surprising; the first dimension is mainly a mixture of lipophilicity and polarity. The second insignificant dimension is mainly *MR*. The c values of the response, y, are proportional to the linear variation of Y explained by the corresponding dimension, i.e. $\sqrt{R^2}$. They define one point per response, and in the

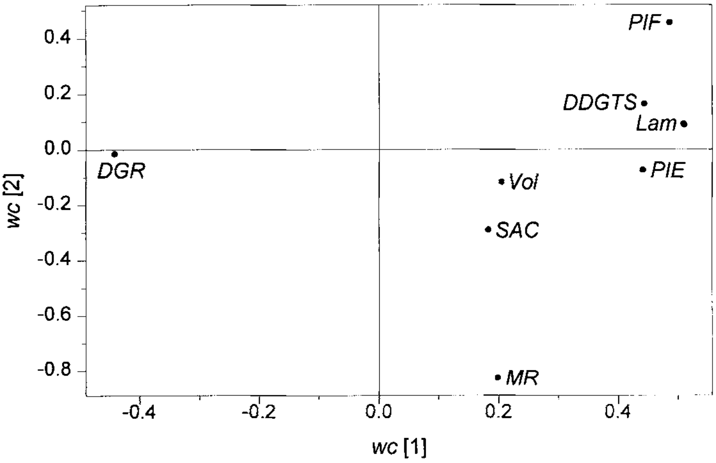


Figure 4. The PLS weights, w and c for the first two dimensions of the first AA model. The second dimension was found to be insignificant, but is included for plotting purposes.

example with a single response, this point (*DDGTS*) is situated in the far upper right-hand quartile of the plot.

The importance of a given X variable for a Y response is obtained by drawing a line from the response "point" in the (correctly scaled) plot through the origin (0, 0) and through to the other side of the axis of origin, see Fig. 5. We shall call this the Y line. A perpendicular line drawn from a X variable on to this line projects the X point onto the Y line. The length of the projection to (0, 0) is proportional to the importance of this X point for a particular Y point and is shown in Fig. 5

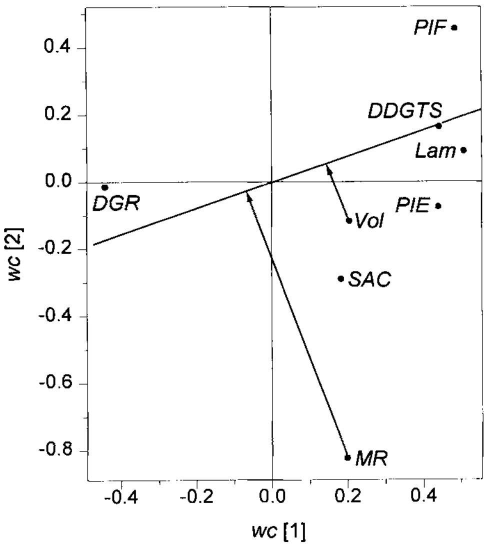


Figure 5. PLS weightings plot as in Fig. 4 with the line from Y = *DDGTS* going through the origin as shown in the plot, together with the projections of X = *MR* and X = *DGR* on to this Y line. The plot has been rescaled to the same length as both t axes.

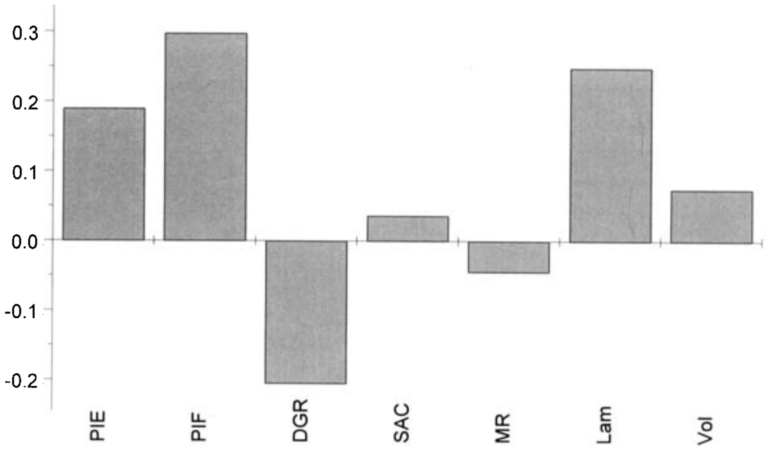


Figure 7. The PLS regression coefficients (**b**) from the AA example shown as functions of the number of components, A. The MLR coefficients are identical to those obtained when $A = 7$.

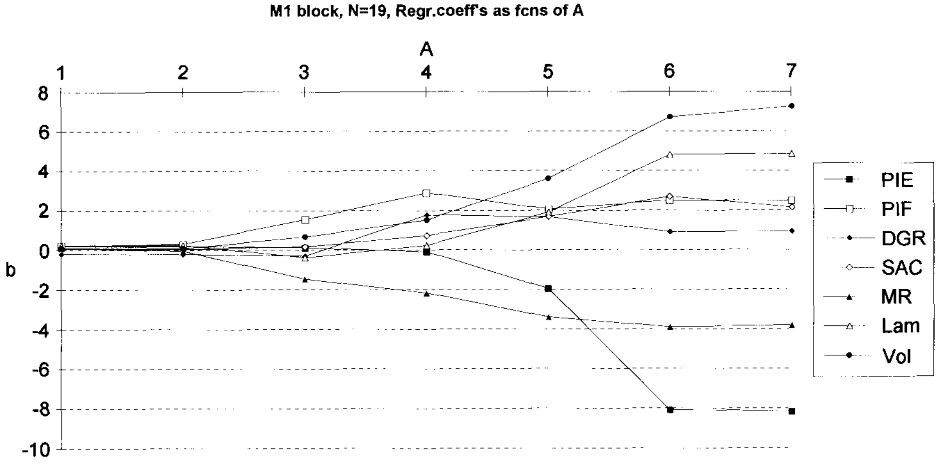


Figure 6. PLS regression coefficients for $A = 2$ components.

for $X = MR$ and $X = Vol$. We can see that MR has a slight negative influence and Vol has a somewhat greater positive influence on Y. These correspond closely to the PLS regression coefficients for $A = 2$ dimensions (Fig. 6).

4.4.4.3 A Comparison of PLS with Multiple Linear Regression (MLR)

In Fig. 7 we see how the PLS regression coefficients (b_{mk}) change when the number of components increase up to $A = 7$, when they are identical to the MLR coefficients. The coefficient of x_3 (DGR) **changes sign** between the PLS model ($A = 1$) and the MLR model ($A = 7$). Moreover, the coefficients of x_4, x_5, and x_7 which are almost zero in the PLS ($A = 1$) model, are large and have opposite signs in the MLR model ($A = 7$), although they are highly correlated to each other.

It is clear that the coefficients in the MLR model are misleading and difficult to interpret, and are very much as a result of the strong correlations between the X variables. PLS, however, stops at $A = 1$, and gives reasonable coefficient values both for $A = 1$ and $A = 2$. Due to the negative correlations between x_1, x_2 and x_3, their coefficients have the same values, but with opposite signs. It is essential to understand that with correlated variables, it is impossible to assign "correct" values to the coefficients, we can only estimate their *joint* contribution to y.

The usual approach taken with MLR and correlated X variables is to select a subset of variables that are not so well correlated. However, this can lead to a misinterpretation of the results, and one tends to forget the non-selected variables in the final model interpretation, although the are usually just as good as candidates for important variables as the ones already selected.

4.4.4.4 Conclusion of the First Analysis

Interpretation of the first round of results is that the PLS model is poor, with an R^2 of only 0.435. A tentative explanation might be that aromatic AAs are different from the other amino acids and that data are, thus, inhomogeneous. To investigate this, a second analysis was undertaken with a reduced data set, $N = 16$, without the aromatic AAs.

Alternatively, like El Tayer et al. [10]. we tried to include a quadratic term in the *Vol* parameter in the model. This gave a slightly better model with $A = 2$ and $R^2 = 0.684$, $Q^2 = 0.540$, but the score plots still indicated groupings in the data set (inhomogeneities), as shown in Fig. 8.

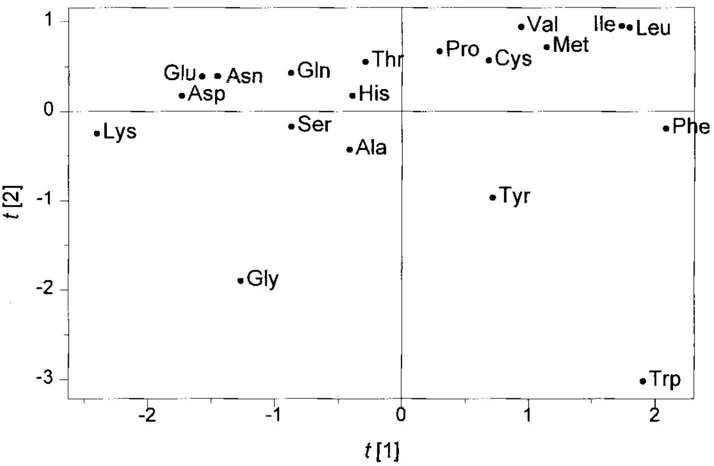


Figure 8. Score plot of t_1 vs t_2 for the PLS model of $N = 19$ AAs with a squared *Vol* term included. The large aromatic AAs are seen to deviate from the other amino acids, as well as from the very small (Gly).

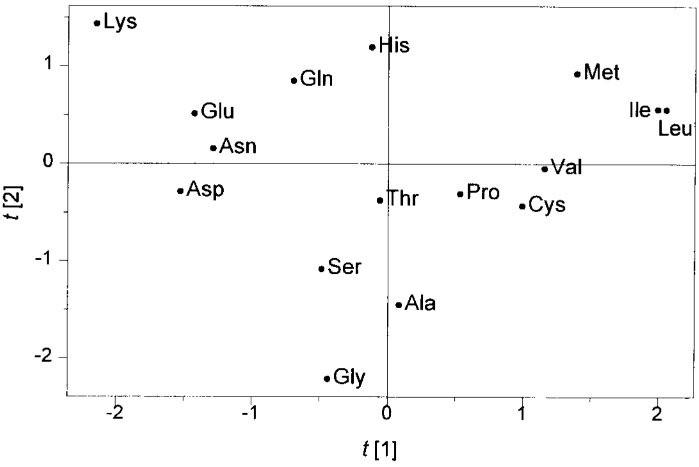


Figure 9. The PLS scores t_1 and t_2 for the $N = 16$ AA example (second analysis). The overlapping points in the right hand corner are Ile and Leu.

4.4.4.5 Conclusion of the Second Analysis

The modeling of $N = 16$ AAs with the same linear model as before produced a substantially better result with $A = 2$ significant components and $R^2 = 0.783$, $Q^2 = 0.706$. The MLR model for these 16 objects gave an R^2 of 0.872, and a Q^2 of 0.608. With only x_2 and x_4 included in the model, MLR gave $R^2 = 0.791$ and $Q^2 = 0.684$, which was very similar to the PLS model ($A = 2$). This marked improvement indicated that the data set was now, indeed, homogeneous and could be properly modeled.

The plot of the X scores (t_1 vs t_2, Fig. 9) shows the 16 amino acids grouped according to polarity from the upper left of the plot to the lower right side, and

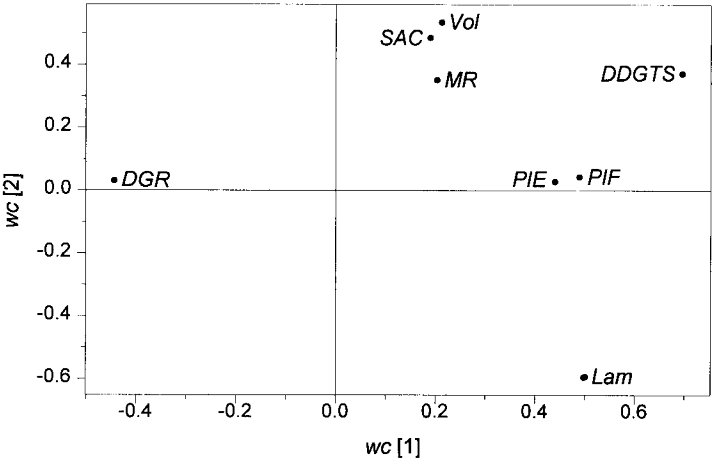


Figure 10. The PLS weights, w and c, for the first two dimensions of the second AA model.

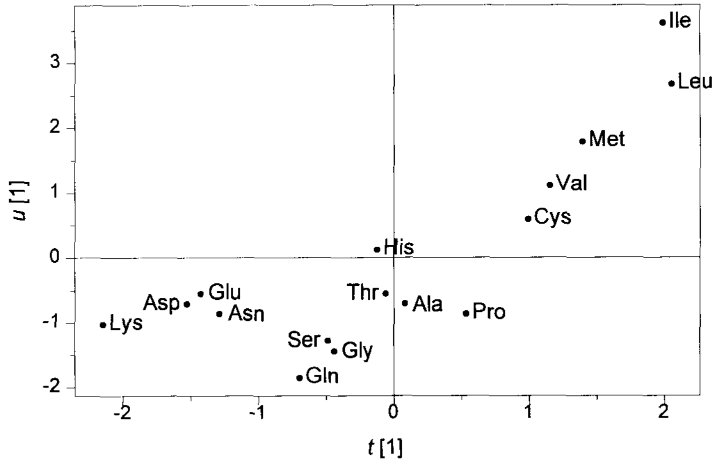


Figure 11. The PLS scores, u_1 and t_1, of the AA example, (second analysis).

according to size and lipophilicity within each grouping. This is consistent with the loading plot (Fig. 10), where we can see the first PLS dimension dominated by lipophilicity and polarity, and the second dimension being a mixture of size and polaritiy, and is similarly to the previous model.

The plot in Fig. 11 of $u_1(y)$ vs t_1 shows, however, a fairly strong curvature, indicating that squared terms in the lipophilicity parameter and, may be also in the polarity parameter, are warranted. In the final analysis, the squares of these four variables were included in the model, which indeed gave better results. Two significant PLS components and one additional component of borderline significance were obtained. The resulting R^2 and Q^2 values were 0.90 and 0.80 for $A = 2$, and for $A = 3$, 0.925 and 0.82, respectively. The $A = 3$ values corresponded to $RSD = 0.92$, and $PRESD$ $(SDEP) = 1.23$, since the SD of Y was 2.989. The full MLR model gave $R^2 = 0.967$, but with a much worse a Q^2 value of 0.09.

Finally, the model was tested with the parameters squared for the size (volume), lipophilicity and polarity descriptors for **all** the $N = 19$ compounds. This gave a PLS model with $A = 2$ or 3, and $R^2 = 0.79$ and $Q^2 = 0.47$, and still the same groupings in the score plots. Hence, it was concluded that the model of the above $N = 16$ AAs, with the parameters squared for the lipophilicity descriptors was the best one.

In order to obtain a picture of the relationship between Y and lipophilicity, polarity, and size of the AAs, a model with just one descriptor per class was developed, plus the lipophilicity parameter was squared. The variables with the largest regression coefficients in the final model were selected as representatives, i.e. $x_2 = PIF$, $x_6 = Lam$, $x_7 = Vol$, and $PIF\hat{}2$. This model was then used to show 3D plots as in Fig. 12.

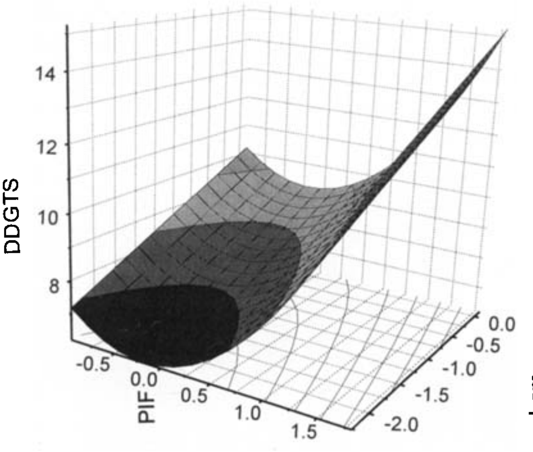


Figure 12. 3D plot of the response *DDGTS* (y) as a function of *PIF* and *Lam* with *Vol* fixed at its average value of 112.925.

4.4.5 Selection of Important Variables

In PLS modeling a variable (x_k) may be important for the modeling of Y. Such variables are identified by large PLS regression coefficients, $\boldsymbol{b}_{mk}$. However, a variable may also be important for the modeling of X, which is identified by large loadings, $\boldsymbol{p}_{ak}$. A summary of the importance of an X variable for *both* Y and X is given by $\boldsymbol{VIP}_k$ (variable importance for the projection, Fig. 13). This is a weighted sum of squares of the PLS weights, $\boldsymbol{w}_{ak}$, with the weights calculated from the amount of Y variance of each PLS component, $\boldsymbol{a}$.

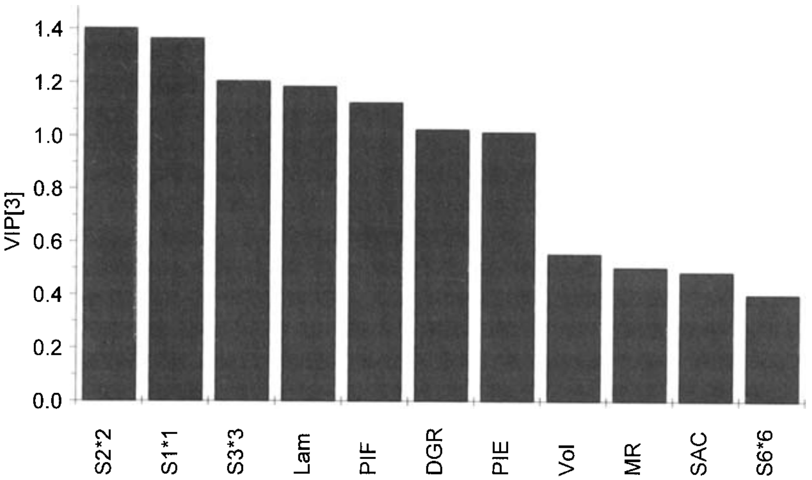


Figure 13. *VIP* of the X variables of the 3 component PLS model, (third analysis). The squares of x_1 = *PIE*, x_2 = *PIF*, x_3 = *DGR*, and x_6 = *Lam* are denoted by S1*1, S2*2, S3*3 and S6*6, respectively.

In data containing a large number of X variables, it is essential to select a subset of variables that really are important. The deletion of variables from the model which have **both** small PLS regression coefficients and small *VIP* values furnishes a pruned model with decent properties. More elaborate strategies, such as GOLPE, are described by Clementi in Chap. 2.3 in [16].

In the final model of the AA example, only one X variable has both small *VIP* values and small b values, namely the square term of *Lam* (the polar descriptor). When this is deleted, a PLS model of Y and the remaining X variables gave almost identical results as the model including *Lam*^2, and the results are, therefore, not shown.

4.4.6 Residuals

The residuals of Y and X are of diagnostic value in determing the quality of the model. A normal probability plot of the Y residuals (Fig. 14) of the final AA model shows a fairly straight line with all values within ±3 *SD*s. In order to be a serious outlier, a point must clearly deviate from this line and be outside of the limit of ±4 *SD*s.

Since there are many X residuals ($N * K$), one needs a summary for each compound in order not to be cluttered with unnecessary detail. This is provided by the residual *SD* of the X residuals of the compound. Because this *SD* is proportional to the distance between the point for the compound and the model plane in X space, it is also often called DMod$X$ (distance to the model in X-space). A DMod$X$ larger than around 2.5 times the overall *SD* of the X residuals

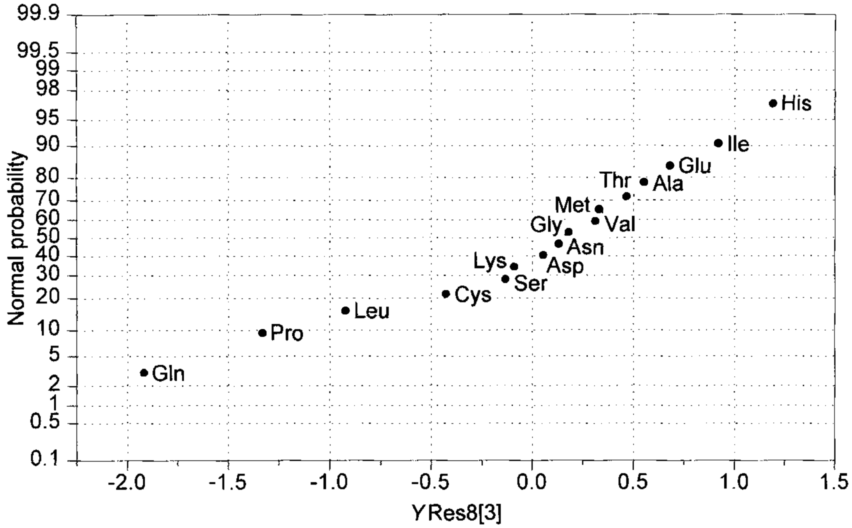


Figure 14. Y residuals of the 3 component PLS model in a normal probability plot (third analysis).

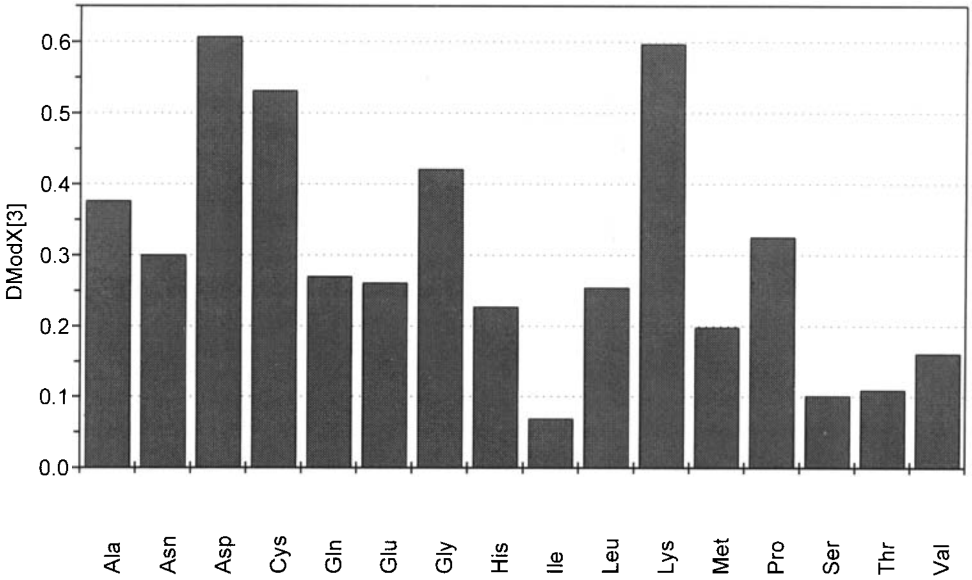


Figure 15. *RSD*s of the *X* residual (DMod*X*) for each compound (third analysis).

(corresponding to an *F*-value of 6.25) indicates that the compound is an outlier. Fig. 15 shows that none of the 19 compounds in the example given has a large DMod*X*. Here the overall *SD* = 0.34.

4.4.7 Conclusions

PLS analysis gave diagnostics (score plots) that indicated inhomogeneity in the data. This was confirmed by the much better model obtained for the $N = 16$ non-aromatic AAs. A remaining curvature in the score plot of u_1 vs t_1 led to the inclusion of squared terms, which gave a very good final model. Only the squared terms for the lipophilicity variables were found to be significant in the final model.

If additional aromatic AAs had been present, a second separate model could have been developed for this type of AAs providing an insight in to how this class differs from non-aromatic AAs. This in a way, corresponds to non-linear modeling; the changes in the relationship between structure and activity, when going from non-aromatic to aromatic AAs, are too large and too non-linear to be modeled by a linear or low degree polynomial model. The use of two separate models which do not directly model the change from one class to another provides a simple approach to deal with these non-linearities.

4.4.8 Summary How to Develop and Interpret a PLS model

1. One must have a good understanding of the given problem, in particular, which biological properties of interest are to be measured and modeled, and which structural features should be varied.
2. Good data, both Y (activity) and X (structural descriptors) must be obtained. Multivariate Y variables provide much more information, because they can first be analyzed separately by PCA. This gives a good idea about the amount of systematic variation in Y, and which Y variables should be analyzed in combination, etc.
3. At the beginning in PLS modeling, the first information obtained is the significant number of components, A, which is an indication of the complexity of the QSAR. This number of components gives the lower bound of the number of structural *effects* that are to be postulated in the system.
4. After obtaining A (see above), the second consideration is how well the model fits the data, i.e. the amount of Y variance (R^2) that is accounted for. If there are several Y variables, one can also obtain an R^2_m value for each Y variable. For each of these values, there is a corresponding Q^2 value ("cross-validated R^2"). The R^2 values give the upper bound of how well the model explains the data and predicts activities for new compounds, and the Q^2 values give the lower bounds for the model.

These parameters, of course, can be also expressed as *RSD*s and *PRESD*s (Predictive Residual *SD*s). If there is any knowledge of the noise in the system being investigated, for example, ±0.3 unitis for log $(1/C)$. These *SD*s, should of course, be similar to the order of magnitude of the noise.

5. The first two or three model dimensions in the score plots (u, t) should be investigated to highlight outliers, curvatures, groupings in the data, or any other problems.

Then, the score plots (t, t) – the windows in X space – should be inspected, again to look for indications of inhomogeneities, groupings, or other patterns. In conjunction with this the weightings plots (w, c) should be used to interpret the patterns and trends seen in the (t, t) plots.

6a. If problems are apparent, i.e. such as too small R^2 and Q^2 values, with outliers or groupings in the score plots, one should try to find a solution. First, plots of residuals (normal probability and DMod$X$ and DMod$Y$) may give an indication of the source of such problems.

Single outliers should be inspected to assess the accuracy of the data, and if this is of no use, be excluded from the analysis, but only if they are non-interesting (i.e., of low activity).

Curvature in plots (u, t) may be improved by including selected squared terms and/or cross-terms in the model.

Then, one returns to point 1 after either, possibly, having transformed the data, modified the model, divided the data into groups, deleted outliers or taken whatever action is warranted.

6b. If no problems are apparent, i.e. R^2 and Q^2 are of the correct magnitude, and the model can be interpreted, one should try to prune the model by deleting unimportant terms, i.e. small regression coefficients and low *VIP* values. Then, a final model is developed, which is interpreted, validated, and for which predictions are made, etc. For the interpretation the weight plots (w, c), coefficient plots, and contour or 3D plots with dominating X variables as the plot coordinates, are invaluable.

4.4.9 Conclusions and Discussion

PLS is an approach to quantitiative modeling of the often complicated relationships between chemical structure and biological activity which is more realistic than MLR, including stepwise selection variants. The reason is that the assumptions which underly PLS – correlations between the X variables, noise in X and model errors – are more in line with reality than the assumptions which underly regression of independent and error free X variables.

The diagnostics in PLS, notably cross-validation and score plots (u/t, and t/t) with the corresponding loading plots, provide information about model complexity and the structure of X data that can not be obtained with ordinary MLR. It will take time for the QSAR community to get used to this additional information and obtaining experience in how to interpret and use this information in QSARs. In particular, a fairly common result in PLS modeling is that the data are inhomogeneous (see the AA example given here), which is rarely observed in MLR. This is mainly because MLR lacks the diagnostic tools for highlighting inhomogeneities in the data. Consequently, there is still the common, but wrongly, hold view that one should always try to squeeze all the data into a single model. With the strong non-linearities that exist in complicated chemical-biological systems, it is warranted to use more than one model to obtain a more accurate picture; non-linearities are typical and sometimes so strong that a single polynomial model could not be constructed. Hence, a flexible approach to QSAR modeling with separate models for different structural classes of compounds is often required, there is no loss of information with this approach in comparison with the single model approach. A new compound is first classified with respect to its X values, and predicted activity values are then obtained by employing the appropriate class model.

A consequence of the greater flexibility and power of PLS in comparison with traditional (stepwise) MLR is that other aspects of the QSAR development are facilitated. Thus, the ability of PLS to analyze *profiles* of activity, makes it easier to devise activity measurements that are relevant to the stated objectives of the investigation; it is easier to assess biological activity by a series of measurements than by a single activity variable.

Similarly, the ability of PLS to handle many collinear structure descriptor variables (X) makes it easier to quantify the variation of structure between compounds, CoMFA and GRID are semi-automated approaches for the quantification of structural variation based on the calculation of thousands of descriptors.

And, finally, the possibility of graphical reprepsentation of PLS parameters and residuals makes it possible to interpret and use the results also in complicated models, thus making QSAR of interesting systems, such as peptides, proteins, nucleic acids, and polysaccharides, accessible to everybody.

Acknowledgements

Support from the Swedish Natural Science Research Council (NFR), and from the Center for Environmental Research in Umeå (CMF) is gratefully received. We would like to thank Dr. Erik Johansson for providing the example with the amino acids cited here.

References

[1] Wold, H., Soft Modeling. The Basic Design and Some Extensions. In: *Systems under Indirect Observation*, Vol. **II**, Jöreskog, K.-G., and Wold, H., eds., North-Holland, Amsterdam (1982)

[2] Wold, S., Ruhe, A., Wold, H., and Dunn III, W. J., *SIAM J. Sci. Stat. Comput.* **5**, 735–743 (1984)

[3] Hellberg, S., Sjöström, M., Skagerberg, B., and Wold, S., *J. Med. Chem.* **30**, 1126–1135 (1987)

[4] Höskuldsson, A., *J. Chemometrics* **2**, 211–228 (1988)

[5] Ståhle L., and Wold, S., *Multivariate Data Analysis and Experimental Design in Biomedical research* (Progress in Medical Chem. Vol **25**) Ellis, G. P., and West, G. B., eds., Elsevier Science, 1988

[6] Wold, S., Johansson, E., and Cocchi, M., PLS – *Partial Least Squares Projections to Latent Structures*. In: *3D QSAR in Drug Design, Theory, Methods, and Applications*, Kubinyi, H., ed., ESCOM Science Publishers, Leiden (1993)

[7] Cramer III, R. D., Patterson, D. E., and Bunse, J. D., *J. Amer. Chem. Soc.* **110**, 5959–5967 (1988)

[8] Goodford, P. J., *J. Med. Chem.* **28**, 849–857 (1985)

[9] Belsley, D. A., Kuh, E., and Welsch, R. E., *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, N. Y., 1980

[10] El Tayar, N. E., Tsai, R.-S., Carrupt, P.-A., and Testa, B., *J. Chem. Soc. Perkin Trans.* **2**, 79–84 (1992)

[11] Wakeling, I. N., and Morris, J. J., *J. Chemometr.* **7**, 291–304 (1993)

[12] Clark, M., and Cramer III, R. D., *Quant. Struct.-Act. Relat.* **12**, 137–145 (1993)

[13] Shao, J., *J. Amer. Stat.-Assoc.* **88**, 486–494 (1993)

[14] Lindgren, F., Geladi, P., and Wold, S., *J. Chemometr.* **7**, 45–59 (1993)

[15] Rännar, S., Geladi, P., Lindgren, F., and Wold, S., *J. Chemometr.* (1994), in press

[16] Van de Waterbeemd, H., ed., *Advanced Computer-Assisted Techniques in Drug Discovery* Methods and Principles in Medicinal Chemistry, Vol. **3**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1995

4.5 Spectral Mapping of Drug-Test Specificities

Paul J. Lewi

Abbreviations

APL	A Programming Language, Iverson's notation
apo	Apomorphine test
ATN	Apomorphine, tryptamine and norepinephrine test
α	*alpha*-adrenergic (norepinephrine) receptor
CFA	Correspondence factor analysis
cpz	Chlorpromazine
D	Dopamine receptor
G	Guanosine
hal	Haloperidol
5HT	Serotonin receptor
nep	Norepinephrine test
PCA	Principal components analysis
SMA	Spectral map analysis
SVD	Singular value decomposition
RC	Rows and columns

Symbols

α	Factor scaling coefficient for rows (compounds) in the context of factor analysis
β	Factor scaling coefficient for columns (tests)
c	Global variance of $\boldsymbol{Z}$
Λ	Diagonal matrix of singular values
$\delta_{kk'}$	Cronecker delta, 1 if $k = k'$ and 0 otherwise
ED_{50}	50 percent effective dose
f_1, f_2	Factor axes
γ	Accuracy of reconstruction of $\boldsymbol{Z}$ using r^* factors
IC_{50}	50 percent inhibitory concentration
i, i'	Indices for compounds
j, j'	Indices for tests
k	Constant
k, k'	Indices for factors
l_{jk}	Loading of test j onto factor k

λ_k^2	Contribution of factor k to the global variance, c
n	Number of compounds
p	Number of tests
π_i	Potency of compound i
r	Number of factors of $\boldsymbol{Z}$
r^*	Number of structural factors of $\boldsymbol{Z}$
s_{ik}	Score of compound i on factor k
S_n	n-dimensional coordinate space of compounds
S_p	n-dimensional coordinate space of tests
S_r	r-dimensional coordinate space of factors
σ_j	Sensitivity of test j
$\boldsymbol{U}$	Matrix of normalized scores, left singular vectors
$\boldsymbol{V}$	Matrix of normalized loadings, right singular vectors
w_i	Normalized weight coefficients for row (compound) i
w_j	Normalized weight coefficients for column (test) j
$\boldsymbol{X}$	Table of observed activities
$x_{..}$	Global mean of X over all compounds and tests
$\tilde{x}_{..}$	Geometric global mean of X over all compounds and tests
x_0	Base value for non-positive substitution
$x_{i.}$	Row mean of X for compound i
$\tilde{x}_{i.}$	Geometric row mean of X for compound i
x_{ij}	Activity of compound i in test j
x_{ij}^*	Small positive value substituting a non-positive value x_{ij}
$x_{.j}$	Column mean of X for test j
$\tilde{x}_{.j}$	Geometric column mean of X for test j
$\boldsymbol{Z}$	Table of transformed activities, specificities
$\lVert z_i \rVert$	Norm of row (compound) i of $\boldsymbol{Z}$
z_{ij}	Specificity between compound i and test j
$\lVert z_i - z_{i'} \rVert$	Contrast between rows (compounds) i, i' of $\boldsymbol{Z}$
$\lVert z_j \rVert$	Norm of column (test) j of $\boldsymbol{Z}$
$\lVert z_j - z_{j'} \rVert$	Contrast between columns (tests) j, j' of $\boldsymbol{Z}$

4.5.1 Activity, Potency, Sensitivity and Specificity

Spectral mapping is an unsupervised multivariate QSAR method. The term multivariate indicates that the method is applicable in the case when several compounds are studied simultaneously in multiple tests. The term "unsupervised" implies that the method does not rely on a specific model for structure-activity. In contrast, supervised methods, such as those based on regression and discriminant analysis, rely on a training set of compounds in order to determine the parameters of the model. Such a specific model is then used for the prediction of the results of newly synthesized compounds, usually within a series of homogeneous chemical structures. In Spectral Map Analysis (SMA), however, no such distinction between the training and prediction set is made, as the method is primarily designed for classification of heterogeneous compounds and for the discovery of structure within a battery of tests, rather than for prediction of biological activity or clinical effects (although it would be possible to use the method for this purpose if it were so required).

Spectral mapping is an exploratory method of analysis, which may help in raising relevant questions about the data, rather than in providing answers to specific questions. As such, it is to be regarded as a preliminary stage in the study of QSAR. The only requirement of the method is that the data are presented in the form of a rectangular table, for example, with rows referring to compounds and with columns denoting tests. (The assignment of rows to compounds and of columns to tests is arbitrary and can be interchanged if so desired.) Each element in the table then expresses the corresponding pharmacological activity of a compound that is produced in a particular test, or vice versa, the activity of a test with a particular compound. Note that the symmetry between compounds and tests is fundamental to the definition of spectral mapping, which is symmetrical with respect to compounds and tests. The numbers in a particular row of the table define the spectrum of the corresponding compound. The numbers in a particular column of the table constitute the spectrum of the corresponding test. From the point of view of an industrial pharmacologist, one may look at the table as describing each compound by means of its spectrum of activities that are produced in a battery of standard tests. From a more academic view point, one can also regard the table as describing each test by means of its spectrum of activities that are obtained in a set of reference compounds. The symmetry dictates that the roles of compounds and tests are interchangeable.

In this context, we define activity as the reciprocal of the effective dose (ED_{50}), e.g. in milligrams of substance per kilogram of bodyweight that is required to produce a stated effect, such as the inhibition of an induced pattern of behavior in half of the animals that received the dose. Spectral mapping, however, is invariant with respect to the units that are chosen for the individual tests. In another context, one may define activity as the reciprocal of the inhibitory concentration (IC_{50}), e.g. in nmoles that is required to inhibit a previously induced effect in half of the test specimens. In general, spectral mapping can be applied to data that are defined on ratio scales, i.e. data that allow for meaningful ratios between them. In particular, one may multiply any column of the table with any positive constant without affecting the result of the analysis. Hence, columns of the table may be defined in different units, although in this context we assume that the units are the same. Missing data are represented by their expected values. In the context of Spectral Map Analysis (SMA), and as will be explained later in Sec. 4.5.3, the expected value for the element at the intersection of a particular row and column of the table is defined as the product of the geometric means of the elements in the corresponding row and column, divided by the geometric global mean over all elements.

Potency of a compound is defined here as the geometric average of the activities produced in all available tests. Similarly, sensitivity of a test is the geometric average of the activities obtained from all available compounds. The potency of a compound and the sensitivity of a test are absolute quantities, which express a notion of size or elevation of a spectrum of activities. Specificity is a relative quantity which is related to the shape or "peakedness" of a spectrum of activities. Two compounds are said to be similar, when they have similarly shaped activity spectra, irrespective of the difference in potency of the two compounds. Two tests are termed as similar, when they also possess similarly shaped activity spectra, irrespective of the difference in sensitivity of the two tests. This means that two compounds are similar, when they have the same specificities for the various standard tests. By virtue of symmetry, two tests are similar when they exhibit the same specificities for the various reference compounds. Specificity is a bipolar (or differential) measure of association between a compound and a test. Specificity is positive, when the association is in the same direction, i.e. when the compound possesses more than an average specificity for the test (and vice versa). It is negative when the association is in opposite directions, i.e. when the compound has less than average specificity for the test (and vice versa). In the former case, the compound and test can be said to attract each other, while in the latter, they repel each other. Spectral mapping is directed towards the analysis of specificities between compounds and tests. In the following section we will define specificity in mathematical terms as the log of the observed activity minus the log potency of the compound and minus the log sensitivity of the test. The distinction between potency, sensitivity and specificity is similar to the one that is made in biology between the size and the shape of animals [1].

Spectral mapping is primarily a graphic method. As its name indicates, it provides a visual display, in the form of a map, of all the specificities between compounds and tests (positive or negative) that are contained in their activity spectra. In the next sections, we will provide a historical account, a case study, and a mathematical description of the method. Because of our background, our point of view is influenced by the design of new therapeutic compounds in a battery of standard tests, but because of the symmetry property, spectral mapping can also be equally applied to the development of a new test using a set of reference comounds.

4.5.2 Historical Background

Spectral mapping has been proposed by the author as a multivariate QSAR method in 1975. The design of the method was the result of favorable circumstances. It first came about at the research laboratory at Janssen in Beerse, which at that time had implemented a number of simple, but highly effective ideas for the statistical analysis of the results of its screening tests. The procedures that resulted from these ideas relied heavily on graphical displays, as will be shown later on. Secondly, the need for a multivariate analysis of the screening results appeared at an auspicious time, when major developments had occurred, notably the biplot graphic for Principal Component Analysis (PCA) by Gabriel [2] and the publication of Correspondence Factor Analysis (CFA) by Benzécri and a group of French data analysts and statisticians [3]. Finally, the availability of APL, a notation for interactive computation designed by Iverson [4], greatly facilitated the formulation and implementation of an alternative method, which embodies the ideas referred to above, but which is more appropriate to pharmacological data. Initially, this approach was called spectral mapping, but was renamed later as Spectral Map Analysis (SMA), when applied to other types of data, especially in the field of marketing (for a more competitive position in the market place) and finance (for performance evaluation). Eventually, the method proved to be fairly general and was applicable to a variety of data that can be produced in the form of a rectangular data table. How this came about is related below.

The starting point of Janssen's research program in 1953 was the search for synthetic opiates, anticholinergics and antihistaminics. In 1957 this led to the screening of a series of propiophenones for morphine-like analgesic activity. The prototype R951 has strong morphine-like properties and is shown in Fig. 1, adapted

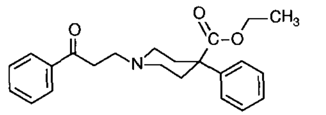
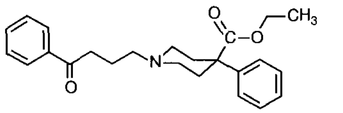
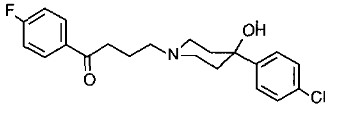
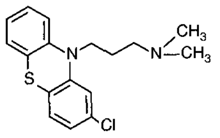
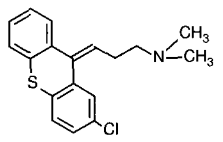
	R 951 morphine-like activity
	R 1187 morphine- and chlorpromazine-like activity
	haloperidol butyrophenone-class neuroleptic
	chlorpromazine phenothiazine-class neuroleptic
	chlorprothixene thioxanthene-class neuroleptic

Figure 1. Change of morphine-like activity of propiophenones into neuroleptic activity of butyrophenones, with haloperidol as the prototype. The structures of chlorpromazine and thioxanthene are shown for comparison (after van Wijngaarden [5]).

from a publication on the early history of the research within Janssen Pharmaceutica [5]. Subsequently, when the length of the alkyl chain between the phenone and the piperidine ring was increased from two to three carbons, this resulted in the formation of the butyrophenone, R1187. The lengthening of the alkyl chain caused a reduction of analgesic activity, but at the same time a new type of activity appeared. It was recognized as being similar to the effect produced by chlorpromazine, a phenothiazine derivative synthesized at Rhône-Poulenc and found by Delay et al. [6] in 1952, to possess antipsychotic properties. Chlorpromazine and related phenothiazines were called neuroleptics by Delay and represented a breakthrough in the treatment of mental illness. This serendipitous discovery at Janssen led in 1958 to the synthesis of the butyrophenone analogue haloperidol which was completely devoid of morphine-like activity and which turned out to be a highly potent and highly specific neuroleptic. Haloperidol is the prototype of the butyrophenone class of neuroleptics, which is structurally distinct from the phenothiazine class as shown in Fig. 1.

In 1961 several different chemical classes of neuroleptics had been discovered, including the thioxanthenes, with chlorprothixene as the prototype, and other minor classes in addition to the previously mentioned phenothiazines and butyrophenones. These neuroleptics were found to antagonize, to varying extents, the effects of apomorphine, amphetamine, tryptamine, norepinephrine and epinephrine (adrenalin) in rats. They also inhibited spontaneous and conditioned motility and produced catalepsy in various degrees [7]. It, thus, appeared as if each neuroleptic possessed a typical pharmacological activity spectrum, which would be analogous to possessing a particular light absorption spectrum. Compounds could, thus, be classified on the basis of their pharmacological activity spectra. Two compounds are thought to be similar if they possess similarly shaped activity spectra, irrespective of their potencies. The process of classification is similar to that of comparing the light absorption spectrum of an unknown compound with the spectra produced by a collection of reference compounds. Similarity of these spectra is assessed on the basis of their shape, irrespective of their average absorption, which is a function of the concentrations of the unknown and reference compounds. Such a continuous classification implies a spatial arrangement of the compounds, such that similar compounds appear close together and, such that dissimilar compounds are at a distance from one another. The arrangement is not necessarily one-dimensional. In fact, in 1961, the classification of the neuroleptics on the basis of the activity spectra was two-dimensional (Fig. 2). Haloperidol shows high specificity for the apomorphine and amphetamine tests; fluanisone, a derived butyrophenone, was found to be specifically active in the norepinephrine and epinephrine tests; floropipamide, also a butyrophenone, now called pipamperone, is specific for the tryptamine test. Regarding the three butyrophenones as vertices or poles of an equilateral triangle, it is possible, on the basis of their spectra, to visually classify chlorpromazine between fluanisone and floropipamide, and to position the phenothiazine derivative, perphenazine, between haloperidol and fluanisone [7]. Note that in 1961 there was still no compound, which could be positioned between haloperidol and floropipamide.

The empirical arrangement of the neuroleptics in Fig. 2 is a tripolar classification on the basis of the shapes of the activity spectra. The colors of the visible spectrum

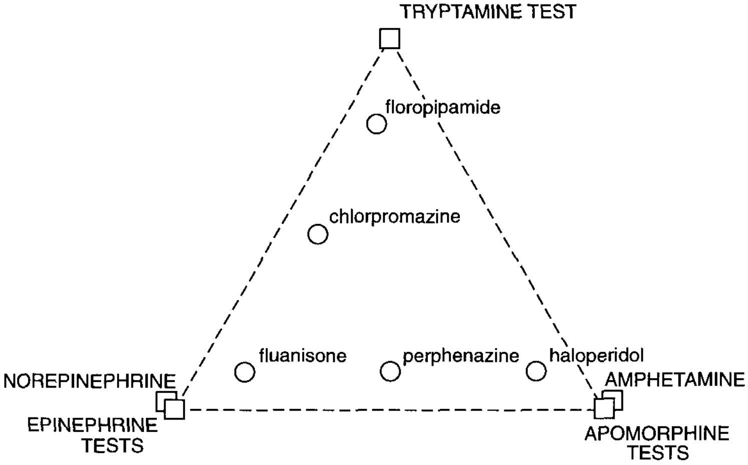


Figure 2. Tripolar empirical classification of the neuroleptics with respect to their specificities in the amphetamine and apomorphine tests (right), the tryptamine test (top) and the norepinephrine and epinephrine tests (left) (after Janssen [7]).

can also be represented in a similar tripolar (or trichromatic) diagram on the basis of their compositions of red, blue and green, as has been already shown by Thobias Mayer [8] in 1758, long before the discovery of the three types of cone cells in the retina of the eye, each with different light absorption characteristics. Such a tripolar classification can be regarded as a form of "avant-garde" spectral mapping. This suggests that geometric thinking about compounds and tests (or more generally about objects and their properties) as points in space is a fundamental thought process. Empirical classifications do not necessarily require coordinate axes for their construction, and often the underlying dimensions are discovered much later. Perhaps the most famous illustration of this is the two-dimensional periodical classification of the chemical elements by D. Mendeleev in 1869, which was based on atomic weight and chemical properties [9]. The modern interpretation of the two dimensions in terms of atomic number and valence electrons was only possible after the development of quantum mechanics. A similar epistemological process took place in the case of the classification of the neuroleptics. However, the time span between the empirical classification and the biological interpretation of its dimensions was much shorter than in the case of the trichromatic classification of colors and of the periodical classification of the elements.

In 1965 Janssen had completed a comprehensive study of 40 neuroleptics in 12 pharmacological tests in rats. Effective doses were determined at 10 consecutive hourly intervals after (subcutaneous) administration of the compounds. The results were published in a seminal paper by Janssen, Niemegeers and Schellekens [10] with the title: "Is it possible to predict the clinical effects of neuroleptics (major tranquilizers) from animal data, part I, neuroleptic activity spectra for rats". The data are reproduced in Table 1 as reciprocal effective doses (kg/mg). The paper of 1965 was declared a citation classic by Current Contents [11]. It contains the basic ideas developed by Janssen at that time for the graphical statistical analysis of

Table 1. Pharmacological activities, defined as reciprocal effective doses (ED_{50} in mg/kg) of 40 neuroleptics in 12 pharmacological tests in rats after subcutaneous administration and including antagonism of the effects of amphetamine, apomorphine, norepinephrine, epinephrine and tryptamine; inhibition of conditioned motility in the jumping box; depression of rearing and ambulation; prevention of traumatic shock; observations of catalepsy and ptosis and inhibition of weight gain [10]. Zero values that appear in this table are random zeros and are due to rounding of the data. The latter can be replaced by small positive values of the order of magnitude of 0.001.

	Catalepsy	Amphetamine	Jumping Box	Apomorphine	Weight Gain	Rearing	Ambulation	Norepinephrine	Epinephrine	Tryptamine	Ptosis	Traum. Shock	Geom. Mean
Aceperone	0.22	0.00	0.08	0.00	0.01	0.03	0.01	2.50	1.25	0.02	0.59	5.00	0.06
Acepromazine	0.13	0.48	1.43	0.07	0.40	0.26	0.29	21.74	10.00	0.17	0.83	62.50	0.92
Acetophenazine	0.67	4.55	3.23	4.55	1.02	0.71	0.67	6.25	1.59	0.04	0.71	1.43	1.19
Amiperone	0.56	9.09	6.25	8.33	1.54	1.43	1.54	0.14	0.03	0.07	0.20	0.13	0.66
Anisoperidone	0.04	1.11	0.26	0.00	0.09	0.20	0.19	2.70	0.91	0.10	0.10	5.00	0.21
Anisopirol	1.18	12.05	9.09	0.53	1.54	0.91	0.50	9.09	2.50	0.20	1.00	3.33	1.73
Benperidol	4.17	37.04	14.29	22.22	16.67	2.86	2.50	5.00	0.83	0.20	0.22	2.00	3.37
Butropipazone	0.13	0.77	0.67	0.10	0.38	0.40	0.22	7.14	3.57	0.17	0.20	16.67	0.62
Butirylperazine	0.50	2.56	1.47	4.00	1.06	0.83	0.67	1.41	0.45	0.03	0.31	0.50	0.71
Chlorpromazine	0.13	0.91	1.08	0.15	0.43	0.27	0.22	1.92	0.63	0.77	0.10	1.25	0.45
Chlorprothixene	0.56	0.67	2.50	0.04	0.24	0.83	0.50	4.17	1.82	4.55	0.33	16.67	0.97
Dixyrazine	0.14	1.20	0.50	0.24	0.22	0.18	0.14	0.06	0.04	0.14	0.07	0.20	0.17
Droperidol	2.78	27.78	33.33	14.29	5.56	1.67	0.77	10.00	5.00	0.10	0.29	12.50	3.60
Floropipamide	0.03	0.20	0.09	0.00	0.10	0.07	0.04	0.40	0.14	0.37	0.10	0.50	0.10
Fluanisone	0.56	3.70	3.33	0.45	1.08	1.43	0.50	14.29	2.70	0.17	0.83	2.00	1.31
Fluphenazine	6.25	10.00	40.00	7.69	3.70	7.14	10.00	1.08	0.43	0.40	0.83	1.00	3.10
Haloperidide	2.00	11.11	5.26	5.88	3.45	3.45	4.55	0.11	0.06	0.83	0.20	0.14	1.16
Haloperidol	5.00	26.32	17.24	5.00	3.70	7.69	4.76	0.48	0.07	0.59	0.83	0.50	2.20
Isospirilene	4.55	32.26	35.71	35.71	10.53	6.67	12.50	0.09	0.01	0.83	0.40	0.25	2.11
Levomepromazine	0.20	0.45	1.25	0.04	0.42	0.20	0.13	2.13	0.77	0.20	0.20	16.67	0.44
Methopromazine	0.15	0.17	1.11	0.02	0.22	0.16	0.19	7.69	1.59	0.12	0.25	12.50	0.40
Methylperidide	1.67	10.53	9.09	14.93	3.33	4.35	4.35	0.19	0.06	0.28	0.20	0.13	1.23
Moperone	1.33	43.48	10.53	4.00	5.56	3.85	1.28	0.91	0.40	0.22	0.22	1.00	1.81
Perphenazine	3.23	6.25	10.00	3.13	5.56	6.67	4.55	2.00	0.63	0.91	0.56	0.33	2.30
Pipamazine	0.63	0.91	1.54	0.10	0.38	1.00	0.63	3.57	1.11	1.14	0.83	8.33	0.97
Prochlorperazine	0.25	2.13	1.25	0.50	0.31	0.77	0.63	0.14	0.06	0.43	0.10	0.17	0.35
Promazine	0.03	0.03	0.10	0.00	0.07	0.04	0.02	1.43	0.59	0.05	0.03	2.00	0.08
Prothipendyl	0.03	0.03	0.40	0.00	0.07	0.03	0.03	0.50	0.33	0.05	0.02	1.43	0.07
Reserpine	1.11	0.03	5.00	0.03	0.67	1.00	0.83	0.03	0.03	0.03	2.22	1.54	0.26
Spiramide	1.67	62.50	20.00	33.33	9.09	2.00	2.86	0.11	0.02	1.00	0.40	0.33	1.77
Spiroperidol	22.22	50.00	83.33	14.29	12.50	18.18	9.09	0.83	0.23	1.59	1.43	1.59	5.92
Tetrabenazine	0.25	0.15	0.71	0.01	0.67	0.22	0.31	0.01	0.01	0.06	0.08	0.13	0.09
Thioperazine	0.77	7.69	5.88	2.13	1.43	1.67	2.70	0.40	0.20	0.05	0.13	0.50	0.84
Thiopropazate	2.78	11.36	9.09	6.67	3.45	7.69	4.55	0.91	0.23	0.91	0.59	1.00	2.33
Thioridazine	0.08	0.14	0.05	0.00	0.03	0.03	0.08	1.82	0.63	0.06	0.20	2.00	0.11
Trabuton	0.13	1.11	0.83	0.11	0.34	0.28	0.38	5.26	3.13	0.22	0.13	0.83	0.49
Trifluperazine	2.50	4.00	8.33	1.82	2.78	3.33	2.56	0.14	0.05	0.12	0.20	0.08	0.78
Triflupromazine	0.56	3.45	3.85	0.56	1.18	1.14	0.77	3.70	1.33	1.14	0.33	7.14	1.40
Trimeprazine	0.04	0.03	0.25	0.00	0.11	0.06	0.06	0.12	0.04	0.22	0.05	3.33	0.08
Triperidol	6.25	40.00	40.00	33.33	12.50	20.00	11.11	3.33	2.00	1.59	0.83	1.67	6.99
Geom. Mean	0.55	1.79	2.63	0.37	0.86	0.80	0.65	0.96	0.34	0.22	0.27	1.35	0.69

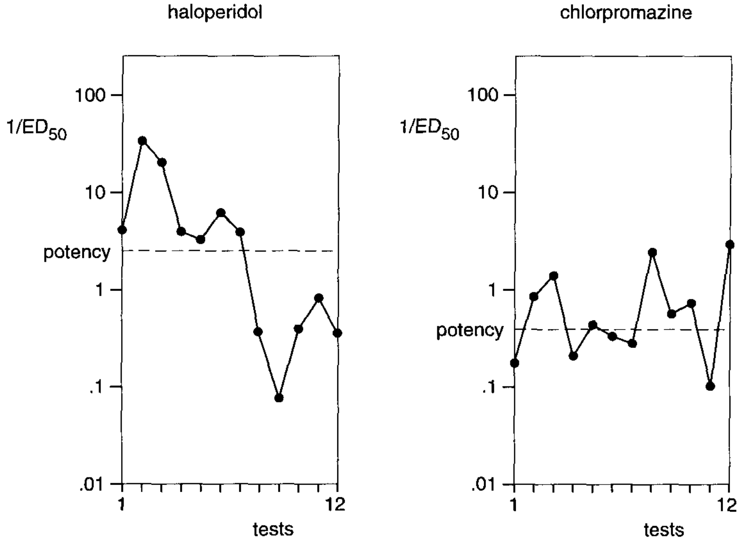


Figure 3. Activity spectra of haloperidol and chlorpromazine in 12 pharmacological tests. The vertical axis represents log reciprocal effective dose, the horizontal line shows the 12 tests in the order as they appear in Table 1.

pharmacological data. Of most importance to our discussion of spectral mapping are the "activity spectra" and the "double-log charts".

In Fig. 3 we have reproduced the activity spectra of haloperidol and chlorpromazine. The vertical scale represents the reciprocal effective doses for each of 12 pharmacological tests: antagonisms of amphetamine, apomorphine, norepinephrine, epinephrine and tryptamine; inhibiton of motility in the jumping box, rearing and ambulation tests; observations of catalepsy (rigid postures) and ptosis (closing of eyelids); reduction of weight gain and prevention of traumatic shock. The order of the tests along the horizontal axis is the same as the ordering in Table 1. This ordering was chosen deliberately and is outlined below. It was observed that the spectra could be arranged along a bipolar (or differential) axis with haloperidol-like compounds at one end and chlorpromazine-like compounds at the other. In fact, the 40 spectra were ordered along this bipolar haloperidol-chlorpromazine contrast. (This contrast can be regarded as a difference of specificities and, in the case of logarithmic transformation of the data, as a log ratio of effective doses. This will be defined more precisely later on.) At the same time, the tests were ordered in Table 1 and in Fig. 3, such that haloperidol-specific tests (such as apomorphine) were on the left side and chlorpromazine-specific ones (such as norepinephrine) appeared on the right side. The concept of symmetry between compounds and tests in spectral mapping was already apparent in this empirical bipolar classification, as both the 40 compounds and the 12 tests were ordered along the same criterion.

In the double logarithmic chart of Fig. 4, neuroleptics are defined as points, which represent their activities in the apomorphine (apo) and norepinephrine (nep) tests. The latter define the coordinate axes of the diagram. A diagonal line is constructed

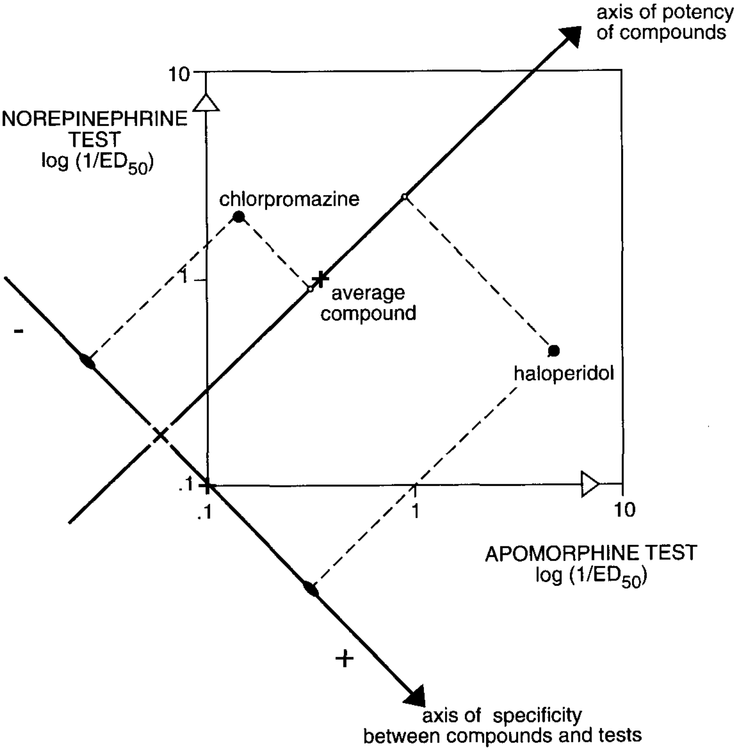


Figure 4. Double-log chart representing (schematically) compounds as points defined by their activities produced in two tests. The diagonal line through the center of the pattern (average compound) represents the axis of potency of the compounds. The axis perpendicular to the latter is the axis of specificity between compounds and tests. The projections of the compounds onto this axis of specificity define their contrasts with respect to the two tests. Haloperidol has a positive contrast, as it is more specific for antagonism of apomorphine than of norepinephrine. Chlorpromazine has a negative contrast. (Data are taken from Table 1.)

through the geometric average of the two tests (indicated by a small cross). It can also be understood that the diagonal represents an axis of potency. Compounds that are projected quite high along this axis tend to possess high activity in both the apomorphine and norepinephrine tests. Those that are projected at the lower end tend to have low activity in both tests. There may be exceptions, however. For example, aceperone is active in the norepinephrine test, but almost devoid of activity in the apomorphine test, as can be seen from the data in Table 1. The projections of the 40 compounds upon the potency axis appear in the order of their potency (or average activity) in the two tests. Compounds below the diagonal possess apo/nep ratios that are larger than average; they are more specific for apomorphine than for norepinephrine. Compounds that lie above the diagonal have apo/nep ratios that are smaller than average; they are more specific for norepinephrine than for apomorphine. It is evident from the geometrical construction in Fig. 4 that the line drawn perpendicularly to the potency axis represents an axis of specificity.

Compounds that are projected at the higher end of this axis have a positive apo/nep contrast. Those that are projected at the lower end have a negative contrast. The original coordinate axes (representing reciprocal effective doses in the apo and nep tests) can be rotated towards a new frame that is defined by the potency axis and the specificity axis. This transformation is an essential operation in spectral mapping. It can be understood that in the case of three tests, we will still obtain a single potency axis (a diagonal line which forms identical angles with all three coordinate axes) and a plane of specificities, which is drawn perpendicularly to the potency axis. The procedure can be generalized for multiple tests, resulting always in a single potency axis and a hyperplane of specificities, which is perpendicular to this axis. The number of dimensions of this hyperplane is one less than the number of tests. Spectral mapping can, thus, be understood to be a decomposition of the activity data into a potency component and one or more components of specificity.

A similar geometrical decomposition can be applied in a double-log diagram, in which the 12 tests are represented as points based on the activities of haloperidol (hal) and chlorpromazine (cpz). The latter form the coordinate axes of the diagram (Fig. 5). The diagonal line then represents the axis of sensitivity, as the projections of the 12 tests upon this axis are represented in the order of their sensitivity (or average activity) for the two compounds. The projections upon the line drawn perpendicularly to the sensitivity axis also represent specificities. Tests that are projected high upon the axis of sensitivity have greater than average ratios of hal/cpz and have positive contrasts. Those at the lower end of the specificity axis have less than average ratios and have negative contrasts. Because of the symmetry requirement of spectral mapping, one cannot dissociate the two diagrams (Figs. 4 and 5). For this reason they are referred to as dual representations. Both should be understood in multiple dimensions rather than in the two-dimensional simplifications, which we have used here. The general multivariate case, however, can only be treated properly in algebraic terms as will be shown below.

The apomorphine/norepinephrine contrast is one of the strongest that can be found in the pharmacological data on the neuroleptics (Table 1). The degree of contrast can be measured by the spread of the projections upon the axis of specificity. If the spread is zero, then all compounds possess the same apo/nep ratio, and, hence, the contrast is zero. The larger the spread, the more variety there is among the apo/nep ratios, and hence, the compounds will show widely varying specificities for either of the two tests. Here, the apo/nep contrast produces a bipolar classification of the compounds, from being highly specific in the apo test towards being highly specific in the nep test. The former compounds are also referred to as being haloperidol-like and the latter chlorpromazine-like. The relevance of this finding was that the bipolar pharmacological classification obtained in rats agreed to a large extent with the bipolar clinical classification of the neuroleptics. The latter was based upon a contrast between the antipsychotic and the tranquilizing properties of the neuroleptics, which were described by Lambert and Revol [12]. In clinical tests was observed that the haloperidol-type compounds were specific for the treatment of delusions and manic states, while the chlorpromazine-type compounds were specific for cases of extreme agitation and confusion. It, thus, appeared to be possible, as claimed by Janssen, et al. [10], to predict the clinical effects of the neuroleptics from their pharmacological

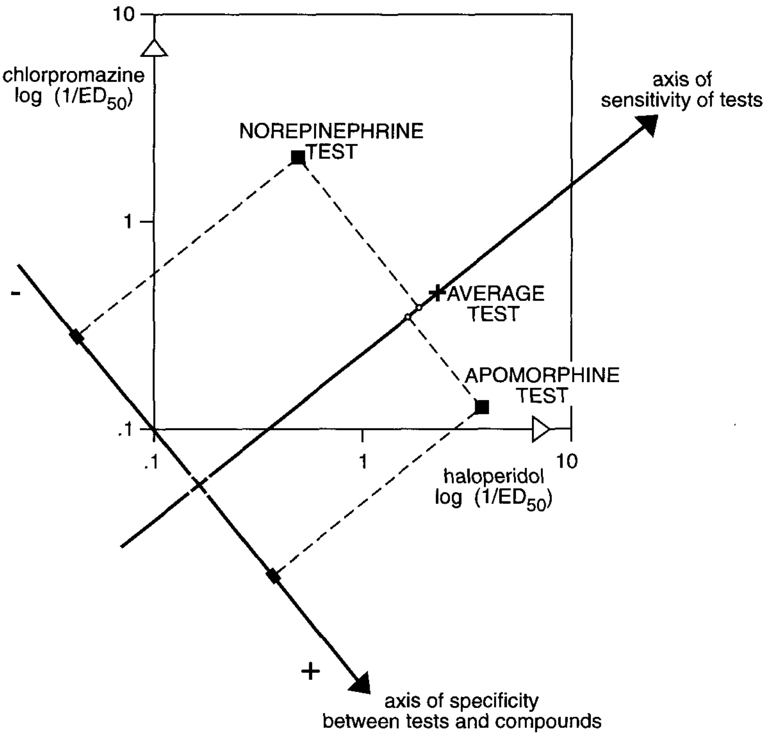


Figure 5. Double-log chart (schematically) representing tests as points defined by their activities exhibited by two compounds. The diagonal line through the center of the pattern (average test) represents the axis of specificity of the tests. The axis drawn perpendicularly to it is the axis of specificity between tests and compounds. The projections of the tests onto this axis of specificity define their contrasts, with respect to the two compounds. The apomorphine test has a positive contrast, as it is more specific for haloperidol-type compounds than for chlorpromazine-type compounds. The norepinephrine test has a negative contrast. (Data are taken from Table 1.)

activity spectra, since the antipsychotic/tranquilizing contrast observed in the clinic correlated well with the apomorphine/norepinephrine contrast in rats. So far, no clinical equivalent could be detected for the tryptamine/norepinephrine or tryptamine/dopamine contrasts that were also apparent on the pharmacological activity spectra (Fig. 2).

The agreement between pharmacological measurements and clinical observations led to speculation about the causes of psychosis and about the mechanism of action of neuroleptics in the brain. It has been known since the early sixties that apomorphine acts as a dopamine agonist, i.e. it mimics the effects of dopamine. Similarly, tryptamine was known as a serotonin agonist. These considerations led to hypotheses that postulated the existence of dopamine and serotonin receptors in the brain, which are implicated in psychosis and which can be blocked by neuroleptics [13, 14, 15]. The concept of receptors, intermediates between drugs and the organisms in which they produce an effect, has been postulated by Emil Fisher [16] as early as 1894. The existence of receptors in the brain that could be blocked by neuroleptics was in 1963 corroborated by the discovery of highly specific apomorphine-blocking neuroleptics in the class of the diphenylbutylamines by Janssen et al. [17]. A great advance was made in the mid-seventies, when it became possible by means of radioactively-labeled compounds, to identify receptors in the brain that bind specifically to dopamine [18, 19], serotonin [20] or norepinephrine [21], and which can be blocked by neuroleptics. These radioactive ligand studies not only provided a biochemical basis for the study of psychosis, they also created the new field of biochemical pharmacology. The neuroleptic receptors are known to be G-coupled proteins that are located in the postsynaptic membranes in the striatum and frontal cortex. In addition to the spectra from animal studies and from clinical observations, biochemical spectra of receptor binding of the neuroleptics became available. It was soon realized that neuroleptics vary considerably in their affinity for the dopamine (D), serotonin (5HT) and norepinephrine (α) receptors. Nowadays, these receptors are known to occur as various types (D1, D2 up to D5, 5HT1, 5HT2 up to 5HT7 and $\alpha1$, $\alpha2$), which in turn, can be divided into several subtypes. Furthermore, it was established that occupation of these receptors *in vitro* correlated well with the *in vivo* antagonism of the effects produced by apomorphine, tryptamine and norepinephrine in rats. Concurrently with these developments, the Janssen screening tests for neuroleptics was reduced to a single test (ATN), in which the agonists apomorphine, tryptamine and norepinephrine were given sequentially following administration of a neuroleptic [22]. These developments provided a rational interpretation of the three poles or vertices in the initial classification of the neuroleptic spectra (Fig. 2) in terms of experimentally confirmed receptors.

An analogy exists between the tripolar classification of the neuroleptics based on their affinities for the apomorphine, tryptamine and norepinephrine receptors, and the afore mentioned trichromatic diagram, which is defined by the three primary colors red, blue and green. Both represent contrasts between the effects produced by receptors (G-protein coupled receptors in the brain and photoreceptors in the cone cells of the retina, respectively). In both cases, the contrasts have been detected before the underlying biochemical mechanisms were discovered, by careful exploration and interpretation of the empirical data. In the case of the neuroleptics as well as in that of the colors, latent (or hidden) variables are responsible for the multivariate contrasts that are experimentally observed. Latent variables are computed from the manifest (or observed) variables by methods which are related to factor analysis. These are usually few in number and their combination enables the reconstruction of the original manifest variables [23]. The tripolar classification of the neuroleptics and the trichromatic diagram were the result of an empirical factor analysis, which did not require multivariate analysis and computers. Not all research programs, however, can be concluded by empirical factor analysis, in which the hidden variables are detected by careful interpretation and graphical analysis, as will be shown below.

In 1974, our laboratory completed a comprehensive study involving 35 reference compounds with antiepileptic properties, based on the following five observations in rats: induction of ataxia and loss of righting, suppression of metrazol-induced tonic extension of the hind and forelegs and clonic seizures [24]. The corresponding activity spectra (such as in Fig. 3) were drawn on cards and spread out on a table. This time, however, it proved to be difficult to classify the spectra along a bipolar line or a tripolar plane, as has been done in the case of the neuroleptics (Fig. 2). Pharmacologists differed in opinion about the similarities and dissimilarities of the spectra and arranged them accordingly in different ways. Indeed, there was no a priori indication as to the nature and the number of underlying factors in this case. Eventually, it was realized that only multivariate analysis could produce an objective mapping of the antiepileptics, with which everybody could agree and which could be used as a starting point for further interpretation.

Rather than attacking the antiepileptic problem immediately, it was decided that we would tackle the case of the neuroleptics first. As the solution had already been derived empirically, it would be easier to determine whether the results of multivariate analysis made sense or not. At that time, in 1974, two basic methods for multivariate analysis of tabulated activity data (such as in Table 1) were available as a first choice. First, one could use Principal Components Analysis (PCA), a well established proven method [25, 26]. PCA produces a graphic display of the compounds (called scores plot) which shows their similarity on the basis of the observed activities in the various tests. This, however, was rather irrelevant to our problem, as we wanted to classify compounds with respect to their specificities in the various tests, i.e. independently of their potencies. Furthermore, PCA also yields a graphic display of the tests (called a loadings plot) which shows the structure of their intercorrelations. This could give us some indication as to the number of structural factors (or hidden variables), which could explain most of the variation in the data. Clearly, PCA does not show specificities between compounds and tests. It is also not symmetric with respect to compounds and to tests. The other method, Correspondence Factor Analysis (CFA), had only recently been described by Benzécri in 1973 [3] and was at that time mainly known by French data analysts and statisticians. Fortunately, we were introduced to CFA by Lacourly [27], a member of Benzécri's group. The revolutionary aspect of CFA was that it showed the specificities between the rows and columns of a table, when in the form of a cross-tabulation (representing parts of a whole) or contingency table (representing counts). The method makes use of the biplot graph developed by Gabriel, [2] in which both the rows and columns are displayed in one and the same graph. The name biplot is derived from the superposition of the scores plot and the loadings plot.

The reading rules of the CFA biplot are simple. A row and a column that are positioned in the same direction, as viewed from the center are considered to have positive specificity. A row and a column viewed from the center in opposite directions have negative specificity. Rows and columns that are at a distance from the center have high specificity, and those that are close to the center are nonspecific. In the case of the pharmacological data of the neuroleptics (Table 1), the CFA biplot produces a classification of the compounds in the form of Fig. 2 and reveals the tripolar structure of the tests. Hence, it supports the three-receptor theory. The problem, however, in a pharmacological application, is that the contrasts that are seen on the CFA biplot cannot be readily interpreted in terms of the ratios between observed effective doses. This follows from the fact that dissimilarities between compounds and between tests are to be interpreted in CFA as distances of *chi* square. The latter is a measure of the relative deviation between the observed and expected values. This follows from a particular transformation in CFA, in which pharmacological activities are transformed into a kind of specificities. We denote the table of observed activities (i.e. reciprocal effective doses) by $\boldsymbol{X}$, and the corresponding table of transformed values by $\boldsymbol{Z}$. The transformation by CFA is then defined by:

$$z_{ij} = \left(x_{ij} - \frac{x_{i.}x_{.j}}{x_{..}}\right) \Bigg/ \frac{x_{i.}x_{.j}}{x_{..}} = \frac{x_{ij}x_{..}}{x_{i.}x_{.j}} - 1 \qquad (1)$$

where x_{ij} and z_{ij} represent the elements corresponding to row i and column j in the tables $\boldsymbol{X}$ and $\boldsymbol{Z}$, and where $x_{i.}$ is the (arithmetic) mean of row i, $x_{.j}$ is the (arithmetic) mean of column j, and $x_{..}$ is the global (arithmetic) mean over all rows and columns of $\boldsymbol{X}$.

The transformation by CFA in Eq. 1 is symmetric with respect to compounds and tests, i.e. they are interchangeable. Note that the quantity $x_{i.}x_{.j}/x_{..}$ is called the expected value of x_{ij}, under the assumption that the observed activities $\boldsymbol{X}$ are multinomially distributed. (The assumption is, in the strictest sense, not fulfilled here, since the observed activities do not represent counts, nor can they be considered as parts of a whole.) In a broader sense, one may consider the transformed values z_{ij} as representing a kind of specificity, since the observed activities x_{ij} are divided by a type of potency, $x_{i.}$ and by a type of sensitivity, $x_{.j}$. The problem that arose, however, was that in the context of CFA, potency and sensitivity are defined as arithmetic means rather than as geometric means of the observed activities, which is incompatible with the idea that biological response is related to the logarithm of the observed effective dose, rather than to the dose itself. This idea stems from early psycho-physical observations and is embodied in the Weber-Fechner law [28], which postulates an approximately linear relationship between response and the logarithm of the stimulus. This law is reflected in the decibel scale for sound energy, the magnitude scale for luminosity, etc. It has also been observed that effective doses are usually more normally distributed on a logarithmic scale than on a linear scale [29].

4.5.3 Spectral Map Analysis

Spectral Map Analysis (SMA) is a straightforward multidimensional extension of the two-dimensional geometrical constructions in Figs. 4 and 5. The original idea was to define a transformation which would transform the observed log $\boldsymbol{X}$ data into specificities, $\boldsymbol{Z}$, and which would have the geometrical effects such as is illustrated in Figs. 6 and 7. In the first of these, we consider each of the n compounds to be represented as points in a p-dimensional coordinate space, S_p, by means of their log activities produced in p tests (Fig. 6). The diagonal line, which forms an equal angle with all p coordinate axes, represents the axis of potency of the compounds, and the plane drawn perpendicularly to this axis, represents the plane

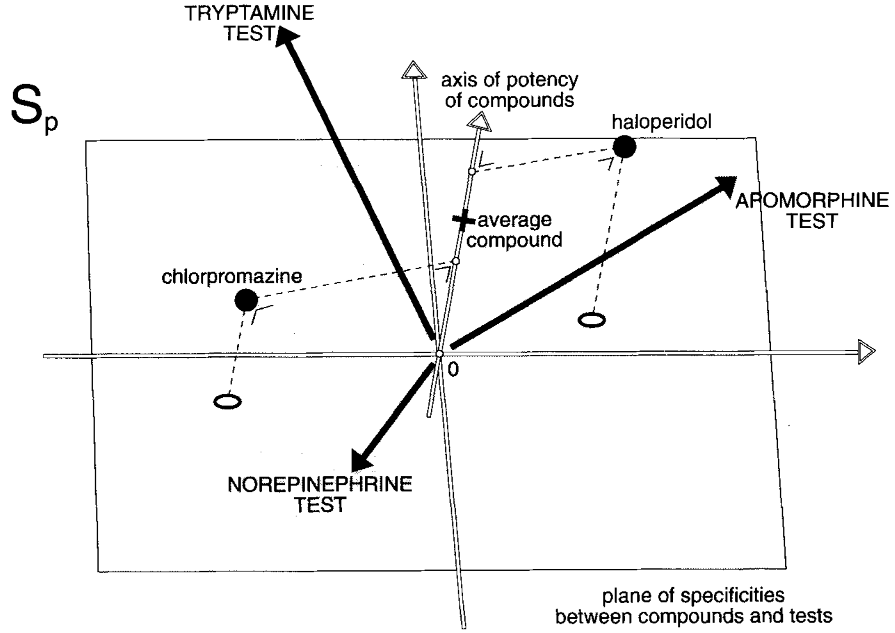


Figure 6. Extension of the double-log chart of Fig. 4 for multiple dimensions. The n compounds are represented as points in p-dimensional space, S_p, spanned by p tests. The diagonal line represents the potencies of the compounds, and the (hyper)plane perpendicular to it defines the specificities between compounds and tests. Spectral mapping performs a Principal Components Analysis (PCA) of the specificities in the (hyper)plane.

of contrasts. For the convenience of the illustration, we have only represented three tests, but in the general case, the plane of contrasts will be a multidimensional (hyper)plane. In the dual representation (Fig. 7) we represent each of the p tests as points in an n-dimensional coordinate space, S_n, by means of their log activities obtained from n compounds. The diagonal line, which forms equal angles with all n coordinate axes, represents the axis of sensitivity of the tests. The plane drawn perpendicularly to this axis represents the plane of specificities. For convenience, we have only considered three compounds, but in the general case, the plane of specificities will be a multidimensional (hyper)plane. It should be understood that the geometrical procedure can be extended to any number of compounds and tests. The two spaces S_n and S_p are referred to as being dual, since one cannot exist without the other. A change in one will automatically result in a change in the other. The geometrical projection can also be defined in the form of an algebraic transformation of the data [30], which in its simplest form can be defined by:

$$z_{ij} = \log x_{ij} - \log \tilde{x}_{i.} - \log \tilde{x}_{.j} + \log \tilde{x}_{..} = \log \frac{x_{ij}\tilde{x}_{..}}{\tilde{x}_{i.}\tilde{x}_{.j}} \qquad (2)$$

where x_{ij} and z_{ij} represent the elements corresponding with row i and column j of the tables X and Z, where $\tilde{x}_{i.}$ is the (geometric) mean of row i, $\tilde{x}_{.j}$ is the (geometric) mean of column j and $\tilde{x}_{..}$ the global (geometric) mean over all rows and columns of X.

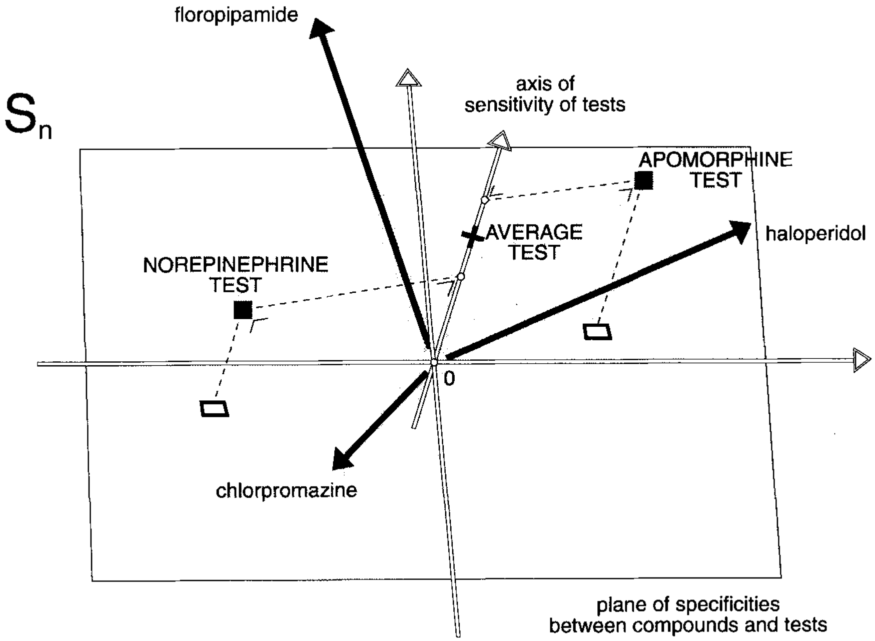


Figure 7. Extension of the double-log chart of Fig. 5 for multiple dimensions. The p tests are represented as points in n-dimensional space, S_n, spanned by n compounds. The diagonal line represents the sensitivities of the tests, and the (hyper)plane perpendicular to it defines the specificities between compounds and tests. Spectral mapping performs a Principal Components Analysis (PCA) of the specificities in the (hyper)plane. Figs. 6 and 7 are dual geometrical representations of one and the same table of data.

The transformation of SMA by Eq. (2) is symmetrical, with respect to the compounds and the tests, i.e. they can be interchanged. Technically, the transformation is referred to as a log double centering of the observed activities, i.e. a simultaneous correction of the data for differences between the corresponding row and column means. The term $\tilde{x}_{i.}$ can be regarded as an estimate of the potency of compound i. Likewise, the term $\tilde{x}_{.j}$ can be considered as an estimate of the sensitivity of test j. The term $\tilde{x}_{..}$ is a constant for the table. Consequently, the expression applies a correction to the log activities for differences in potency of the compounds and for differences in sensitivity of the tests. The result can, thus, be interpreted as specificities between compounds and tests, according to our earlier definition.

We now define contrast as the difference between two specificities, and by virtue of the symmetry principle, we obtain two types of contrasts. The contrast between compounds i and i' with respect to test j follows from Eq. (2):

$$z_{ij} - z_{i'j} = \log x_{ij} - \log \tilde{x}_{i.} - (\log x_{i'j} - \log \tilde{x}_{i'.}) = \log \frac{x_{ij}}{x_{i'j}} - \log \frac{\tilde{x}_{i.}}{\tilde{x}_{i'.}} \qquad (3)$$

and all the other terms cancel out. This shows that contrasts can be defined in terms of log ratios of effective doses.

Similarly, the contrast between tests j and j' with respect to compound i also follows on from Eq. (2):

$$z_{ij} - z_{ij'} = \log x_{ij} - \log \tilde{x}_{.j} - (\log x_{ij'} - \log \tilde{x}_{.j'}) = \log \frac{x_{ij}}{x_{ij'}} - \log \frac{\tilde{x}_{.j}}{\tilde{x}_{.j'}} \tag{4}$$

and all the other terms cancel out. Again, this shows that contrasts can be defined in terms of log ratios of effective doses.

We can define the root mean square (*rms*) contrast between compounds i and i' as the norm of the individual contrasts of compounds i and i', with respect to all tests:

$$\|z_i - z_{i'}\| = \left(\frac{1}{p}\sum_{j}^{p}(z_{ij} - z_{i'j})^2\right)^{1/2} \tag{5}$$

where z_i and $z_{i'}$ denote vectors of p elements with the general element z_{ij} and $z_{i'j}$. Similarly, we define the *rms* contrast between test j and j' as the norm of the individual contrast of tests j and j' with respect to all n compounds:

$$\|z_j - z_{j'}\| = \left(\frac{1}{n}\sum_{i}^{n}(z_{ij} - z_{ij'})^2\right)^{1/2} \tag{6}$$

where z_j and $z_{j'}$ denote vectors of n elements with general element z_{ij} and $z_{ij'}$. For the sake of completeness, we define the *rms* specificity of compound i as the norm of the individual specificities of compound i for all p tests j:

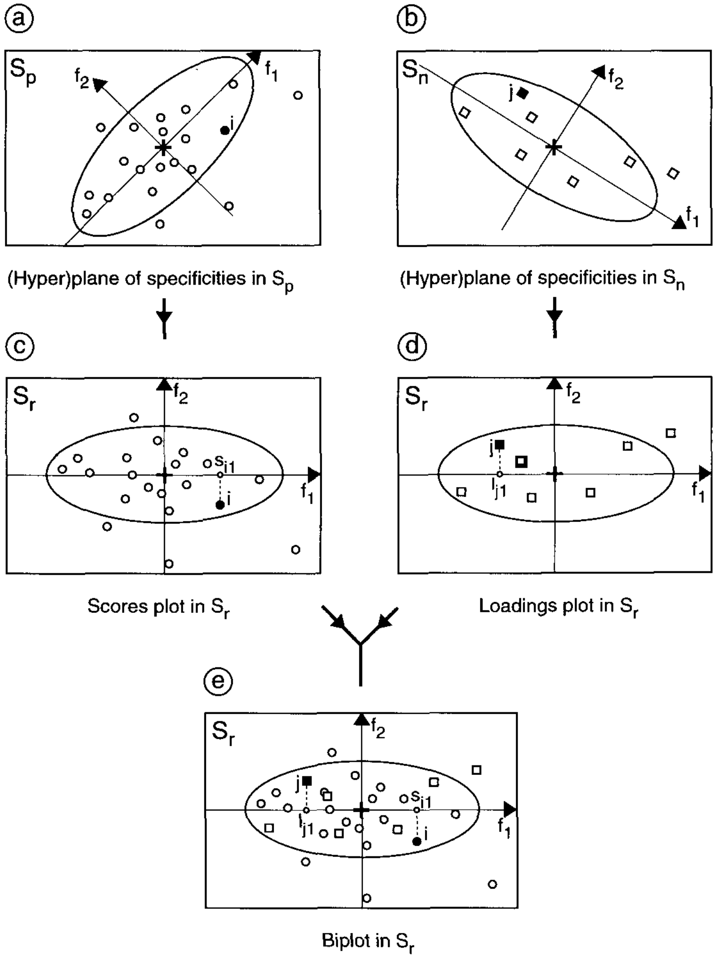
$$\|z_i\| = \left(\frac{1}{p}\sum_{j}^{p} z_{ij}^2\right)^{1/2} \tag{7}$$

In a similar fashion, we define the *rms* specificity of a test j as the norm of the individual specificities of test j for all n compounds i:

$$\|z_j\| = \left(\frac{1}{n}\sum_{i}^{n} z_{ij}^2\right)^{1/2} \tag{8}$$

The reason why root mean square (*rms*) values are taken, rather than mean values, is that the mean specificities over all compounds, or over all tests, are zero, as follows from the definition of specificities in Eq. (2). The same applies to contrasts, since a contrast has been defined as the difference between two specificities. Since the (hyper)planes containing the projections of the compounds and tests are usually multidimensional, one may attempt to reduce the apparent high dimensionality by means of Principal Component Analysis (PCA). Hence, one can describe SMA as a log double centered PCA, or as PCA of specificities. In the present context of SMA, we will refer to a principal component of specificity by the more general term of factor. This avoids confusion with the results of the more usual PCA, which is applied to column standardized data.

Fig. 8a is a schematic representation of the projection of the n compounds (rows) in the (hyper)plane of specificities within the coordinate space, S_p, as defined by the p tests (columns). Fig. 8b shows the projections of p tests (columns) in the (hyper)plane of specificities within the dual coordinate space, S_n, as defined by n



(Hyper)plane of specificities in S_p

(Hyper)plane of specificities in S_n

Scores plot in S_r

Loadings plot in S_r

Biplot in S_r

Figure 8. Illustration of the dual geometrical representation of the data in an $n \times p$ table of transformed data (specificities). Panel a shows n compounds in test space, S_p. Panel b shows p tests in compound space, S_n. The axes (f_1 and f_2) represent the principal components or factors of the patterns. Corresponding factors are common in the two spaces. The elliptic contours are a schematic representation of the probability density contours of the patterns of points. The diagram corresponds to the planes of specificity in Figs. 6 and 7. Panels c and d show the compounds and tests in the rotated space, S_r, spanned by the common factors. The projection of compound i onto factor 1 defines the score s_{i1}. The projection of test j onto the same factor 1 defines the loading s_{j1}. In panel e, the scores and loadings plots are combined into a single biplot.

compounds (rows). By convention, we represent compounds (rows) by circles, and tests (columns) by squares in our graphical representations. The factors are the axes of inertia of the patterns of the points in the dual (hyper)planes. In the case of an ellipsoidal structure, these factors can be regarded as axes of symmetry of the patterns. It is important, although not easy, to realize that corresponding factors in the dual representations are common. For example, factor 1 of the pattern of compounds is the same as factor 1 in the pattern of tests, etc. (We will show in the mathematical section how to extract factors from tabulated data.) The importance of a factor is measured by its variance (dispersion). Factors are ordered in decreasing order of their contribution to the global variance of the transformed data, i.e. of the specificities between compounds and tests in $\boldsymbol{Z}$. The global variance, c, which in this context is also referred to as global interaction, is defined by the expression:

$$c = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} z_{ij}^2 \tag{9}$$

The number of factors that can be extracted from the transformed data is equal to the minimal number of dimensions that are required to represent the compounds and the tests in the (hyper)planes of specificity. This minimal number of dimensions is called the rank, r, of the transformed data matrix $\boldsymbol{Z}$. It is, at the most, equal to the smallest of the number of rows or number of columns minus one. (The loss of one dimension is due to the removal of the potency and sensitivity from the observed activities). The factors are, by construction, orthogonal which implies that they are uncorrelated. As a consequence, the sum of the variances contributed by all r factors is equal to the global variance, c, in the transformed data. If λ_k^2 denotes the contribution of factor k to the global variance, then we obtain:

$$c = \sum_{k=1}^{r} \lambda_k^2 \tag{10}$$

Once the orientations of the factors are known, it is possible to rotate the (hyper)planes of specificity towards the orthogonal factors, as is shown in Fig. 8c and d. The net result is that the original spaces, S_p and S_n, are rotated into a common factor space, S_r. The coordinates of the n rows (compounds) along r common factors in S_r are conventionally called factor scores and are compiled in the $n \times r$ matrix of factor scores, $\boldsymbol{S}$. A plot, such as in Fig. 8c, of the n rows (compounds) in a low-dimensional factor space is referred to as a scores plot. In a similar manner, one refers to the coordinates of the p column (tests) along the r common factors by the conventional name of factor loadings, which are compiled in the $p \times r$ matrix of factor loadings, $\boldsymbol{L}$. A plot, such as in Fig. 8d, of the p columns (tests) in a low-dimensional factor space is referred to as a loadings plot. Since the factors are common to both the scores and loadings plots, it is feasible to combine the two into a so-called biplot as shown in Fig. 8e. The biplot derives its name from the two entities (rows and columns) that are represented in one and the same plot, spanned by common factors [2]. The scaling of the factor coordinates in $\boldsymbol{S}$ and $\boldsymbol{L}$ is such that they are related to the specificities in $\boldsymbol{Z}$ by the matrix product:

$$z_{ij} = \sum_{k=1}^{r} s_{ik} l_{jk} \tag{11}$$

where s_{ik} is the coordinate (score) of compound i along factor k, and where l_{jk} is the coordinate (loading) of test j along factor k. The above formula also defines the singular value decomposition (SVD) of table $\boldsymbol{Z}$ [31]. From the relation in Eq. (11), follow two expressions for contrasts in terms of the factor coordinates:

$$z_{ij} - z_{i'j} = \sum_{k=1}^{r} (s_{ik} - s_{i'k})\, l_{jk} \tag{12}$$

$$z_{ij} - z_{ij'} = \sum_{k=1}^{r} s_{ik}(l_{jk} - l_{j'k}) \tag{13}$$

We will now briefly provide a geometrical interpretation of the double centered PCA biplot (or biplot of specificities). A salient feature of the double centered biplot is that both the pattern of rows (compounds) and the pattern of columns (tests) are centered about the origin of factor space. In the usual column-standardized PCA biplot, only the pattern of the rows is centered about the origin. (In the illustrations of Fig. 8 the origin of factor space is indicated by a small cross.)

The distance of a compound or test from the center of the plot is proportional to their *rms* specificity, as defined by Eqs. (7) and (8). A compound that is at a distance from the center has specificities (positive and negative) for two or more tests. A test that is at a distance from the center also has specificities (positive and negative) for two or more compounds. The center itself represents the aspecific compound and test. The distance between two compounds, or between two tests, is proportional to the *rms* contrasts between them, according to Eqs. (5) and (6). Two compounds that are at a distance from one another have contrasts (positive and negative) for two or more tests. Two tests that are at a distance from one another exhibit contrasts (positive and negative), with respect to two or more compounds. Compounds that are coincident on the plot have zero contrast; their spectra of activity are similarly shaped (although they may be different in potency). Likewise, tests that are coincident on the plot have zero contrast; their spectra of activity are similarly shaped (although they may differ in sensitivity).

The projection of a point, representing a test j, onto an axis through the center and through a compound i is proportional to the specificity z_{ij} between the compound and test. The same is true for the projection of a point representing a compound i onto an axis through the center and a test j (Figs. 9a and 9b). This follows from Eq. (11). An axis through the center and through a point, representing a compound or test, is called a unipolar axis. It reproduces the specificities in the table of transformed data, $\boldsymbol{Z}$.

The projection of a point, representing a test j, onto an axis through two compounds i and i', reproduces the contrast between them. The same property holds for the projection of a point, representing a compound i, and an axis through two tests j and j' (Figs. 9c and 9d). This follows from Eqs. (12) and (13). An axis through two compounds or through two tests is called a bipolar axis. It reproduces contrasts in the form of differences between specificities in the table of transformed data, $\boldsymbol{Z}$. It should be remembered that a contrast is defined here as a difference of specificities, and that because of the logarithmic reexpression, these can be interpreted as log ratios in accordance with Eqs. (3) and (4).

A fundamental problem with the biplot is its impossibility to exactly reproduce the distances between rows and between columns, and at the same time, allow projections between the rows and columns. In other words, it is not possible to

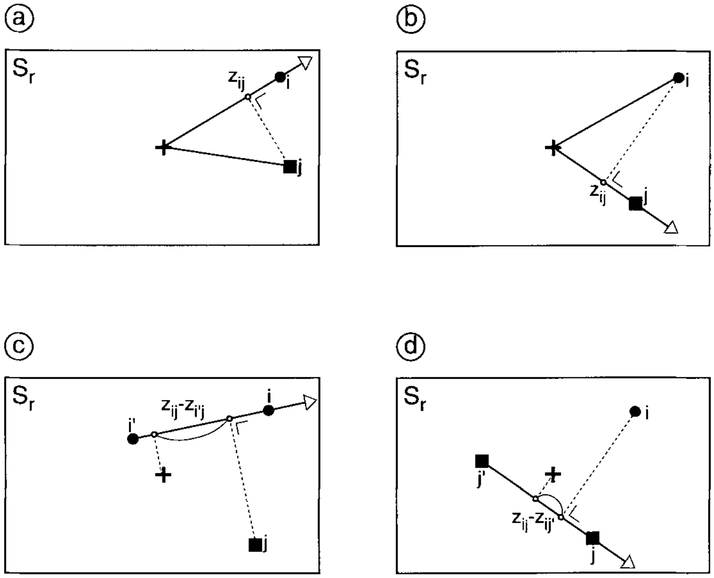


Figure 9. Reading rules of the spectral map.
a) projection of a test j onto a unipolar axis through a compound i and the center (+) reproduces the specificity z_{ij},
b) projection of a compound j onto a unipolar axis through a test j and the center (+) reproduces the specificity z_{ij},
c) projection of test j onto a bipolar axis through two compounds i and i' reproduces the contrast $z_{ij} - z_{i'j}$
d) projection of a compound i onto a bipolar axis through two tests j and j' reproduces the contrast $z_{ij} - z_{ij'}$.

exactly reproduce simultaneously the *rms* specificities (or *rms* contrasts) and the specificities (or contrasts) themselves. What is exactly reproduced depends on the choice of two so-called factor scaling coefficients, α and β, one for the coordinates of the rows and the other for the coordinates of the columns in factor space. In SMA we opted for the exact reproduction of specificities at the expense of the reproduction of the *rms* specificities. Our choice of factor scaling coefficients for row and column coordinates is symmetric and is defined by $\alpha = \beta = .5$, as explained in the mathematical section. In a two-dimensional biplot, the degree of distortion is proportional to the quartic root of the ratio of the contributions $(\lambda_1^2/\lambda_2^2)$ to the global variance of the two factors that span the biplot. The effect of this distortion is to increase the spread along the second factor of the spectral map. The distortion is minimal when λ_1 is close to λ_2, but increases when λ_2 is much smaller than λ_1. (It is assumed here that factors are arranged in decreasing order of the magnitude of their contribution to the global variance, hence $\lambda_1 \geqq \lambda_2$.)

Often one finds that only the first few factors account for the structure in the data. The remaining factors then represent noise, artifacts or information that is not relevant to the problem at hand. If there are r^* structural factors and $r - r^*$ residual ones, then the accuracy γ of the representation in the reduced factor space, S_{r^*}, is given by:

$$\gamma = \sum_{k=1}^{r^*} \lambda_k^2/c \tag{14}$$

where λ_k^2 has been defined before as the contribution of factor k to the global variance, c, of the transformed data, $\boldsymbol{Z}$.

4.5.4 Spectral Map of the Neuroleptics

Fig. 10 shows the spectral map derived from Table 1. The horizontal axis represents the first structural factor which accounts for 73% of the global variance of the transformed data (specificities), the vertical axis represents another 12% of the global variance. The third factor, which is perpendicular to the plane of the map, contributes only 6% and is, therefore, not regarded as being relevant. Hence, the two-factor SMA biplot represents 85% of the global variance, with the residual 15% being in higher dimensions of factor space. In this case, we have, at the cost of 15% residual variance, the advantage of representing the specificities between compounds and tests in a planar representation with sufficient accuracy. Not all cases are as straightforward as this one, however, and some may require the inclusion of the third factor for a 3-dimensional perspective. In more complicated cases, it may be necessary to split the table and perform two or more separate analyses, such that each produces a two- or three-dimensional spectral map which accounts for a sufficient amount of variance in the subdivided table.

The reading rules of this SMA biplot are as follows. First, circles represent compounds and squares represent tests, according to an earlier convention. Secondly, the areas of the circles and squares are related to the potency and sensitivity of the compounds and tests, respectively, as defined by the geometric means of the rows and columns of Table 1. The third rule defines the positions of the compounds and tests on the biplot. Compounds that are close together on the map possess similar activity spectra, irrespective of their differences in potency (e.g. haloperidol and spiroperidol; chlorpromazine and triflupromazine). These compounds exhibit little contrast in the various tests. Tests that are close together on the map have similar activity spectra, irrespective of their differences in sensitivity (e.g. amphetamine and apomorphine; norepinephrine and epinephrine). These tests produce no contrasts in the various compounds. Compounds and tests that appear in the same direction from the center (which is indicated by a small cross) have positive specificity for one another (e.g. benperidol and apomorphine; fluanisone and norepinephrine; floropipamide and tryptamine). These compounds and tests are said to attract each other. Compounds and tests that appear on opposite sides of the center have negative specificity. These compounds and tests are said to repel each other.

The horizontal factor, which accounts for 73% of the specificities, is determined by the contrast between amphetamine and apomorphine, on the one hand, and epinephrine, norepinephrine and traumatic shock, on the other hand. The vertical

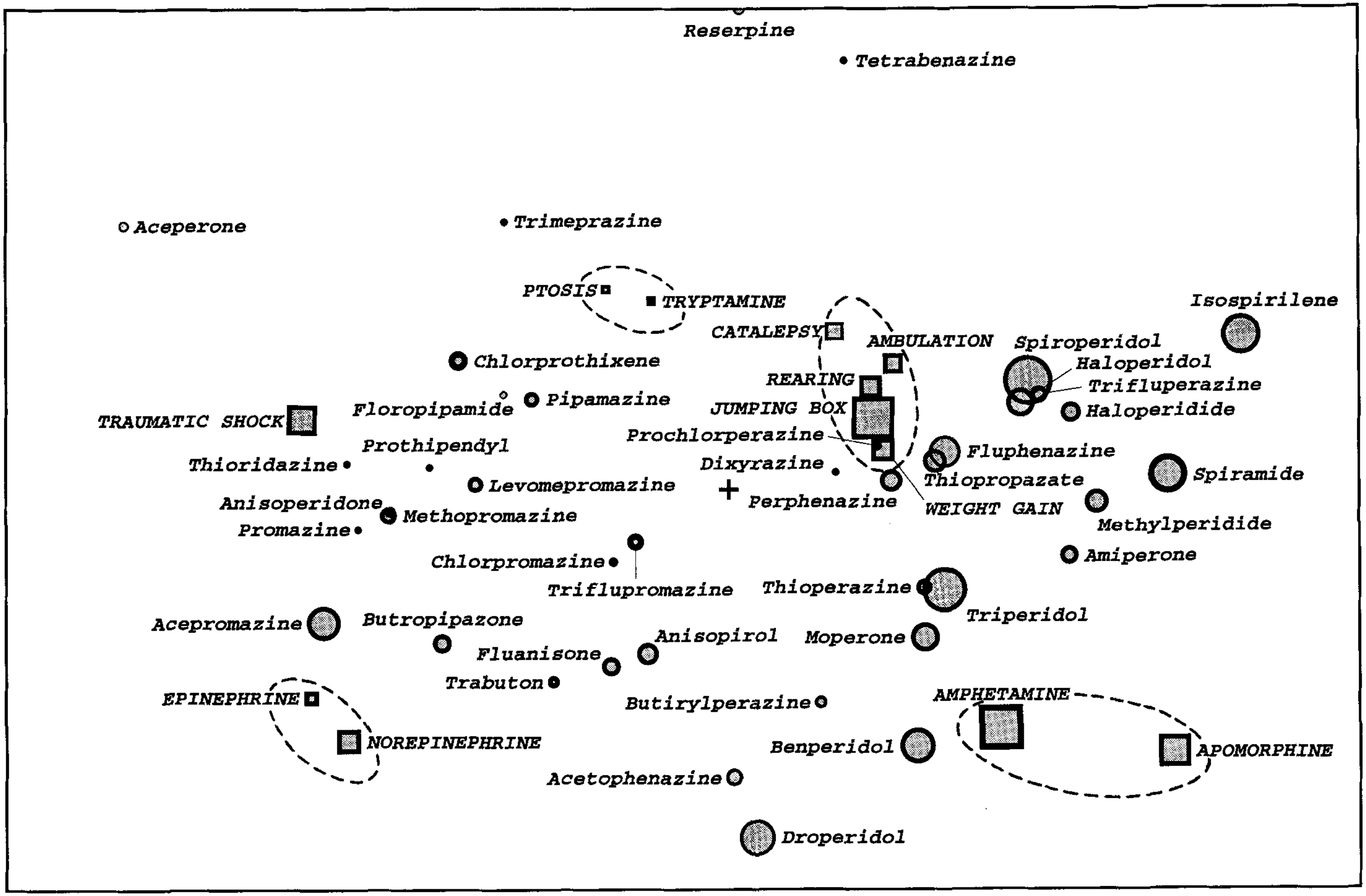


Figure 10. Spectral map of pharmacological activities of 40 neuroleptics in 12 pharmacological tests, as defined by Table 1. Circles represent compounds and squares represent tests. Areas of circles are proportional to the geometric mean activity (potency) of the compounds, as shown in the marginal column of Table 1. Areas of squares are proportional to the geometric mean activity (sensitivity) of the tests, as shown in the marginal row of Table 1. Distances of circles and squares from the center (+) of the biplot are related to their root mean square specificity. The interpretation of this spectral map is given in the text. The three poles of the map are formed by the amphetamine and apomorphine tests, the epinephrine and norepinephrine tests, and the tryptamine and ptosis tests, respectively.

factor, which only contributes 12%, is determined by the contrast of the former ones with tryptamine and ptosis. This interpretation of the spectral map agrees, to a large extent, with the result of empirical factor analysis undertaken in 1961 [7] and is reproduced schematically in Fig. 2. As predicted, chlorpromazine lies between fluanisone and floropipamide, and perphenazine lies between haloperidol and fluanisone. The spread along the vertical direction is exaggerated, however, due to the large discrepancy between the contributions of the first and second factors. The apparent spread is inflated by an amount which is equal to the quartic root of 73/12, or about 157% of the real spread. As already stated, this is caused by the particular choice of scaling coefficients of the factor coordinates ($\alpha = \beta = .5$) which favors projections (specificities) rather than reproductions of distances (*rms* specificities). A prominent feature on the map is the similarity of the motility tests, including ambulation, rearing, jumping box and catalepsy. Together with the other similarities (amphetamine and apomorphine; norepinephrine and epinephrine; ptosis and tryptamine), this redundancy in the data suggests that a large part of the information on the map could have been produced with fewer tests. In the case of a two-factor biplot, the minimal number of tests is three. Hence, we have to select three tests which each have maximal specificity (distance from the center), and which also have maximal contrast between them (distance between their representative points on the map). If a selection is to be made between one or more equivalent tests (e.g. epinephrine and norepinephrine), the one with the greatest sensitivity is taken. A justifiable choice comprises apomorphine, norepinephrine and tryptamine. These are precisely the specific agonists of the three receptors that have been identified in brain tissue and that bind to neuroleptics, namely the dopaminergic, serotonergic and adrenergic receptors. Although there were no highly specific neuroleptics among the 40 that have been studied in 1965 by Janssen, et al. [10], the three-receptor model that emerged from this analysis was in agreement with the knowledge to date (Fig. 2). The three tests (apomorphine, norepinephrine and tryptamine) that are defined as the minimal set are called poles of the map. They are also referred to as marker variables [32]. A reanalysis of Table 1, using only these three poles, is shown in the SMA biplot of Fig. 11. The horizontal and vertical factors account for 81 and 19% of the global variance of the transformed data (specificities), which amounts to 100% in total. The reading rules of this spectral map are the same as those of the previous one, except for the addition of three bipolar axes. Each of these bipolar axes defines a contrast, namely, apomorphine/norepinephrine (horizontal), tryptamine/norepinephrine (pointing upwards) and apomorphine/tryptamine (pointing downwards). The axes are provided with tick marks and logarithmically spaced scale values, which express the corresponding ratios of effective doses. By perpendicular projection onto a bipolar axis, one can read off the corresponding contrast of each compound. For example, the contrast of isospirilene in the apomorphine and norepinephrine tests amounts to about 400 (exactly 393). This means that isospirilene is about 400 times more active in the apomorphine test, when compared to the norepinephrine test. The contrast of floropipamide in the tryptamine and norepinephrine test is about 1 (exactly 1.08), which means that the compound is about equally active in both tests. But since the norepinephrine test is about 4.4 times more sensitive than the tryptamine test (Table 1), floropipamide

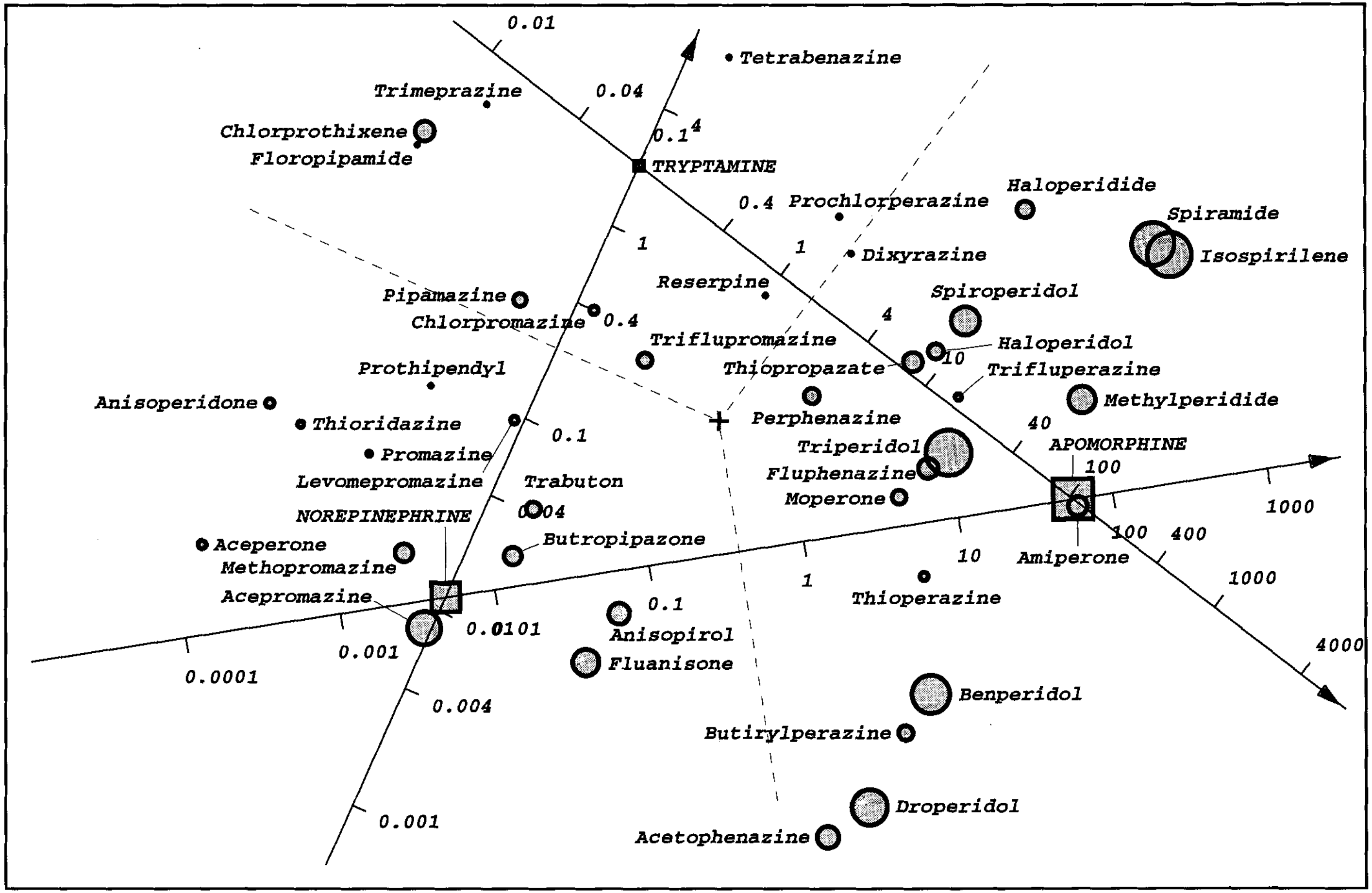


Figure 11. Spectral map of pharmacological activities of 40 neuroleptics in three selected pharmacological tests, as defined in Table 1. The three selected tests are poles of the spectral map of Fig. 10. The three bipolar axes through the tests represent contrasts. Such values are expressed as ratios between activity values in Table 1. The projecition of the center (+) onto a bipolar axis defines the origin of the axis of contrast. All other conventions are the same as in Fig. 10. Note the fair agreement between this map and Fig. 10.

has a highly positive contrast with respect to the tryptamine and norepinephrine tests. Note that the perpendicular projection of the center (+) onto a bipolar axis defines the point of zero contrast. This point separates the regions of positive contrast (in the direction of the arrow) and of negative contrast (in the opposite direction of the arrow). The agreement between the complete analysis based on 12 tests (Fig. 10), and the selective analysis using three poles (Fig. 11), is fair, with a few exceptions (e.g. reserpine and aceperone). This analysis by SMA also supports the three-receptor model suggested by the earlier empirical factor analysis (Fig. 2). Both analyses have been produced by SPECTRAMAP [33], a program for multivariate data analysis with emphasis on graphical representation of the results.*

4.5.5 Mathematical Description of SMA

Let X be an $n \times p$ table of activities (reciprocal effecitive doses or inhibitory concentrations) with the general element, x_{ij}, at the intersection of row i with column j. By convention, the row index, i, refers to one of the n compounds, and the column index, j, labels one of the p tests, although the compounds and tests can be interchanged.

The first step in the analysis is a transformation of the observed activities X into specificities by means of logarithmic reexpression followed by double centering. The choice of the base of the logarithms is not relevant, and for the sake of simplicity we adopted natural logs (base e). The result is an $n \times p$ table of specificities between compounds and tests, $\boldsymbol{Z}$, with the general element, z_{ij}:

$$z_{ij} = \log x_{ij} - \frac{1}{p}\sum_{j}^{p} \log x_{ij} - \frac{1}{n}\sum_{i}^{n} \log x_{ij} + \frac{1}{np}\sum_{i}^{n}\sum_{j}^{p} \log x_{ij} \tag{15}$$

or equivalently:

$$z_{ij} = \log x_{ij} - \log \tilde{x}_{i.} - \log \tilde{x}_{.j} + \log \tilde{x}_{..} = \log \frac{x_{ij}\tilde{x}_{..}}{\tilde{x}_{i.}\tilde{x}_{.j}}$$

where $\tilde{x}_{i.}$, $\tilde{x}_{.j}$ and $\tilde{x}_{..}$ represent the geometric row, column and global means of the elements in X.

The second step involves the application of Singular Value Decomposition (*SVD*) to the table $(np)^{-1/2}\boldsymbol{Z}$, which yields the $n \times r$ matrix, U, of normalized scores, the $p \times r$ matrix, V, of normalized loadings, and the $r \times r$ diagonal matrix, Λ, which contains the associated singular values. The decomposition is defined by means of:

$$(np)^{-1/2} z_{ij} = \sum_{k}^{r} u_{ik} v_{jk} \lambda_{kk} \tag{16}$$

* SPECTRAMAP is a commercial PC software product and a registered trademark of Janssen Pharmaceutica N.V.

where r is the number of singular values that are different from zero, where u_{ik} and v_{jk} are general elements of $\boldsymbol{U}$ and $\boldsymbol{V}$, and where λ_{kk} represents a diagonal element of $\boldsymbol{\Lambda}$. The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are mutually orthogonal, which implies the following:

$$\sum_{i}^{n} u_{ik} u_{ik'} = \delta_{kk'} \tag{17}$$

$$\sum_{j}^{p} v_{jk} v_{jk'} = \delta_{kk'} \tag{18}$$

where the Cronecker symbol, $\delta_{kk'}$, represents 1 if $k = k'$, and 0 otherwise. The columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ are arranged in decreasing order of their associated singular value in $\boldsymbol{\Lambda}$. The corresponding columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ represent the orientations of the common factors in row and column space, respectively. Hence, *SVD* defines an r-dimensional common factor space. The contribution of each factor to the global variance c of $\boldsymbol{Z}$ can be shown to be equal to the square of the associated singular value. Since factors are uncorrelated we have the following:

$$c = \sum_{k}^{r} \lambda_k^2 = \frac{1}{np} \sum_{i}^{n} \sum_{j}^{p} z_{ij}^2 \tag{19}$$

The normalized factor scores, $\boldsymbol{U}$, and the normalized factor loadings, $\boldsymbol{V}$, are also referred to as left- and right-singular vectors, or as normalized row and column principal components. *SVD* of a rectangular table can be computed by means of the Golub and Reinsch algorithm [31]. If only a small number of dominant factors are required, then use can be mase of the iterative NIPALS algorithm, designed by Wold [34].

The coordinates of the n rows (compounds) along the r factors are compiled in the $n \times r$ factor score matrix, $\boldsymbol{S}$, with general elements s_{ik}. Likewise, the coordinates of the p columns (tests), along the same r factors, are compiled in the $p \times r$ factor loadings matrix, $\boldsymbol{L}$, with general element l_{jk}. The factor scores and loadings, $\boldsymbol{S}$ and $\boldsymbol{L}$, are derived from the normalized scores and loadings, $\boldsymbol{U}$ and $\boldsymbol{V}$, by means of appropriate scaling:

$$s_{ik} = n^{1/2} u_{ik} \lambda_{kk}^{\alpha} \tag{20}$$

$$l_{jk} = p^{1/2} v_{jk} \lambda_{kk}^{\beta} \tag{21}$$

where α and β are the factor scaling coefficients for the scores and loadings, respectively.

It should be noted that we employed here a generalized notation, which requires the introduction of the constant weights, $n^{1/2}$ and $p^{1/2}$, for rows and columns of $\boldsymbol{Z}$ in the definition of *SVD*. The reason for our choice is that this notation can be more readily generalized to variable weighting of rows and columns, as will be shown below in the discussion of generalized *SVD*. In the usual case of constant weighting, one may omit the constants related to $n^{1/2}$ and $p^{1/2}$ without violating the validity of the expressions. In this way, one obtains the usual formulas for ordinary *SVD*.

A plot of the rows (compounds) in a low-dimensional factor space is called a scores plot. A plot of the columns (tests) in the same low-dimensional factor space is called a loadings plot. The joint representation of both rows and columns in a common low-dimensional factor space is called biplot [2]. The interpretation of the biplot depends largely on the choice of the factor scaling coefficients, α and β. Briefly, if $\alpha = 1$, then distances between points representing rows and the origin of space can be interpreted as root mean square (*rms*) values of the rows in the table $\boldsymbol{Z}$. If $\beta = 1$, then distances between points representing columns and the origin of space can be interpreted as root mean square (*rms*) values of the columns of the table $\boldsymbol{Z}$. If $\alpha + \beta = 1$, then perpendicular projections of points, representing rows, upon unipolar axes through the columns and the origin reproduce the values in the table $\boldsymbol{Z}$. It is not possible to find a pair (α, β), which allows an interpretation of the distances between rows and between columns, and at the same time, of the projections. In SMA we define the factor scaling coefficients symmetrically as $\alpha = \beta = .5$, which is in favor of the projections at the expense of the distances. The reproduction of the specificities in $\boldsymbol{Z}$ by perpendicular projection can be verified by means of the matrix product of $\boldsymbol{S}$ with (the transpose of) $\boldsymbol{L}$:

$$\sum_k^r s_{ik} l_{jk} = (np)^{1/2} \sum_k^r u_{ik} \lambda_{kk} v_{jk} = z_{ij} \tag{22}$$

using Eqs. (20) and (21), and the assumption that $\alpha + \beta = 1$.

The distortion introduced on the biplot by our choice of factor scaling coefficients, $\alpha = \beta = .5$, is most prominent in the direction of the lower order factors. The degree of distortion can be quantified as follows. Let us assume a biplot which is spanned by the first two factors, with contributions to the global variance λ_1^2 and λ_2^2. The apparent spread along the second factor is then $(\lambda_2/\lambda_1)/^{1/2}$, while the exact spread should be λ_2/λ_1. The degree of distortion is then $(\lambda_1/\lambda_2)^{1/2}$, or $(\lambda_1^2/\lambda_2^2)^{1/4}$.

In this section, we assumed that all compounds and all tests carry an equal weight in the analysis. If a total weight (or mass) of one is assigned to all compounds and to all tests, then each compound carries a constant weight equal to $1/n$, and each test is given a constant weight equal to $1/p$. In the general case of variable weighting, we assigned to each row (compound) a variable weight, w_i, and to each column (test) a variable weight, w_j, such that:

$$\sum_i^n w_i = 1 \quad \text{and} \quad \sum_j^p w_j = 1 \tag{23}$$

These substitutions lead to the definition of generalized *SVD*. In our application, one may choose the weights w_i to be proportional to the potency of compound i, and the weight w_j may be made proportional to the sensitivity of test j. A weighted SMA is then defined by substitution in the previous Eqs. (15) to (22) of the constant $1/n$ by w_i, and of the constant $1/p$ by w_j. (Care should be taken that the variable weights w_i and w_j are within the control of the summations over rows and columns, respectively.) Because of the logarithmic reexpression in SMA, data must be positive. A small number of random zeros can be tolerated, however. (Random zeros arise from the lack of precision of measurements, or from limitation of the sample size.

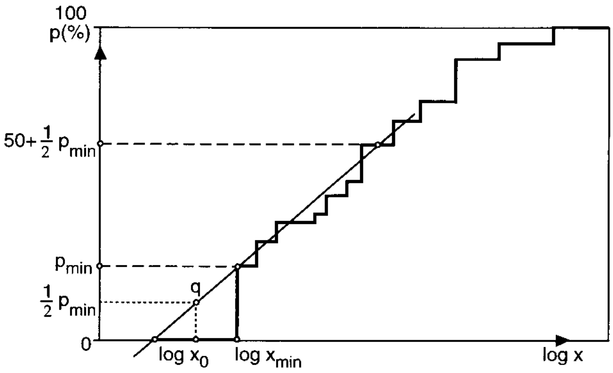


Figure 12. Diagram showing the procedure for obtaining the base number x_0 which is used in the calculation of small positive substitution values for random zeros in a rectangular data table X. The value $x_{\min}$ is the smallest positive value in the table and $p_{\min}$ is the corresponding percentage with respect to the total number of elements in the table.

They contrast with structural zeros, such as in binary or presence/absence data.) The procedure for replacing random zeros is outlined below.

In order to determine substitution values for random zeros, we first constructed the distribution of all values in the original table X on a logarithmic scale (Fig. 12). This distribution is truncated at the lower end, since we assumed the presence of a number of zero values in the table. Let us assume that $x_{\min}$ is the smallest positive value in the table and that the corresponding ordinate value is $p_{\min}$. (In Fig. 12 it is assumed that about 20% of the data are non-positive.) The next step is to extrapolate the lower tail of the distribution by means of a linear regression applied between $p_{\min}$ and the point with an ordinate value at $50 + p_{\min}/2$, halfway between $p_{\min}$ and 100 percent. Next, a point q is determined on the regression line with an ordinate value at $p_{\min}/2$. The abscissa which corresponds to this point q represents the base value, x_0, of the zero substitution. In the final step one computes the substitution value x_{ij}^* at the intersection of a particular row i and column j from:

$$x_{ij}^* = x_0 \frac{\tilde{x}_{i.}\tilde{x}_{.j}}{\tilde{x}_{..}} \tag{24}$$

where $\tilde{x}_{i.}$, $\tilde{x}_{.j}$ and $\tilde{x}_{..}$ are the geometric row, column and global means of the data table X, respectively. It is required further that the substitution value x_{ij}^* is not larger than the smallest positive value $x_{\min}$ in the table divided by 2. The effect of this operation is that random zeros are substituted automatically in a consistent way such that:

$$\frac{x_{ij}^*}{x_{i'j}^*} = \frac{\tilde{x}_{i.}}{\tilde{x}_{i'.}} \quad \text{and} \quad \frac{x_{ij}^*}{x_{ij'}^*} = \frac{\tilde{x}_{.j}}{\tilde{x}_{.j'}} \tag{25}$$

for all i, i' and j, j'.

4.5.6 Discussion

Spectral mapping is based upon a general model, which relates the observed biological activities to the potency of a compound, the sensitivity of a test and the specificity of the compound for that test. In general terms, this relationship can be written as:

$$\log x_{ij} = z_{ij} + \log \pi_i + \log \sigma_j + k \tag{26}$$

where π_i represents the potency of compound i, σ_j the sensitivity of test j, x_{ij} and z_{ij} the activity and specificity of compound i in test j, and where k is a constant. The potency, π, when measured in binding studies (*in vitro*), depends upon the availability of the compound in the immediate vicinity of the receptors. In animal tests (*in vivo*), the potency is also a function of the pharmacokinetic properties of the compound, i.e. its bioavailability, rate of transport into the target tissue, rate of accumulation in fat, adsorption to circulating proteins, metabolism and elimination. A particular compound may be more efficient than others in reaching the appropriate receptors, to which it can bind. In this case, a smaller dose of this compound is required to produce a pharmacological effect, which in turn makes the particular compound more active in the battery of tests. The sensitivity, σ, depends upon the intensity of the response by an organism following the activation of a receptor to which it is coupled. A particular effect may be readily produced, while others require much more stimulation. In this case, a lesser dose of this compound will be required to produce the given effect. As a consequence, this increases the activities of the various compounds in the particular test. The general model is represented schematically in Fig. 13.

The potency, π_i, of compound i is estimated as the geometric mean activity $\bar{x}_{i.}$ of row i in the table of observed activities, X. Similarly, the sensitivity, σ_j, of test j is

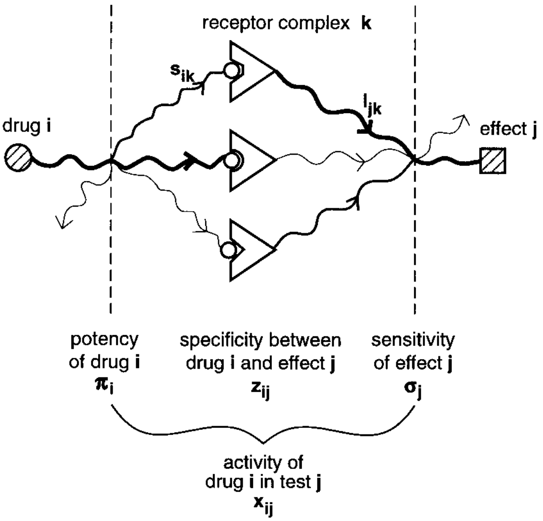

receptor complex k
s_ik
l_jk
drug i
effect j
potency of drug i π_i
specificity between drug i and effect j z_ij
sensitivity of effect j σ_j
activity of drug i in test j x_ij

Figure 13. General model which underlies the procedure of spectral mapping. The potency π_i of a compound i is determined by the physico-chemical and pharmacokinetic properties of the compound i. The sensitivity σ_j of a test j is governed by the physiological and psychological properties of the organism in which the effect is observed. The scores s_{ik} are the coupling coefficients between compound i and receptor complex, k. The loadings l_{jk} are the coupling coefficients between test j and receptor complex, k.

estimated from the geometric mean activity $\tilde{x}_{.j}$ of column j in the same table $\boldsymbol{X}$. Finally, the constant k turns out to be related to the global geometric mean $\tilde{x}_{..}$. In addition, we can decompose the specificities $\boldsymbol{Z}$ between n compounds and p tests into n scores, $\boldsymbol{S}$, of compounds, along r^* structural factors, and into p loadings, $\boldsymbol{L}$, of tests along the same r^* structural factors. This has been explained before. Substitutions of these relations into Eq. (26) leads to an expression for the general model in terms of observed activities:

$$\log x_{ij} \approx \sum_{k}^{r^*} s_{ik} l_{jk} + \log \tilde{x}_{i.} + \log \tilde{x}_{.j} - \log \tilde{x}_{..} \tag{27}$$

The number of structural factors, r^*, has been shown to be equal to the number of operative receptors minus one. It is not possible, therefore, to associate the factors directly to the individual receptors that bind to the compounds and that trigger the responses in the tests. In the neuroleptic case of Table 1, we found that the first factor of the biplot (Fig. 10) is determined by the dopamine and norepinephrine receptors, while the second factor receives contributions from the three receptors, i.e. dopamine, norepinephrine and serotonin. A factor can, thus, be seen as a complex of receptors, each of which contributes to the factor in varying degrees. One can interpret the score s_{ik} as the degree of coupling of compound i with receptor complex k. The degree of coupling, s_{ik}, depends on the ability of the compound to fit to the receptors of complex k, and this, of course, depends upon its steric and electronic properties. In a similar fashion, one can associate the score l_{jk} as the degree of coupling of receptor complex k with test j. The degree of coupling, l_{jk}, is determined by the biochemical and neurological pathways that link the receptor complex k to the organ that produces the observed effect of test j. The varying degrees of coupling are incorporated in the general model of Fig. 13 by means of connecting lines of variable thickness.

It is difficult to assess how original the idea of spectral mapping really was when it was first applied in 1975 to pharmacological data. The method of double centering had been known by psychologists as a combination of *R*-mode and *Q*-mode factor analysis (the distinction being related to column centering or row centering of the data table). The effect of double centering was already recognized by Cronbach and Gleser [35] as a removal of the size component of spectra, but they did not make use of logarithms. A logarithmic double centered transformation has been proposed by the Danish statistician Georg Rasch in 1963 [36] but without a factor analysis of the resulting specificities. Factor analysis has been applied to log double centered data in the social and psychological fields according to Andersen [36] as early as 1966, which is before the discovery of the biplot by Gabriel [2]. Kazmierczak [37], who referred to the method as logarithmic analysis, pointed out that the idea of the log double centered approach was first proposed by the English statistician Udny Yule in the form of an invariance principle. (Every row or column of the data table can be replaced by one that is proportional to it, without affecting the result of the analysis.) Goodman [38] in a review of factor analytic methods of contingency tables and cross-tabulations referred to the log double centered approach as the saturated RC (rows and columns) association model and as the log bilinear model, which is contrasted with Correspondence Factor Analysis (CFA). The analogy between the estimated from the geometric mean activity $\tilde{x}_{.j}$ of column j in the same table $\boldsymbol{X}$. Finally, the constant k turns out to be related to the global geometric mean $\tilde{x}_{..}$. In addition, we can decompose the specificities $\boldsymbol{Z}$ between n compounds and p tests into n scores, $\boldsymbol{S}$, of compounds, along r^* structural factors, and into p loadings, $\boldsymbol{L}$, of tests along the same r^* structural factors. This has been explained before. Substitutions of these relations into Eq. (26) leads to an expression for the general model in terms of observed activities:

$$\log x_{ij} \approx \sum_{k}^{r^*} s_{ik} l_{jk} + \log \tilde{x}_{i.} + \log \tilde{x}_{.j} - \log \tilde{x}_{..} \tag{27}$$

The number of structural factors, r^*, has been shown to be equal to the number of operative receptors minus one. It is not possible, therefore, to associate the factors directly to the individual receptors that bind to the compounds and that trigger the responses in the tests. In the neuroleptic case of Table 1, we found that the first factor of the biplot (Fig. 10) is determined by the dopamine and norepinephrine receptors, while the second factor receives contributions from the three receptors, i.e. dopamine, norepinephrine and serotonin. A factor can, thus, be seen as a complex of receptors, each of which contributes to the factor in varying degrees. One can interpret the score s_{ik} as the degree of coupling of compound i with receptor complex k. The degree of coupling, s_{ik}, depends on the ability of the compound to fit to the receptors of complex k, and this, of course, depends upon its steric and electronic properties. In a similar fashion, one can associate the score l_{jk} as the degree of coupling of receptor complex k with test j. The degree of coupling, l_{jk}, is determined by the biochemical and neurological pathways that link the receptor complex k to the organ that produces the observed effect of test j. The varying degrees of coupling are incorporated in the general model of Fig. 13 by means of connecting lines of variable thickness.

It is difficult to assess how original the idea of spectral mapping really was when it was first applied in 1975 to pharmacological data. The method of double centering had been known by psychologists as a combination of R-mode and Q-mode factor analysis (the distinction being related to column centering or row centering of the data table). The effect of double centering was already recognized by Cronbach and Gleser [35] as a removal of the size component of spectra, but they did not make use of logarithms. A logarithmic double centered transformation has been proposed by the Danish statistician Georg Rasch in 1963 [36] but without a factor analysis of the resulting specificities. Factor analysis has been applied to log double centered data in the social and psychological fields according to Andersen [36] as early as 1966, which is before the discovery of the biplot by Gabriel [2]. Kazmierczak [37], who referred to the method as logarithmic analysis, pointed out that the idea of the log double centered approach was first proposed by the English statistician Udny Yule in the form of an invariance principle. (Every row or column of the data table can be replaced by one that is proportional to it, without affecting the result of the analysis.) Goodman [38] in a review of factor analytic methods of contingency *tables* and cross-tabulations referred to the log double centered approach as the saturated RC (rows and columns) association model and as the log bilinear model, which is contrasted with Correspondence Factor Analysis (CFA). The analogy between the two approaches has been underlined by Escoufier and Junca [39] and Greenacre [40]. In the case of contingency tables, one can show that the results of SMA and CFA converge, if the specificities in the data are small and provided that the potencies and sensitivities are homogeneous. This follows from the fact that Eq. (1) approximates Eq. (2):

$$z_{ij} = \frac{x_{ij}x_{..}}{x_{i.}x_{.j}} - 1 \approx \log \frac{x_{ij}x_{..}}{x_{i.}x_{.j}} \approx \log \frac{x_{ij}\tilde{x}_{..}}{\tilde{x}_{i.}\tilde{x}_{.j}} \tag{28}$$

provided that $x_{ij} \approx x_{i.}x_{.j}/x_{..} \approx \tilde{x}_{i.}\tilde{x}_{.j}/\tilde{x}_{..}$.

A comparison between the performance of SMA and CFA has been made by Thielemans, et al. [41] in the context of epidemiological contingency tables. The scope of SMA, however, is not limited to contingency tables and cross-tabulations, and can be extended to so-called measurement tables. In the latter, columns may be expressed in different units, and the data need not be parts of a whole, i.e. must not add up to meaningful totals, such as in Table 1. Perhaps, the originality of SMA lies in the interpretation of the biplot in terms of bipolar axes, which represent contrasts between rows and between columns and which can be expressed as ratios of elements in the original data table, such as in Fig. 11 [33, 42, 43, 44].

The question that arises is whether these tools of analysis and computation can provide fundamental knowledge that would not be obtainable from careful observation and interpretation. Perhaps they may not, but in any case, they can speed up and enrich the process of interpretation by pointing towards unexpected contrasts, by stimulating relevant questions and by identifying blind alleys. If the biplots indicate that there is little structure in the data then, probably, keen and diligent interpretation will add little to this. On the other hand, if a striking pattern is observed on the biplot, this may point toward an interesting hypothesis which must be confirmed by collateral information and subsequent testing. An illustration of this is given below.

In our virology department an unexpected discovery was made by SMA when analyzing a table of inhibitory concentrations of 15 antiviral compounds in cultures of 100 rhinovirus subtypes (responsible for the symptoms of common cold). The biplot revealed two neatly separated groups of serotypes, each with different specificity for either compounds with aliphatic or for compounds with polycyclic structural fragments. This led to the hypothesis of two structurally different proteins on the viral envelope, which form so-called grooves or canyons, and which function as receptors for antiviral compounds. Once a compound docks into the groove (like a key in a lock), the viral envelope cannot open properly and, hence, the genetic material cannot be used to replicate the virus inside the host cell. Antiviral compounds must, of course, be able to "lock-up" both groups of rhinoviruses in order for the infection not to spread further. As a consequence, they must possess an aliphatic element on one side and a polycyclic structure on the other in order to function as a double key [45, 46]. Of course, the success of the application of SMA depended to a large extent on the interpretation by the virologist of the main factor in terms of a contrast between two distinct receptors.

The approach of SMA can also be extended to multiblock analyses, for example, when the same set of compounds has been studied in different settings such as in animal pharmacology, in radioactive ligand binding experiments and in the clinic. In such a situation, one may be interested in correlations between the different test situations and their predictive capability [47].

In summary, we state that SMA is an effective method for discovering contrasts in pharmacological dose-response data, provided that these are related to specific biological pathways (such as receptors) that mediate between the administration of a drug and the observed effects.

Acknowledgement

The author is grateful to A. Biermans and K. Sprangers for the preparation of this manuscript, to A. Megens, K. Engelen and T. Amery for proofreading, to J. Van Hoof and K. Van Reusel for the computer programming and to J. Van Mierlo for help with the figures.

References

[1] Jolicoeur, P., and Mosimann, J. E., *Growth* **24**, 339 – 354 (1960)
[2] Gabriel, K. R., *Biometrika* **58**, 453 – 467 (1971)
[3] Benzécri, J.-P., *L'Analyse des Correspondances. L'Analyse des Données, Vol.* II. Dunod, Paris, 1973
[4] Iverson, K., *A Programming Language*, J. Wiley, New York, 1962
[5] Van Wijngaarden, I., *Pharm. Intern.*, 26 – 28 (February 1980)
[6] Delay, J., Deniker, P., and Harl, J. M., *Ann. Méd. Psych.* **110**, 112 – 131 (1952)
[7] Janssen, P. A. J., *Arzneim. Forsch./Drug Res.* **11**, 819 – 824; 932 – 938 (1961)
[8] Mayer, T., *De Affinitate Colorum Dissertatio.* Unpublished notes, Göttingen (1758). In: *Opera inedita Tobiae Mayeri*, Published by Lichtenberg, G.C., Göttingen, 1775.
[9] Kolodkine, P., *Dmitri Mendeléiev et la Loi Périodique*, Seghers, Paris, 1963
[10] Janssen, P. A. J., Niemegeers, C. J. E., and Schellekens, K. H. L., *Arzneim. Forsch./Drug Res.* **15**, 104 – 117 (1965)
[11] Garfield, E., *Current Contents* **27**, 17 (1986)
[12] Lambert, P. A., and Revol, K., *Presse Méd.* **68**, 1509 (1960)
[13] Carlsson, A., and Lindqvist, M., *Acta Pharmacol. Toxicol.* **20**, 140 – 144 (1963)
[14] Woolley, D., *The Biological Bases of Psychoses*, J. Wiley, New York, 131 (1962)
[15] de Vries, S. E., and Megens, A., *Janssen Medical Scientific News* **8**, 191 – 194 (1993)
[16] Fisher, E., *Ber. Deutsch. Chem. Gesellsch.* **27**, 2985 – 2993 (1894). Cited by Holmstedt, B., and Liljestrand, G., *Readings in Pharmacology*, Raven Press, New York, 251 (1981)
[17] Janssen, P. A. J., Niemegeers, C. J. E., Schellekens, K. H. L., Dresse, A., Lenaerts, F. M., Pinchard, A., Schaper, W. K. A., Van Nueten, J. M., and Verbruggen, F. J., *Arzneim. Forsch./Drug Res.* **18**, 261 – 287 (1968)
[18] Seeman, P., Chau-Wong, M., Tedesco, J., and Wong, K., *Proc. Natl. Acad. Sci. USA* **72**, 4376 – 4380 (1975)
[19] Creese, I., Burt, D. R., and Snyder, S. H., *Life Sci.* **17**, 993 – 1002 (1975)
[20] Leysen, J., and Laduron, P., *Arch. Int. Pharmacodyn. Thérap.* **230**, 337 – 339 (1977)
[21] Greenberg, D. A., U'Prichard, D. C., and Snyder, S. H., *Life Sci.* **29**, 557 – 562 (1976)
[22] Niemegeers, C. J. E., Lenaerts, F. M., Artois, K. J. K., and Janssen, P. A. J., *Arch. Int. Pharmacodyn. Therap.* **227**, 238 – 253 (1977)
[23] Kvalheim, O. M., *The Latent Variable (Factor) Approach to the Analysis of Multivariate Data: History, Philosophy and Scientific Implications.* In: *Understanding and History in Arts and Science.* Studies in honor of Richard Holten Pearce, University of Bergen (1990)

[24] Desmedt, L. K. C., Niemegeers, C. J. E., Lewi, P. J., and Janssen, P. A. J., *Arzneim. Forsch./Drug Res.* **26**, 1592–1603 (1976)

[25] Dunteman, G. H., *Principal Components Analysis*, Sage, Newbury Park, 1989

[26] Joliffe, I. T., *Principal Components Analysis*, Springer, New York, 1986

[27] Lacourly, G., and Lebart, L., *Analyse Multidimensionale Interactive d'un Ensemble de Données*, Rapport CEGOS Informatique et CREDOC, Paris, 1974

[28] Fechner, G. T., *Elemente der Psychophysik*, (1907). *Elements of Psychophysics*. Reprinted by: Davis, D. H., ed., Holt, Rinehart and Winston, New York, 1966

[29] Finney, D. J., *Probit analysis*, 3rd Ed., Cambridge Univ. Press, Cambridge, 1971

[30] Lewi, P. J., *Arzneim. Forsch./Drug Res.* **26**, 1295–1300 (1976)

[31] Golub, G. H., and Reinsch, C., *Numer. Math.* **14**, 403–420 (1970)

[32] Kvalheim, O. M., *Chemom. Intell. Lab. Syst*, **4**, 11–25 (1988)

[33] Lewi, P. J., *Eur. J. Med. Chem.* **21**, 155–162 (1986)

[34] Wold, H., *Soft Modeling by Latent Variable, the Non-Linear Iterative Partial Least Squares (NIPALS) Algorithm*. In: *Perspectives in Probability and Statistics*, Gani, J., ed., Academic Press, London, 1975, p. 117–142

[35] Cronbach, L. J., and Gleser, G. C., *Psychol. Bull.* **50**, 456–473 (1953)

[36] Andersen, E. B., *The Discrete Measurement Model of Finite Order with Applications to a Social Psychological Data Set*, Copenhagen, Danish Government, Printing Office, 1966 (in Danish). Cited by Andersen, E. B., In: Discussion of paper by Goodman, L. A. *Int. Statistical Review* **54**, 243–309 (1986)

[37] Kazmierczak, J. B., *Revue Stat. Appl.* **33**, 13–24 (1985)

[38] Goodman, L. A., *Int. Stat. Rev.* **54**, 243–309 (1986)

[39] Escoufier, Y., and Junca, S., *Int. Stat. Rev.* **54**, 279–283 (1986)

[40] Greenacre, M. J., *Theory and Applications of Correspondence Analysis*, Academic Press, London, 1984

[41] Thielemans, A., Lewi, P. J., and Massart, D. L., *Chemom. Intell. Lab. Syst.* **3**, 277–300 (1988)

[42] Lewi, P. J., *Multivariate Analysis in Industrial Practice*, Research Studies Press, J. Wiley, Chichester, 1982

[43] Lewi, P. J., *Chemom. and Intel. Lab. Syst.* **5**, 105–116 (1989)

[44] Lewi, P. J., *SPECTRAMAP, Introduction to Multivariate Analysis of Rectangular Data Tables, with special Emphasis on Biplots*. Janssen Pharmaceutica Report (1993), Beerse, Belgium

[45] Andries, K., Dewindt, B., Snoeks, J., Wouters, L., Moereels, H., Lewi, P. J., and Janssen, P. A. J., *J. Virology* **64**, 1117–1123 (1990)

[46] Lewi, P. J., Van Hoof, J., and Andries, K., *The Role of Exploratory Data Analysis in the Development of Novel Antiviral Compounds*. In: *Scientific Computing and Automation (Europe)*, Karjalainen, E. J., ed., Elsevier, Amsterdam (1990) p. 97–103

[47] Lewi, P. J., Vekemans, B., and Gypen, L. M., *Partial Least Squares (PLS) for the Prediction of Real-Life Performance from Laboratory Results*. In: *Scientific Computing and Automation (Europe)*, Karjalainen, E. J., ed., Elsevier, Amsterdam (1990) p. 199–209

4.6 Display of Multivariate Data Using Non-Linear Mapping

James Devillers

Abbreviations and Symbols

CFA	Correspondence factor analysis
C_i	Individual contribution to the distances in the non-linear map
d_{ij}	Euclidean interpoint distance in the display space
d_{ij}^*	Euclidean interpoint distance in the original space
$(d_{ij}^*)^p$	Weighting factor
E	Mapping error
E_i	Individual mapping error
NLM	Non-linear mapping
PC(s)	Principal component(s)
PCA	Principal components analysis
QSAR	Quantitative structure-activity relationship
SAR	Structure-activity relationship

4.6.1 Linear and Non-Linear Methods in SAR and QSAR studies

Linear multivariate methods such as principal components analysis (PCA) and correspondence factor analysis (CFA) are now widely used in medicinal chemistry and related disciplines for deriving structure-activity relationships (SAR), [see e.g. Refs. 1 – 4]. However, it is obvious that biological activities of molecules may not be just related to topological and/or physico-chemical descriptors by means of linear relationships. Therefore, non-linear multivariate methods can also lead to interesting SAR conclusions [5 – 7]. Besides the linear methods, they can be an additional or complementary source of information about the relationships in the data. Among the non-linear methods available for multivariate data analysis, [see e.g. Refs. 8, 9], the non-linear mapping (NLM) method [10] is very useful for the reduction of dimensionality and visualization of multivariate data [11 – 15]. Under these conditions, the aim of this paper is first, to present the statistical principles of the NLM method, and then to underline the heuristic capability of this particular statistical analysis in medicinal chemistry.

4.6.2 Non-Linear Mapping Algorithms

The non-linear mapping (NLM) method was designed by Sammon [10] to visually represent a set of points defined in an n-dimensional space by a configuration of the data in a lower d-dimensional display space ($d = 2$ or 3). The principal feature of this method is that it tries to preserve as much as possible the distances between the points in the display space similar to the actual distances in the original space. The procedure for performing this transformation can be summarized as follows.

A.

Interpoint distances in the original space are computed. The Euclidean distance is the most widely used, but any distance measure is suitable for NLM, as long as it is monotonic and the derivative of the mapping error (E) exists (e.g. the Hamming distance which can save valuable time [16]).

B.

An initial configuration of the points (generally random) in the display space is chosen. Several authors have proposed to use the co-ordinates of points of the first principal components (PCs) as the initial configuration [12, 17]. However, it is always highly recommended to perform several trials, either with random configurations, or with the other PC co-ordinates [12].

C.

A mapping error (E) is calculated from the distances in the two spaces. The original mapping error (E) calculation for NLM, devised by Sammon [10] on the basis of the Euclidean distance, is stated as follows (Eq. (1)):

$$E = \frac{1}{\sum\limits_{i<j}^{N} d_{ij}^{*}} \sum_{i<j}^{N} \frac{(d_{ij}^{*} - d_{ij})^2}{d_{ij}^{*}} \tag{1}$$

where d_{ij}^{*} and d_{ij} are the Euclidean interpoint distances in the original and display spaces, respectively. The procedure proposed by Sammon [10] has been significantly modified by Kowalski and Bender [18]. They defined the mapping error function, E, as shown in Eq. (2), where $(d_{ij}^{*})^{\rho}$ is a weighting factor:

$$E(\rho) = \sum_{i<j}^{N} \frac{(d_{ij}^{*} - d_{ij})^2}{(d_{ij}^{*})^{\rho}} \tag{2}$$

Indeed, the ρ value may be adjusted so as to underline particular features in the data set. A value of $\rho = 2$ corresponds to an equal weighting of small and large distances. When $\rho = -2$, the larger distances are preserved at the expense of the smaller distances [18]. The above equation, Eq. (2), with $\rho = 2$ is one of the most widely used equations. However, other types of mapping errors have been used [12].

D.

Co-ordinates of points in the display space are iteratively modified by means of a non-linear procedure, so as to minimize the mapping error. Various minimization algorithms can be used. Thus, for example, Sammon [10] preferred the "steepest descent procedure", Kowalski and Bender [18] adopted the Polak-Ribière method, and Klein and Dubes [19] proposed simulated annealing. The algorithm terminates when no significant decrease in the mapping error is obtained over the course of several iterations [12, 19, 20].

Additional information on the conceptual and theoretical aspects of the NLM method can be found in an earlier paper by the author and coworkers [12].

4.6.3 Interpretation of Non-Linear Maps

The problem of the quality of the representation of each observation holds for all linear (e.g. PCA) and non-linear mapping methods. It is obvious that the interpretation of the maps must always be accompanied by an inspection of some valuable statistical parameters (e.g. absolute and relative contributions in the case of CFA [3]). For the interpretation of non-linear maps, two statistical parameters describing the quality of the representation (E_i) and the contribution to the distances of each point (C_i) were recently proposed [12].

In the case of Sammon's error [10], E_i, which estimates the goodness of fit for each observation, is calculated from Eq. (3):

$$E_i = \frac{1}{2 \times \sum_{i<j}^{N} d_{ij}^*} \sum_{j}^{N} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \tag{3}$$

By definition, the sum of all the individual mapping errors equals the total mapping error, E.

The statistical parameter, C_i, can be defined (Eq. (4)) as the sum of all distances between a point i and all the others in the display space, divided by the sum of all distances between all points in the display space [12]. The sum of the C_i values equals one.

$$C_i = \frac{1}{\sum_{i,j}^{N} d_{ij}} \sum_{j}^{N} d_{ij} \tag{4}$$

4.6.4 Drawbacks and Limitations

The NLM method has two major drawbacks. First, the maps obtained are unique. This means that new objects cannot be directly plotted onto the map without recomputation, since interpoint distances are interdependent. Furthermore, maps depend on the initial configuration, in the display space, since the minimization process finds the nearest local minimum rather than the global minimum. Second, even if the cost of computation is constantly decreasing, it is important to stress

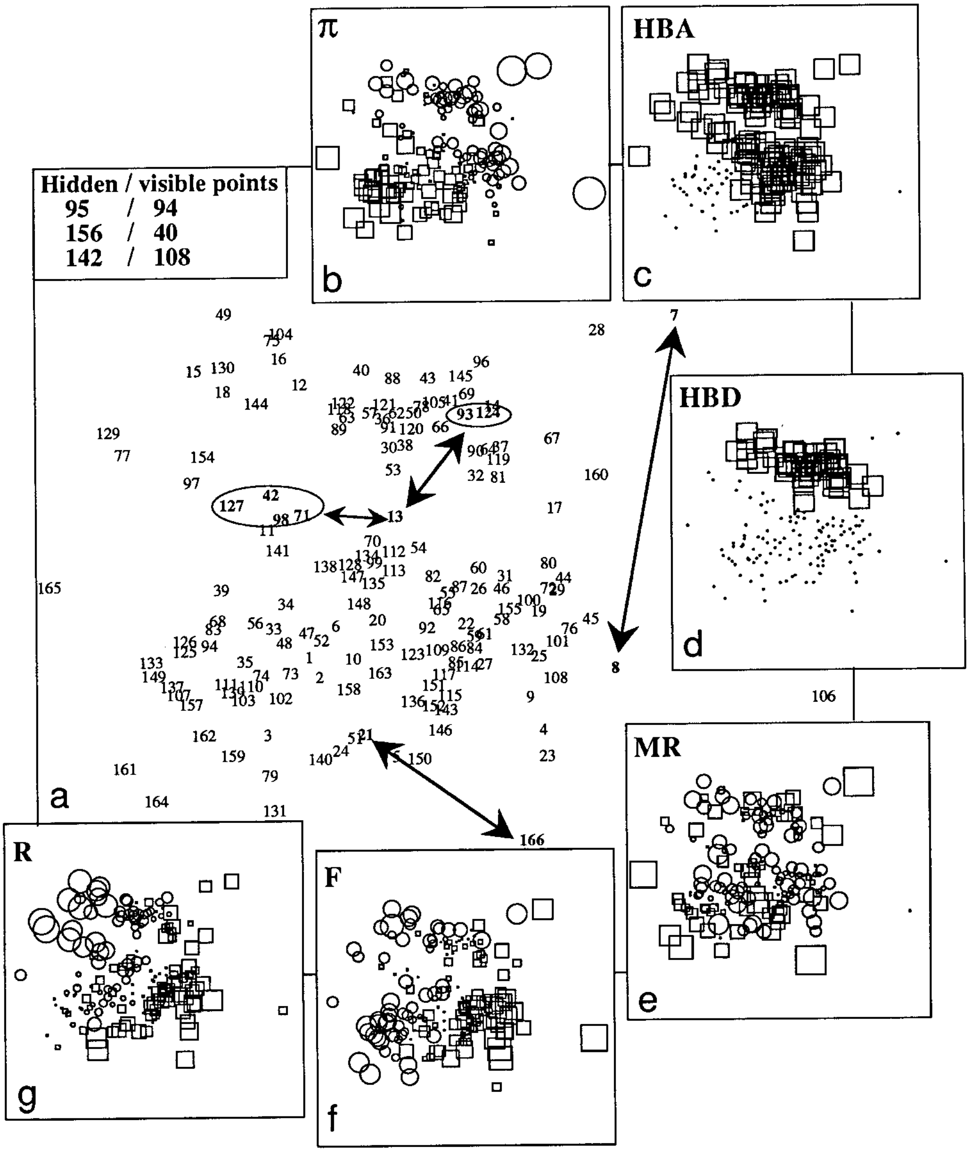


Figure 1. a) Non-linear map of the 166 aromatic substituents described by six substituent constants (π, *HBA*, *HBD*, *MR*, *F*, and *R*). b) to g) Plot of the scaled values of the six parameters used to describe the substituents on each point of the non-linear map. Squares (positive values) and circles (negative values) are proportional in size to the magnitude of the parameters. In Fig. 1c and 1d the dots indicate the substituents which do not have the ability to accept and donate H-bonds, respectively. See Table 1 for the numbering of substituents.

that the computation time to obtain a good NLM configuration using microcomputers, can be long, when the number of objects considered is large.

Some solutions have been devised to overcome the above two problems. Thus, for example, it has been proposed to use PC scores as an initial configuration for solving the problem of uniqueness of the maps obtained. It has also been shown that when new objects, not too dissimilar from the original training set, were introduced, the maps obtained tended to be quite stable [11]. In order to reduce the time for computation, procedures, using two or more base points as reference for the map, have been devised [21 – 23].

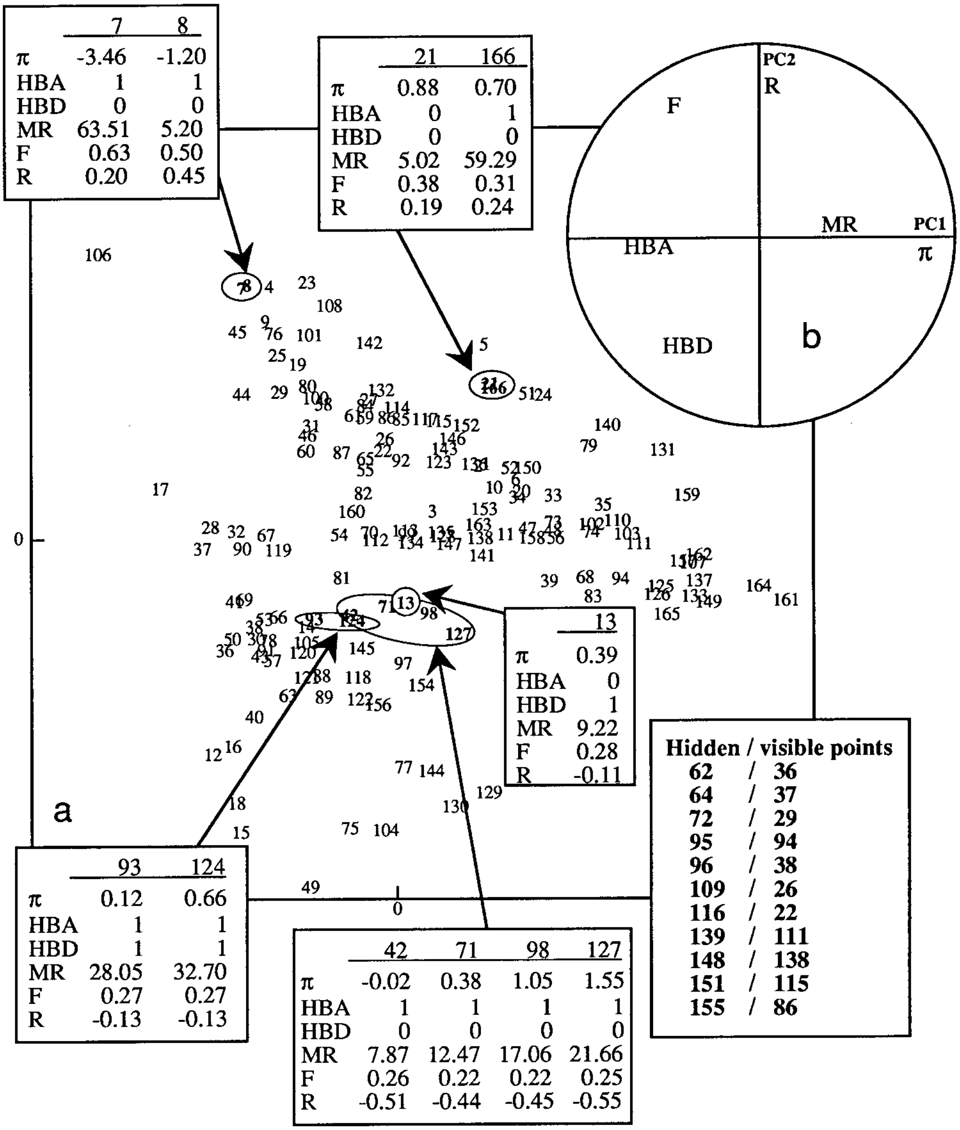
4.6.5 An Illustrative Example in Medicinal Chemistry

In medicinal chemistry, the selection of optimal test series is essential, since the design of a new drug is extremely costly (ca. $ 100 million [24]). A considerable amount of work has been directed towards this aim and many authors have proposed different methods, based on simple 2D plots, decision trees or multivariate analyses (for a review of these methods see Ref. 24). Recently [25], we have shown that the NLM method was particularly suitable for the rational selection of test series and for deriving structure-activity relationships from graphical representations. Fig. 1a clearly illustrates this principle and represents the non-linear map of 166 aromatic substituents (Table 1), described by six substituent constants (i.e. π, HBA, HBD, MR, F and R), encoding hydrophobic, steric, and electronic effects [26]. With a low mapping error (i.e. 6.4e-2), we can advance that the main information contained in the original data matrix (166×6) is summarized on Fig. 1a. This can also be shown by calculating the E_i values (not given here) and by plotting the values used for the NLM analysis (i.e. centered and reduced for π, MR, F and R; reduced for HBA and HBD) onto the non-linear map by means of squares (positive values) and circles (negative values), whose sizes are proportional to the magnitude of the parameters studied. Indeed, Fig. 1b to 1g allow a clear interpretation of the location of the points in Fig. 1a. It is important to note that a PCA performed on the standardized data does not allow all the information contained in the original data matrix to be summarized in one plane only. Thus, for example, Fig. 2a, which represents the $PC1 - PC2$ plane, shows substituent **7** as being near to substituent **8**, while they are actually different and should be distant from each other as shown on Fig. 1a. The same remark can be made for substituents **21** and **166** as well as for the series **13/42**, **71**, **98**, and **127/93** and **124** (compare Fig. 2a with Fig. 1a).

From these results, it is obvious that for selecting test series with high information content, the use of Fig. 1a can be particularly suitable, since this selection can be performed by a simple visual inspection of the map.

Table 1. List of the 166 aromatic substituents [26]

N°	substituent	N°	substituent	N°	substituent
1	Br	**2**	Cl	**3**	F
4	SO_2F	**5**	SF_5	**6**	I
7	IO_2	**8**	NO	**9**	NO_2
10	NNN	**11**	H	**12**	OH
13	SH	**14**	$B(OH)_2$	**15**	NH_2
16	NHOH	**17**	SO_2NH_2	**18**	$NHNH_2$
19	5-Cl-1-Tetrazolyl	**20**	$N=CCl_2$	**21**	CF_3
22	OCF_3	**23**	SO_2CF_3	**24**	SCF_3
25	CN	**26**	NCS	**27**	SCN
28	CO_2^-	**29**	1-Tetrazolyl	**30**	NHCN
31	CHO	**32**	CO_2H	**33**	CH_2Br
34	CH_2Cl	**35**	CH_2I	**36**	NHCHO
37	$CONH_2$	**38**	$CH=NOH$	**39**	CH_3
40	$NHCONH_2$	**41**	$NHC=S(NH_2)$	**42**	OCH_3
43	CH_2OH	**44**	$SOCH_3$	**45**	SO_2CH_3
46	OSO_2CH_3	**47**	SCH_3	**48**	$SeCH_3$
49	$NHCH_3$	**50**	$NHSO_2CH_3$	**51**	CF_2CF_3
52	$C\equiv CH$	**53**	$NHCOCF_3$	**54**	CH_2CN
55	$CH=CHNO_2$-(*trans*)	**56**	$CH=CH_2$	**57**	$NHC=O(CH_2Cl)$
58	$COCH_3$	**59**	$SCOCH_3$	**60**	$OCOCH_3$
61	CO_2CH_3	**62**	$NHCOCH_3$	**63**	$NHCO_2CH_3$
64	$C=O(NHCH_3)$	**65**	$CH=NOCH_3$	**66**	$NHC=S(CH_3)$
67	$CH=NNHC=S(NH_2)$	**68**	CH_2CH_3	**69**	$CH=NNHCONHNH_2$
70	CH_2OCH_3	**71**	OCH_2CH_3	**72**	SOC_2H_5
73	SC_2H_5	**74**	SeC_2H_5	**75**	NHC_2H_5
76	$SO_2C_2H_5$	**77**	$N(CH_3)_2$	**78**	$NHSO_2C_2H_5$
79	$P(CH_3)_2$	**80**	$PO(OCH_3)_2$	**81**	$C(OH)(CF_3)_2$
82	$CH=CHCN$	**83**	Cyclopropyl	**84**	COC_2H_5
85	$SCOC_2H_5$	**86**	$CO_2C_2H_5$	**87**	$OCOC_2H_5$
88	$CH_2CH_2CO_2H$	**89**	$NHCO_2C_2H_5$	**90**	$CONHC_2H_5$
91	$NHCOC_2H_5$	**92**	$CH=NOC_2H_5$	**93**	$NHC=S(C_2H_5)$
94	$CH(CH_3)_2$	**95**	C_3H_7	**96**	$NHC=S(NHC_2H_5)$
97	$OCH(CH_3)_2$	**98**	OC_3H_7	**99**	$CH_2OC_2H_5$
100	SOC_3H_7	**101**	$SO_2C_3H_7$	**102**	SC_3H_7
103	SeC_3H_7	**104**	NHC_3H_7	**105**	$NHSO_2C_3H_7$
106	$N(CH_3)_3^+$	**107**	$Si(CH_3)_3$	**108**	$CH=C(CN)_2$
109	1-Pyrryl	**110**	2-Thienyl	**111**	3-Thienyl
112	$CH=CHCOCH_3$	**113**	$CH=CHCO_2CH_3$	**114**	COC_3H_7
115	$SCOC_3H_7$	**116**	$OCOC_3H_7$	**117**	$CO_2C_3H_7$
118	$(CH_2)_3CO_2H$	**119**	$CONHC_3H_7$	**120**	$NHCOC_3H_7$
121	$NHC=OCH(CH_3)_2$	**122**	$NHCO_2C_3H_7$	**123**	$CH=NOC_3H_7$
124	$NHC=S(C_3H_7)$	**125**	C_4H_9	**126**	$C(CH_3)_3$
127	OC_4H_9	**128**	$CH_2OC_3H_7$	**129**	$N(C_2H_5)_2$
130	NHC_4H_9	**131**	$P(C_2H_5)_2$	**132**	$PO(OC_2H_5)_2$
133	$CH_2Si(CH_3)_3$	**134**	$CH=CHCOC_2H_5$	**135**	$CH=CHCO_2C_2H_5$
136	$CH=NOC_4H_9$	**137**	C_5H_{11}	**138**	$CH_2OC_4H_9$
139	C_6H_5	**140**	$N=NC_6H_5$	**141**	OC_6H_5
142	$SO_2C_6H_5$	**143**	$OSO_2C_6H_5$	**144**	NHC_6H_5
145	$NHSO_2C_6H_5$	**146**	2,5-Dimethyl-1-pyrryl	**147**	$CH=CHCOC_3H_7$
148	$CH=CHCO_2C_3H_7$	**149**	Cyclohexyl	**150**	2-Benzthiazolyl
151	COC_6H_5	**152**	$CO_2C_6H_5$	**153**	$OCOC_6H_5$
154	$N=CHC_6H_5$	**155**	$CH=NC_6H_5$	**156**	$NHCOC_6H_5$
157	$CH_2C_6H_5$	**158**	$CH_2OC_6H_5$	**159**	$C\equiv CC_6H_5$
160	$CH=NNHCOC_6H_5$	**161**	$CH_2Si(C_2H_5)_3$	**162**	$CH=CHC_6H_5$-(*trans*)
163	$CH=CHCOC_6H_5$	**164**	Ferrocenyl	**165**	$N(C_6H_5)_2$
166	$P=O(C_6H_5)_2$				



	7	8
π	-3.46	-1.20
HBA	1	1
HBD	0	0
MR	63.51	5.20
F	0.63	0.50
R	0.20	0.45

	21	166
π	0.88	0.70
HBA	0	1
HBD	0	0
MR	5.02	59.29
F	0.38	0.31
R	0.19	0.24

	13
π	0.39
HBA	0
HBD	1
MR	9.22
F	0.28
R	-0.11

Hidden	/ visible points
62	/ 36
64	/ 37
72	/ 29
95	/ 94
96	/ 38
109	/ 26
116	/ 22
139	/ 111
148	/ 138
151	/ 115
155	/ 86

	93	124
π	0.12	0.66
HBA	1	1
HBD	1	1
MR	28.05	32.70
F	0.27	0.27
R	-0.13	-0.13

	42	71	98	127
π	-0.02	0.38	1.05	1.55
HBA	1	1	1	1
HBD	0	0	0	0
MR	7.87	12.47	17.06	21.66
F	0.26	0.22	0.22	0.25
R	-0.51	-0.44	-0.45	-0.55

Figure 2. a) Score map (*PC*1 – *PC*2 plane). *PC*1 and *PC*2 account for 32% and 26% of the total variance, respectively. b) Correlation circle. See Table 1 for the numbering of substituents.

4.6.6 Software Availability

The number of statistical packages, including the NLM method, is rather small. Among them, we can cite ARTHUR [27], DISCLOSE [28], ISPAHAN [29], and STATQSAR [30]. This last package integrates a special module which is dedicated to the use of NLM techniques in SAR and QSAR studies. Besides these computer programs, some easily implementable algorithms are available in the literature [10, 12]. Similarly, Zitko [31] published the listings of two programs written in HP 3000 BASIC.

4.6.7 Concluding Remark

Numerous studies deal with the comparison of the performances of the NLM method with other non-linear and/or linear mapping techniques [10, 29, 32–34]. From the published results, it is often difficult to obtain a fair estimation of their relative efficiency and usefulness, since they are most often presented from artificially generated data sets or from well-known real data sets (e.g. iris data [35]). In all cases, from a practical point of view [e.g. 12–15, 25], we think that the NLM method should be seen as a valuable additional tool in the kit of the classical multivariate analyses available for deriving structure-activity and structure-property relationships.

References

[1] Ojasoo, T., Doré, J. C., Gilbert, J., and Raynaud, J. P., *J. Med. Chem.* **31**, 1160–1169 (1988)

[2] van de Waterbeemd, H., El Tayar, N., Carrupt, P. A., and Testa, B., *J. Comput.-Aided Mol. Design* **3**, 111–132 (1989)

[3] Devillers, J., and Karcher, W., Correspondence Factor Analysis as a Tool in Environmental SAR and QSAR Studies. In: *Practical Applications of Quantitative Structure-Activity Relationships (QSAR) in Environmental Chemistry and Toxicology*. Karcher, W., and Devillers, J., eds., Kluwer Academic Publishers, Dordrecht (1990) p. 181–195

[4] Domine, D., Devillers, J., Chastrette, M., and Karcher, W. *Pestic. Sci.* **35**, 73–82 (1992)

[5] Rose, V. S., Croall, I. F., and MacFie, H. J. H., *Quant. Struct. – Act. Relat.* **10**, 6–15 (1991)

[6] Wiese, M., and Schaper, K. J., *SAR QSAR Environ. Res.* **1**, 137–152 (1993)

[7] Domine, D., Devillers, J., Chastrette, M., and Karcher, W., *SAR QSAR Environ. Res.* **1**, 211–219 (1993)

[8] Gifi, A., *Nonlinear Multivariate Analysis*, John Wiley, Chichester, 1990

[9] Pao, Y. H., *Adaptive Pattern Recognition and Neural Networks*, Addison Wesley, Reading, 1989

[10] Sammon, J. W., *IEEE Trans. Comput.* **C-18**, 401–409 (1969)

[11] Livingstone, D. J., *Pestic. Sci.* **27**, 287–304 (1989)

[12] Domine, D., Devillers, J., Chastrette, M., and Karcher, W., *J. Chemometrics* **7**, 227–242 (1993)

[13] Chastrette, M., Devillers, J., Domine, D., and de Saint Laumer, J. Y., *New Tools for the Selection and Critical Analysis of Large Collections of Data.* (13th International CODATA Conference, Beijing, 19–22 October 1992)

[14] Devillers, J., Domine, D., Chastrette, M., and Karcher, W., *Multivariate Analyses in Pesticide Research* (5th International Congress on Ecotoxicology of Pesticides, Riva del Garda, 22–25 October 1992)

[15] Domine, D., Devillers, J., Garrigues, P., Budzinski, H., Chastrette, M., and Karcher, W., *Sci. Total Environ.* **155**, 9–24 (1994)

[16] White, I., *IEEE Trans. Comput.* **C-21**, 220–221 (1972)

[17] Hudson, B., Livingstone, D. J., and Rahr, E., *J. Comput.-Aided Mol. Design* **3**, 55–65 (1989)
[18] Kowalski, B. R., and Bender, C. F., *J. Am. Chem. Soc.* **95**, 686–692 (1973)
[19] Klein, R. W., and Dubes, R. C., *Pattern Recognition* **22**, 213–220 (1989)
[20] Valko, K., Cserhati, T., and Forgacs, E., *J. Chromatogr.* **550**, 667–675 (1991)
[21] Forina, M., Armanino, C., Lanteri, S., and Calcagno, C., *Ann. Chim.* **73**, 641–657 (1983)
[22] Lin, C. H., and Chen, H. F., *Anal. Chem.* **49**, 1357–1363 (1977)
[23] Drack, H., *Anal. Chem.* **50**, 2147 (1978)
[24] Pleiss, M. A., and Unger, S. H., The Design of Test Series and the Significance of QSAR Relationships. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Ramsden, C.A., ed., Pergamon Press, Oxford (1990) p. 561–587
[25] Domine, D., Devillers, J., and Chastrette, M., *J. Med. Chem.* **37**, 973–980 (1994)
[26] Hansch, C., and Leo, A., *Substituent Constants for Correlation Analysis in Chemistry and Biology*, John Wiley, New York, 1979
[27] Harper, A. M., Duewer, D. L., Kowalski, B. R., and Fasching, J. L., *ACS Symp. Ser.* **52**, 14–52 (1977)
[28] Bawden, D., *Anal. Chim. Acta* **158**, 363–368 (1984)
[29] Gelsema, E. S., and Eden, G., *Pattern Recognition* **12**, 127–136 (1980)
[30] STATQSAR; CTIS, Lyon, France (1993)
[31] Zitko, V., *Multidimensional Data Display by Nonlinear Mapping*, Report n°. 1428 (1986), Biological Station, St. Andrews, Canada
[32] Pykett, C. E., *Electronics Lett.* **14**, 799–800 (1978)
[33] Siedlecki, W., Siedlecka, K., and Sklansky, J., *Pattern Recognition* **21**, 411–429 (1988)
[34] Siedlecki, W., Siedlecka, K., and Sklansky, J., *Pattern Recognition* **21**, 431–438 (1988)
[35] Fisher, R. A., *Ann. Eugen.* **7**, 179–188 (1936)

4.7 The Use of Canonical Correlation Analysis

Martyn Glenn Ford and David William Salt

Abbreviations

CCA	Canonical correlation analysis
CV	Canonical variate
MRA	Multiple regression analysis
PCA	Principal component analysis
QSAR	Quantitative structure-activity relationship

Symbols

a_{ij}	Coefficient of the *j*-th response variable in the *i*-th CV
$B1$	Verloop steric parameter (breadth 1)
$B4$	Verloop steric parameter (breadth 4)
b_{ij}	Coefficient of the *j*-th descriptor variable in the *i*-th CV
C_{XY}	Correlation matrix between descriptor and response variables
CH_2	H-nmr chemical shift of the benzylic methylene
$CNVRFi$	*i*-th canonical variate first set
$CNVRSi$	*i*-th canonical variate second set
ED_{50}	Dose required to affect 50% of treated insects
$\mathrm{k_e}$	Elimination rate constant
k_p	Penetration rate constant
KD_{50}	Dose required to knock down 50% of treated insects
κ	Livingstone's charge transfer constant
L	Verloop steric parameter (length)
LD_{50}	Dose required to kill 50% of treated insects
log *cis*	Proportion of *cis*-isomer in the mixed ester
λ	Pharmacokinetic distribution coefficient
λ_i	*i*-th largest eigenvalue of C_{XY}
MTC	Molar threshold concentration
p	Number of descriptor variables
π	Partition term
q	Number of response variables
R^2	Coefficient of determination
$R_{C_{y.x}}$	Redundancy coefficient

R_{ci}	Canonical correlation between W_i and Z_i
$[R_X(j)]^2$	Proportion of variance in descriptor variables accounted for by the j-th CV
$r_{X_i}(j)$	Canonical loading of the ith descriptor variable onto the j-th CV
$[R_Y(j)]^2$	Proportion of variance in response variables accounted for by the j-th CV
$r_{Y_i}(j)$	Canonical loading of the i-th response variable on the jth CV
s	The smaller of p and q
S_2	Steric effect
S_4	Steric parameter describing esterification of the substituted benzyl alcohol
σ	Hammett constant
W_i	i-th canonical variate first set
X_i	i-th descriptor variable
χ_j	j-th test statistic for Bartlett's test
x_{ij}	i-th value of the j-th descriptor variable
Y_i	i-th response variable
Z_i	i-th canonical variate second set

4.7.1 Introduction

In studies of structure/activity relationships, several biological responses may be measured. For example, different potencies, each related to a different biological response (e.g. $1/ED_{50}$, $1/LD_{50}$, $1/KD_{50}$, etc.) could be estimated for each test compound, examined during the development of a new drug or agrochemical. It may then be necessary to determine whether relationships exist between two sets of variables, the biological potencies (Set 1) and the chemical/molecular properties (Set 2).

One approach is to employ canonical correlation analysis (CCA). CCA is a technique which determines the linear combination of the response variables that is maximally correlated (ordinary product moment) with a linear combination of the predictor variables. Unlike multiple regression, where the potencies are analyzed independently with one model for each response, thus ignoring any correlation structure amongst the different potencies. CCA utilizes this shared information and affords an analysis of all response variables simultaneously.

4.7.2 Formulation of the Problem

The variables in the response group are designated as $Y_1, Y_2, \ldots, Y_q$ and the variables in the descriptor group as $X_1, X_2, \ldots, X_p$. The principle of the method is to calculate a linear combination of the q response variables:

$$W_1 = a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1q}Y_q \tag{1}$$

and a linear combination of the p predictor or descriptor variables:

$$Z_1 = b_{11}X_1 + b_{12}X_2 + \ldots + b_{1p}X_p \tag{2}$$

where the coefficients $a_{11}, a_{12}, \ldots, a_{1q}$ and $b_{11}, b_{12}, \ldots, b_{1p}$ are estimated from the data. These coefficients are chosen, so that the pairwise correlation between W_1 and Z_1 will be as large as possible. The idea is that if this maximum correlation, R_{C1}, is significantly large, then there is evidence of an association between the two sets of variables. W_1 and Z_1, given by Eqs. (1) and (2), are referred to as canonical variates (CV), and the (maximum) correlation between them (R_{C1}) is known as the canonical correlation. These canonical variates are equivalent to the principal components (PCs) produced in principal component analysis (PCA), or the latent variables produced in PLS, with the exception that the criterion for their selection has altered. Whereas all three techniques produce linear combinations of the original variables, CCA does so not with the object of accounting for as much variance as possible within one set of variables (PCA) or maximising the covariance $X'Y$ of the data (PLS), but in order to maximize the correlation between the two sets of variables, X and Y.

The techniques of CCA, PCA and PLS are analogous in several other respects. PCA, for example, selects a first PC that accounts for a maximum amount of variance in a given set of variables, and then computes a second PC accounting for as much as possible of the variance left unaccounted for by the first PC, and so forth. PLS selects successive pairs of latent variables, one member of each pair being constructed from the X and Y sets, respectively, to have maximum covariance. CCA follows a similar procedure. The first pair of CVs, W_1 and Z_1 (Eqs. (1) and (2)), are selected so as to give the highest intercorrelation possible, given the nature of the variables involved. A second pair of CVs, (W_2, Z_2) is then selected to account for a maximum amount of the relationship between the two sets of variables left unaccounted for by the first pair of CVs, and so forth. In practice, the number of pairs of canonical variates (W_i, Z_i), extracted by the analysis, will be equal to the smaller of q and p. Thus, the linear relationships,

$$\begin{aligned} W_1 &= a_{11}Y_1 + a_{12}Y_2 + \ldots + a_{1q}Y_q \\ W_2 &= a_{21}Y_1 + a_{22}Y_2 + \ldots + a_{2q}Y_q \\ &\ldots\ldots\ldots\ldots \\ W_s &= a_{s1}Y_1 + a_{s2}Y_2 + \ldots + a_{sq}Y_q \end{aligned}$$

and

$$\begin{aligned} Z_1 &= b_{11}X_1 + b_{12}X_2 + \ldots + b_{1p}X_p \\ Z_2 &= b_{21}X_1 + b_{22}X_2 + \ldots + b_{2p}X_p \\ &\ldots\ldots\ldots\ldots \\ Z_s &= b_{s1}X_1 + b_{s2}X_2 + \ldots + b_{sp}X_p \end{aligned} \tag{3}$$

can be found, where s is the smaller of q and p. The pairs of canonical variates are derived in decreasing order of importance, so that the (canonical) correlation R_{C1} between the first pair of CVs (W_1, Z_1) is a maximum; the correlation R_{C2} betweeen the second pair (W_2, Z_2) is a maximum, subject to these variables being uncorrelated with (W_1, Z_1), and $R_{C2} < R_{C1}$; the correlation R_{C3} between W_3 and Z_3 is a maximum, subject to (W_3, Z_3) being uncorrelated with W_1, Z_1, W_2 and Z_2, and $R_{C3} < R_{C2}$, and so forth.

4.7.3 Features of Canonical Correlation Analysis (CCA)

4.7.3.1 Procedure for CCA

CCA takes, as its b asic input, two groups of variables (standardized to zero mean and unit variance), each of which can be given theoretical meaning as a group. These two groups will generally comprize n observations on a set of response variables $(Y_i, i = 1, 2, \ldots, q)$, and a set of descriptor/predictor variables $(X_i, i = 1, 2, \ldots, p)$, so that the data matrix will look like the following

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} & y_{11} & y_{12} & \cdots & y_{1q} \\ x_{21} & x_{22} & \cdots & x_{2p} & y_{21} & y_{22} & \cdots & y_{2q} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & y_{n1} & y_{n2} & \cdots & y_{nq} \end{bmatrix}$$

The method of extracting the successive pairs of CVs involves an eigenvalue-eigenvector problem, which has some similarity with that for obtaining the PCs in PCA. The eigenvalues (R_C^2) and associated eigenvectors constructed by the CCA are based on the combined $(p + q) \times (p + q)$ correlation matrix, $\boldsymbol{C}_{XY}$, between the descriptor variables and the response variables, where

$$\boldsymbol{C}_{XY} = \left[\begin{array}{ccc:ccc} p \times p & \text{matrix} & \boldsymbol{R}_{XX} & p \times q & \text{matrix} & \boldsymbol{R}_{XY} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ q \times p & \text{matrix} & \boldsymbol{R}'_{XY} & q \times q & \text{matrix} & \boldsymbol{R}_{YY} \end{array}\right].$$

From this matrix an $s \times s$ matrix $\boldsymbol{R}_{YY}^{-1}\boldsymbol{R}'_{XY}\boldsymbol{R}_{XX}^{-1}\boldsymbol{R}_{YY}$ can be constructed, and the eigenvalues of this matrix $\lambda_1 > \lambda_2 > \ldots > \lambda_s$ are the squares of the canonical correlations between the pairs of canonical variates, i.e., $R_{\mathrm{Ci}} = \sqrt{\lambda_i}$, and represents the amount of variance in the canonical variate, W_i, that is accounted for by the other canonical variate, Z_i. The corresponding eigenvectors allow the canonical variate coefficients (a_{ij} and b_{ij} in Eq. (3)) to be calculated.

4.7.3.2 Canonical Weights and Canonical Loadings

Interpretation of the canonical variates is necessary, if a picture of the association between the two variables sets is to be formulated. Canonical weights and canonical loadings have been used to assess the relationship between the original variables and the canonical variates. Canonical weights, a_{ij} and b_{ij}, in Eq. (3), are analogous to the coefficients in multiple linear regression analysis (MRA) and indicate the contribution of each variable to the variance of the respective canonical variate. They, therefore, define the location and direction of the canonical vector, and identify the variables used in its construction. However, as in MRA, the canonical weights may be highly unstable due to multicollinearity. Thus, some variable may have a small or negative weight because it is strongly correlated to some other variable(s) in the model. Canonical loadings are thought by many to be more useful in identifying the nature of the canonical relationships. Canonical loadings give the simple product moment correlation of the original variable and its respective CV, and reflects the degree to which the variable is represented by a CV. The canonical loadings can easily be found by correlating the raw variable scores with the canonical variate scores. Canonical variate scores are analogous to PC scores in PCA.

4.7.3.3 Proportion of Explained Variance

A large canonical correlation between pairs of variates does not necessarily indicate a useful and interpretable solution. For example, if only one or two variables have a high association with the canonical variable, and, thus, have high loadings, the total amount of variance in the response set of variables (Y), accounted for by the canonical variate, can be small. In such cases, there is generally no relationship between the two sets of variables, since the canonical structure indicates only a specific relationship between one or two prediction and response variables.

The proportion of explained variance in the Y set that is accounted for by a particular variate is given by the average squared loading of the response variables on that particular variate, i.e.,

$$[R_Y(j)]^2 = \sum_{i=1}^{q} [r_{Y_i}(j)]^2/q \tag{4}$$

where $[R_Y(j)]^2$ denotes the proportion of variance in the response set variables accounted for by the j-th canonical variate, and $r_{Y_i}(j)$ is the canonical loading of the i-th response variable on the j-th canonical variate. Similarly, the proportion of variance in the predictor set of variables (X) accounted for by the j-th canonical variate is given by:

$$[R_X(j)]^2 = \sum_{i=1}^{p} [r_{X_i}(j)]^2/p\,. \tag{5}$$

4.7.3.4 Redundancy Coefficient

In many QSAR studies, it is useful to know how much of the variance in the response set is accounted for by the predictor set. One might think that R_C^2 provides this information. However, atthough the squared canonical correlation coefficients do have some interpretations of the variance, they give the variance shared by the CVs and not the variance shared by the original X and Y variables.

Stewart and Love [1] have proposed an index, called the redundancy coefficient, which represents the amount of variance in the response set that is redundant to the variance in the predictor set. This redundancy coefficient, denoted by $R_{\mathrm{C}_{Y/X}}$, is given by

$$R_{\mathrm{C}_{Y/X}} = \sum_{j=1}^{s} \lambda_j [R_Y(j)]^2 \tag{6}$$

and is the sum of the product of the proportion of explained variance in the Y set that is accounted for by a particular CV with its associated eigenvalue $(=(R_{\mathrm{C}_j})^2)$. Stewart and Love showed that $R_{\mathrm{C}_{Y/X}}$ is equivalent to regressing each Y variable, in turn, on all the X variables, and then averaging the q resulting square multiple correlation coefficients. Thus, $R_{\mathrm{C}_{Y/X}}$ expresses the proportion of variance in the Y set, which is explained by the X set. A redundancy coefficient, $R_{\mathrm{C}_{X/Y}}$, can also be constructed and represents the amount of variance in the X set of variables that is redundant to the variance in the Y set.

4.7.3.5 Hypothesis Testing

Bartlett [2] has outlined a procedure for testing the statistical significance of the canonical correlations, when the sample size is large, and so determined how many significant relationships exist between the two sets of variables.

To test if there is at least one significant canonical correlation, the following test statistic is calculated,

$$\chi_0 = -\left[n - 1 - \frac{(p + q + 1)}{2}\right] \sum_{i=1}^{s} \ln(1 - \lambda_i) \tag{7}$$

where n is the number of observations, for which data are available. The distribution of χ_0 is approximately *chi*-squared, so that if χ_0 is greater than a selected percentage point of the *chi*-squared distribution with pq degrees of freedom, then it may be concluded that at least one of the canonical correlation coefficients is significantly large. However, if χ_0 is not significantly large, then there is no evidence of any relationship between the X and Y variables.

If χ_0 is significantly large, then the following test statistic is calculated:

$$\chi_1 = -\left[n - 1 - \frac{(p + q + 1)}{2}\right] \sum_{i=2}^{s} \ln(1 - \lambda_i) \tag{8}$$

χ_1 has the effect of the first canonical correlation removed, and is approximately *chi*-squared distributed with $(p - 1)(q - 1)$ degrees of freedom. This process continues, until it is found that the remaining correlations are no longer significant. The test statistic to examine the significance of all, but the first j canonical coefficients is:

$$\chi_j = -\left[n - 1 - \frac{(p + q + 1)}{2}\right] \sum_{i=j+1}^{s} \ln(1 - \lambda_i) \tag{9}$$

and has $(p - j)(q - j)$ degrees of freedom.

4.7.4 The Application of CCA to QSAR Problems

Compared with other multivariate procedures, CCA has been used in relatively few QSAR studies. A search of the Science Citation Database, for example, identified only one publication, describing the use of the technique to investigate the structure/activity relationships of drugs. Nevertheless, a number of studies have been undertaken in this laboratory, some of which will be described to indicate the potential of the method as a tool for investigating structure/activity relationships. There are no examples of the use of CCA, which establish its value for predicting the biological activity of drugs or agrochemicals, although at least one preliminary investigation has been undertaken [3].

4.7.4.1 Pharmacokinetics of Pyrethroid Insecticides in Insects

A mathematical model describing insect pharmacokinetics [4] was used to generate data, describing the penetration and elimination of a series of pyrethroid analogues, the (±)-*cis/trans*-methylbenzyl-chrysanthemates, applied to mustard beetles. The three parameters of the model, k_p (penetration rate constant), k_e (elimination rate constant) and λ (a parameter describing the relative affinities of the compound for the outside and inside of the organism) were estimated for each compound using pharmacokinetic profiles (mass of insecticide vs time), based on experimental results obtained by rinsing applied insecticide from the surface of treated insects, and grinding and extracting the washed insects with a suitable solvent a various times after treatment [5].

The physico-chemical properties of the compounds were described by four variables, viz. π (partition term), σ (Hammett constant), S_2 (a steric effect) and log *cis* (the proportion of *cis*-isomer in the mixed esters). The influence of chemical structure on pharmacokinetics was investigated using canonical correlation analysis with log k_e, log k_p and log λ as the first set of variables, and π, σ, S_2 and log *cis* as the second set. The correlation matrix for this analysis is presented in Table 1, and is based on data for 14 of the 22 possible compounds, plus the parent

Table 1. Interparameter correlations, describing the molecular properties of the methylbenzyl-(±)-*cis/trans*-chrysanthemates and their pharmacokinetic behavior, following topical application to mustard beetles, *Phaedon cochleariae* Fab.

	log k_e	log k_p	log λ	π	σ	S_2	log *cis*
log k_e	1.000						
log k_p	0.359	1.000					
log λ	−0.090	−0.590	1.000				
π	−0.567	0.088	−0.598	1.000			
σ	0.565	−0.040	0.475	−0.948	1.000		
S_2	−0.291	−0.078	−0.451	0.356	−0.112	1.000	
log *cis*	−0.277	−0.379	0.123	0.134	−0.122	0.185	1.000

Table 2. The standardized coefficients and correlations (loadings) of the pharmacokinetic and molecular properties of the methylbenzyl-(±)-*cis*/*trans*-chrysanthemates with the canonical variates

Variable	Standardized coefficients Canonical Variate			Canonical loadings Canonical Variate		
Set 1	*CNVRF*1	*CNVRF*2	*CNVRF*3	*CNVRF*1	*CNVRF*2	*CNVR*3
$\log k_e$	0.597	0.028	−0.906	0.634	0.378	−0.674
$\log k_p$	0.336	0.970	0.861	0.005	0.999	0.034
$\log \lambda$	0.923	−0.0333	0.850	0.671	−0.609	0.423
Set 2	*CNVRS*1	*CNVRS*1	*CNVRS*1	*CNVRS*1	*CNVRS*2	*CNVRS*3
π	−0.121	3.287	−0.160	−0.925	0.179	0.323
σ	0.643	2.859	−0.783	0.819	−0.079	−0.565
S_2	−0.547	−0.858	−0.800	−0.662	−0.137	−0.734
log *cis*	0.004	−0.687	0.189	−0.193	−0.755	0.114
Can. Corr.	0.930	0.503	0.254			
Eigenvalue	0.866	0.253	0.065			

compound benzyl-(±)-*cis*/*trans*-chrysanthemate. The results of the CCA, which was performed using the BMDP program 6M, are summarized in Table 2.

Applying Bartlett's test (Eq. (7)) we have:

$$\chi_0 = -[15 - 1 - (4 + 3 + 1)/2]\,[\ln(1 - 0.866) + \ln(1 - 0.253) + \ln(1 - 0.065)] = 23.65 \qquad (7)$$

which, with 4 × 3 = 12 degrees of freedom, is significant at the 5% level. Consequently, there is evidence to suggest that there is a relationship between the two variable sets. We can now test whether a significant relationship exists between the two sets of variables after the effect of the first canonical variate pair have been removed. Using Eq. (8) we find that $\chi_1 = 3.59$, which, with $(4 - 1)(3 - 1) = 6$ degrees of freedom, is not significant. Similarly, $\chi_2 = 0.67$ (with 2 degrees of freedom). We can conclude, therefore, that only one canonical correlation is significant.

From Table 2 it can be seen that $\log k_e$ and $\log \lambda$ from the first set, and π, σ and S_2 from the second set, load fairly strongly onto the first CV. The signs of the coefficients indicate that an increase in the value of this variate is associated with a) a decrease in π, b) an increase in σ, and c) a decrease in S_2. A plot of the canonical scores for W_1(*CNVRF*1) and Z_1(*CNVRS*1) are given in Fig. 1, where the pattern of points reflects the canonical correlation of 0.930(R_{C1}).

An increase in the size of k_e and λ would result in a faster net flow of material through the two compartments of the pharmacokinetic model. Thus, an increase in the magnitude of the first CV describes a reduced residence time within the insect, which would then be exposed to a high level of internal toxicant (large λ) for a reduced

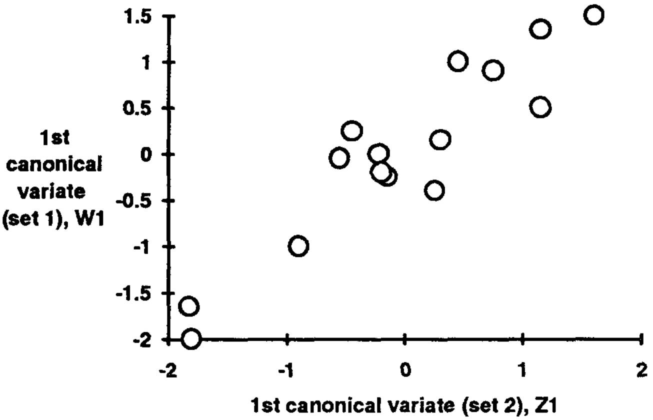


Figure 1. A plot of the canonical scores for $W_1(CNVRF1)$ and $Z_1(CNVRS1)$, extracted from data describing the relationship between the pharmacokinetic and the molecular properties of the methylbenzyl-(±)-*cis*/*trans*-chrysanthemates (R_{C1} = 0.93).

elapsed time (high rate of elimination, k_e). Although the other two pairs of CVs are not significant, careful interpretation of Bartlett's χ_0 statistic is necessary if Type II errors (acceptance of false null hypotheses) are to be avoided. Thus, it is interesting to note that the second CV associates log k_p (with a loading of 0.999) almost entirely with log *cis* (loading = 0.755). This result suggests that the second CV may represent the conductivity of the insect cuticle to movement of insecticide across this barrier, and that the rate cuticular penetration increases in proportion to the relative amount of *trans*-isomer applied.

The redundancy calculations using Eq. (6) show that despite the high canonical correlation linking the two variable sets, only 38.7% of the variance in the pharmacokinetic parameters is accounted for by the physico-chemical variables.

4.7.4.2 The Relationship between the Physico-Chemical Properties of Pyrethroids and Pharmacokinetics, Pharmacodynamics and Toxicity

In a subsequent study from the same laboratory, CCA was used to investigate the structure/activity relationships of a second series of substituted benzyl (±)-*cis*/*trans*-cyclopropane-1-carboxylates, the QSAR compounds [3]. These compounds were selected using procedures, designed to produce a training series with good distributional properties and low interparameter associations, and were, therefore, well suited to a QSAR study. Because pyrethroids have a rapid knockdown (K_d) effect in insects followed much later by lethality (L), two potencies ($1/KD_{50}$ and $1/LD_{50}$) were determined for each compound in the series. These are respectively the estimates of the inverse of the dose required to knockdown and kill 50% of the treated insects.

The association between the set of variables describing knockdown and insecticidal activities (log KD_{50} and log LD_{50}: Set 1) and the physico-chemical properties of the

QSAR pyrethroids were investigated. The physico-chemical properties of the pyrethroids were the following: five Verloop steric parameters, viz. $2L$, $3B1$, $4B1$, $4B4$, $5B1$, where the numeral prefix indicates the position of a substituent on the benzyl ring; Livingstone's $\varkappa$ parameter, derived from NMR studies and describing charge-transfer interactions [3]; Hammett's σ constant; a dummy variable, S_4, describing the sterification of the substituted benzyl alcohol with an acid moiety, $S_4 = 0$ for chrysanthemates or 1 for chlorsanthemates; and CH_2, the NMR chemical shift (ppm) of the benzylic methylene protons (Set 2). Two significant pairs of canonical variates, *CNVRF1/CNVRS1* and *CNVRF2/ CNVRS2*, were identified (Bartlett's test, $p < 0.00001$), and used [3] to identify associations between the response variables (Set 1) and the descriptor variables (Set 2).

The first pair, *CNVRF1/CNVRS1*, is based on the sum of the weighted log LD_{50} and log KD_{50} estimates, whereas the second pair, *CNVRF2/CNVRS2*, is based on a difference between these estimates, i.e. a contrast between the two sets of potencies.

The canonical variates *CNVRF1/CNVRS1* were correlated with experimental estimates of the threshold concentration (log *MTC*) of pyrethroid required to elicit abnormal activity in an isolated nerve preparation, the crayfish stretch receptor, and is, therefore, related to the pharmacodynamic activity of the pyrethroids (Eqs. (10) and (11)).

$$\begin{array}{lccc r}
 & n & r & F & \\
CNVRF1 = 5.38 + 0.85 \log MTC & 16 & 0.80 & 25.4 & (10) \\
CNVRS1 = 6.59 + 1.03 \log MTC & 6 & 0.91 & 20.4 & (11)
\end{array}$$

The second pair of variates (*CNVRF2/CNVRS2*) were correlated with the pharmacokinetic parameters, k_e, k_p and λ, estimated for these compounds following topical application (4 µg/insect) to mustard beetles, *Phaedon cochleariae*.

$$\begin{array}{lccc r}
 & n & r & F & \\
CNVRF2 = 2.20 - 2.74k_e - 0.05k_p - 0.70\lambda & 16 & 0.91 & 20.4 & (12) \\
CNVRS2 = 1.77 - 2.31k_e - 0.05k_p - 0.58\lambda & 6 & 0.95 & 5.6 & (13)
\end{array}$$

This interpretation of the two pairs of variates enabled Szydlo [3] to identify relationships between the physico-chemical properties of the QSAR compounds and their pharmacokinetic behavior and pharmacodynamic activity, PA (Table 3).

Table 3. Examples of the relationships between the physico-chemical properties of pyrethroid insecticides and their pharmacokinetic and pharmacodynamic behavior identified by CCA [3]

Property	Pharmacokinetics	Pharmacodynamics
$4B1$	An increase in $4B1$ reduces flow through the insect	Large dimensions correlate positively with pharmacodynamic activity (PA)
κ	Electron donation reduces reduced k_e, k_p or λ	Relatively unimportant
S_4	Dichlorovinyl substitution results in a reduced k_e, k_p or λ	Dichlorovinyl substitution enhances pharmacodynamic activity (PA)

4.7.4.3 Variable Deletion Procedures and CCA

In order to obtain a parsimonious QSAR model, based on CCA, Szydlo [7] developed a procedure for deleting descriptors in a stepwise manner by using an approach commonly employed in multiple regression. His search for a reduced set of descriptors was undertaken using an initial set of 22 physico-chemical variables. This set was characterized by substantial multicollinearity, but could be reduced to eleven variables without loss of information using the coefficient of determination (R^2) as a diagnostic to identify those variables which could be deleted from the set because they contained no unique information (i.e. R^2 is equal to unity) and are, therefore, redundant. Each descriptor was regressed on all other descriptors in the set (Set 2), and those with R^2 values equal to unity were eliminated. The remaining 11 variables, therefore, contained some unique information which could be related to insecticidal and knockdown activities.

The association of this reduced set, with the two toxicity parameters log KD_{50} and log LD_{50}, was investigated using CCA. An initial canonical correlation resulted in two toxicological canonical variates ($TF1$ and $TF2$), which summarized the relationship between the biological and molecular properties of the training series. However, severe multicollinearity still characterized the physico-chemical variable set, indicating the presence of redundant variables. In order to reduce the complexity of the predictor set, whilst retaining its association with the two toxicological potencies, backward stepping was employed as follows. The variable, whose removal resulted in the smallest reduction in the CCA eigenvalue, R_C^2, was discarded and the process repeated in a second step using the remaining variables. Stepwise deletion was stopped, when the reduction in R_C^2 was considered to indicate a significant loss of information on the basis of an F-statistic. The F-test employed compares the change in variance explained, after removing a variable, with the residual variance before deletion of a variable,

$$F = \frac{(R_{\mathrm{C}k_1}^2 - R_{\mathrm{C}k_2}^2)}{(k_1 - k_2)} \bigg/ \frac{(1 - R_{\mathrm{C}k_1}^2)}{(n - k_1 - 1)} \tag{14}$$

Table 4. Summary of the backward stepping procedure to identify the most parasimonious canonical variates, describing the relationship between the toxicity and physico-chemical properties of the QSAR pyrethroids [7]

Number of variables	Variable removed	Eigenvalue of the 1st canonical variate	F	Eigenvalue of the 2nd canonical variate	F
11		0.969		0.904	
10	F	0.969		0.899	
9	$5L$	0.964		0.897	
8	R	0.945		0.886	3.20
7	$3B1$	0.941	1.73	0.814	19.58
6	log MW	0.934	3.25	0.733	
5	$4L$	0.896	19.80	0.687	

Table 5. Squared multiple correlations of each variable in the predictor set with all other variables in that set, before and after backward stepping [7]

Variable	Original 11 variables R^2	"Best" 6 variables R^2
$3B1$	0.996	
$3B4$	0.992	0.383
$4L$	0.993	0.709
$4B1$	0.993	0.733
$5L$	0.997	
$6L$	0.996	0.191
κ	0.995	0.270
F	0.995	
R	0.998	
S_4	0.991	0.090

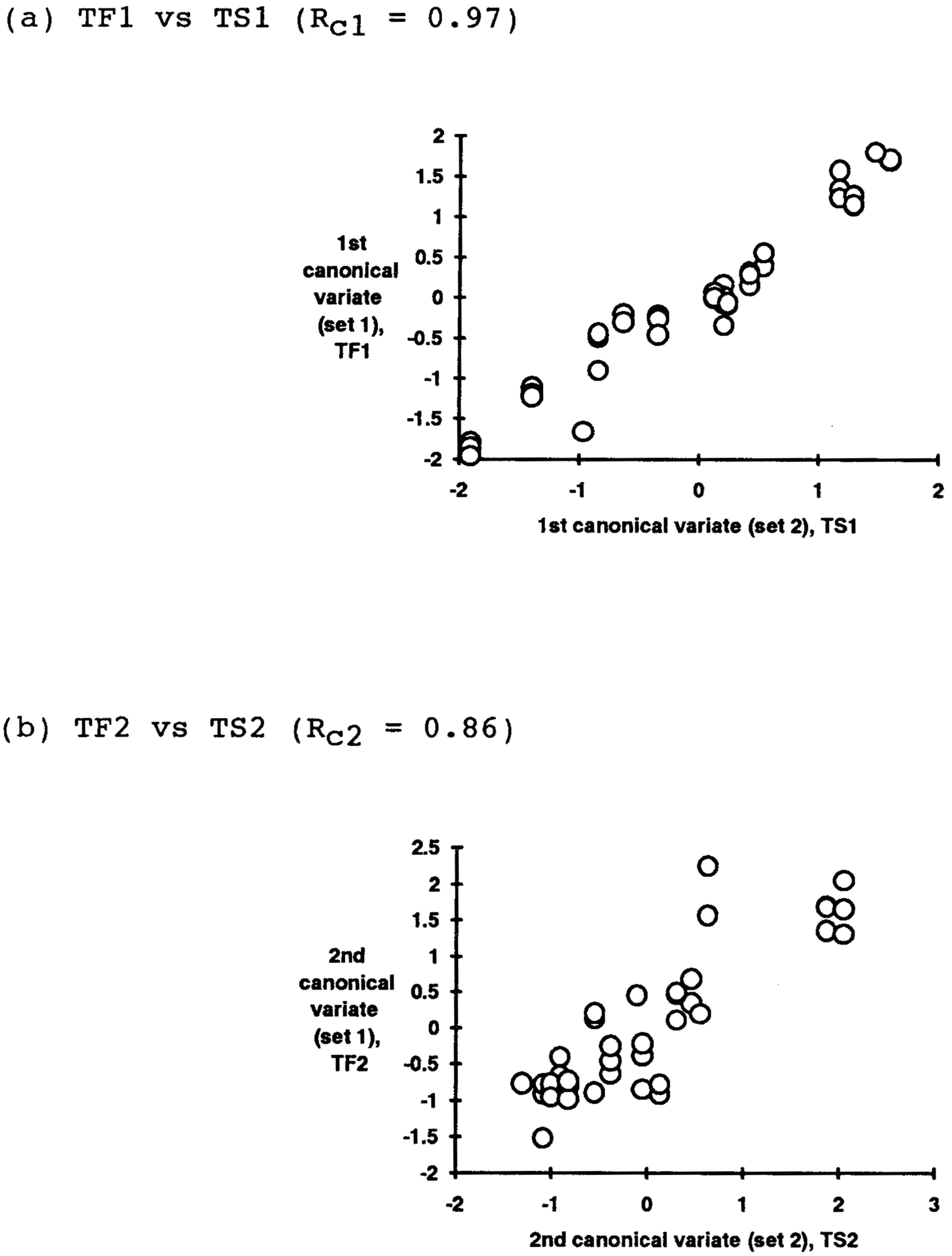
where k_1 is the number of variables at the present step, and k_2 is the number after the next step when one further variable has been deleted. Thus, this F-statistic measures whether deletion of a particular variable produces a significant change in the variation shared between the response and descriptor sets.

This procedure was carried out for each of the canonical variates to yield the "best" equation based on the following six descriptors viz. $3B4$, 4L, $4B1$, $6L$, κ and S_4. A summary of this procedure is presented in Table 4. The squared multiple correlations of each variable in the second set, with all other variables in that set computed before and after backward stepping, reflecting both the degeneracy and multicollinearity of this data, is described in Table 5. Table 6 presents loadings and

Table 6. Standardized coefficients and correlations (loadings) of the original variables with the canonical variates estimated for the QSAR pyrethroids, topically applied to mustard beetles, *Phaedon cochleariae* Fab. [7]

Variable	Loadings		Standardized coefficients	
	$TF1$	$TF2$	$TF1$	$TF2$
$\log LD_{50}$	**0.998**	−0.055	0.949	−0.960
$\log KD_{50}$	**0.711**	**0.703**	0.074	1.348
	$TS1$	$TS2$	$TS1$	$TS2$
$3B4$	**−0.347**	0.117	−0.744	−0.045
$4L$	−0.134	**0.389**	0.414	−0.416
$4B1$	**−0.332**	**0.492**	−0.759	1.086
$6L$	**−0.318**	−0.062	−0.331	−0.355
κ	0.263	**0.704**	0.517	0.862
S_4	**−0.582**	0.077	−0.552	0.087

Entries in bold indicate significant loading ($p < 0.05$)



(a) TF1 vs TS1 (R_{C1} = 0.97)

(b) TF2 vs TS2 (R_{C2} = 0.86)

Figure 2. A plot of the canonical scores for *TF*1, *TF*2, *TS*1 and *TS*2 extracted from QSAR data, describing the insecticidal activity of the aromatic substituted benzyl-(±)-*cis*/*trans*-cyclopropane carboxylates.

standardized coefficients for the "best" equations for $T1$ and $T2$, and Fig. 2 shows the observed canonical associations.

Bartlett's test for the significance of the canonical variates obtained for the six predictor variables gave the following results:

1st canonical variate ($T1$):	eigenvalue = 0.934
Chi-squared = 139.48 (12 d.o.f.),	tail probability < 0.001
2nd canonical variate ($T2$):	eigenvalue = 0.733
Chi-squared = 45.52 (12 d.o.f.),	tail probability < 0.001

Thus, two significant variates are obtained using a reduced set of six physico-chemical properties. It should be noted that different combinations of the response and descriptor variables load (Table 6) onto the two variates, $T1$ and $T2$; some variables, e.g. log KD_{50} and $4B1$, are associated with both variates, while others, e.g. log LD_{50} and $6L$, are associated with only one variate. Elimination of redundant variables, including many of the Verloop parameters, wich for the QSAR pyrethroids are highly correlated has reduced, though not eliminated, the problems of degeneracy and multicollinearity (Table 4) and identified a parsimonious set of variates to describe the structure/activity relationships of these insecticides.

Because the above procedure can be applied to each of the canonical variates under review, decisions on variable deletion will need to take account of the influence of a variable on the set of significant variates. It may well be that a variable which makes no real contribution to the first variate, has a major influence on the second or third. Thus, sets of canonical variates have to be considered if type II errors are to be avoided. It is also advizable to keep a check on the redundancy, within and between the data sets, during the deletion procedure. The coefficient of variation for each variable regressed on all other variables within a set (Sec. 4.7.4.3), and the redundancy coefficient of Stewart and Love [1] (Sec. 4.7.3.4) can provide useful diagnostics for this task.

4.7.4.4 Mapping the Toxicological and Physico-Chemical Hyperspaces

Szydlo [7] obtained bivariate maps of the toxicological and physico-chemical hyperspaces of the QSAR pyrethroids by plotting $TF1$ against $TF2$, and $TS1$ against $TS2$, respectively. There are more points in the toxicological space compared with the physico-chemical space, reflecting the replication of the toxicity estimates, usually 3 per compound. Because the set of toxicological estimates, KD_{50} and LD_{50}, is of low dimensionality and is common to both analyzes, the maps based on the canonical variates $TF1$, $TF2$, $TS1$ and $TS2$ (Fig. 3) [3] are very similar to those ($CNVRF1$, $CNVRF2$, $CNVRS1$ and $CNVRS2$) reported by Szydlo et al. [3]. They can be used to identify the contributions of pharmacodynamics ($TF1/TS1$; $CNVRF1/CNVRS1$) and pharmacokinetics ($TF2/TS2$; $CNVRF2/CNVRS2$) to the overall knockdown and killing potencies of each compound in the series [3, 6]. Eqs. (10) and (11), and the associated CCA loadings (Table 6), for example, suggested that low values of $TF1/TS1$ corresponded to high pharmacodynamic ($-\log MTC$), knockdown

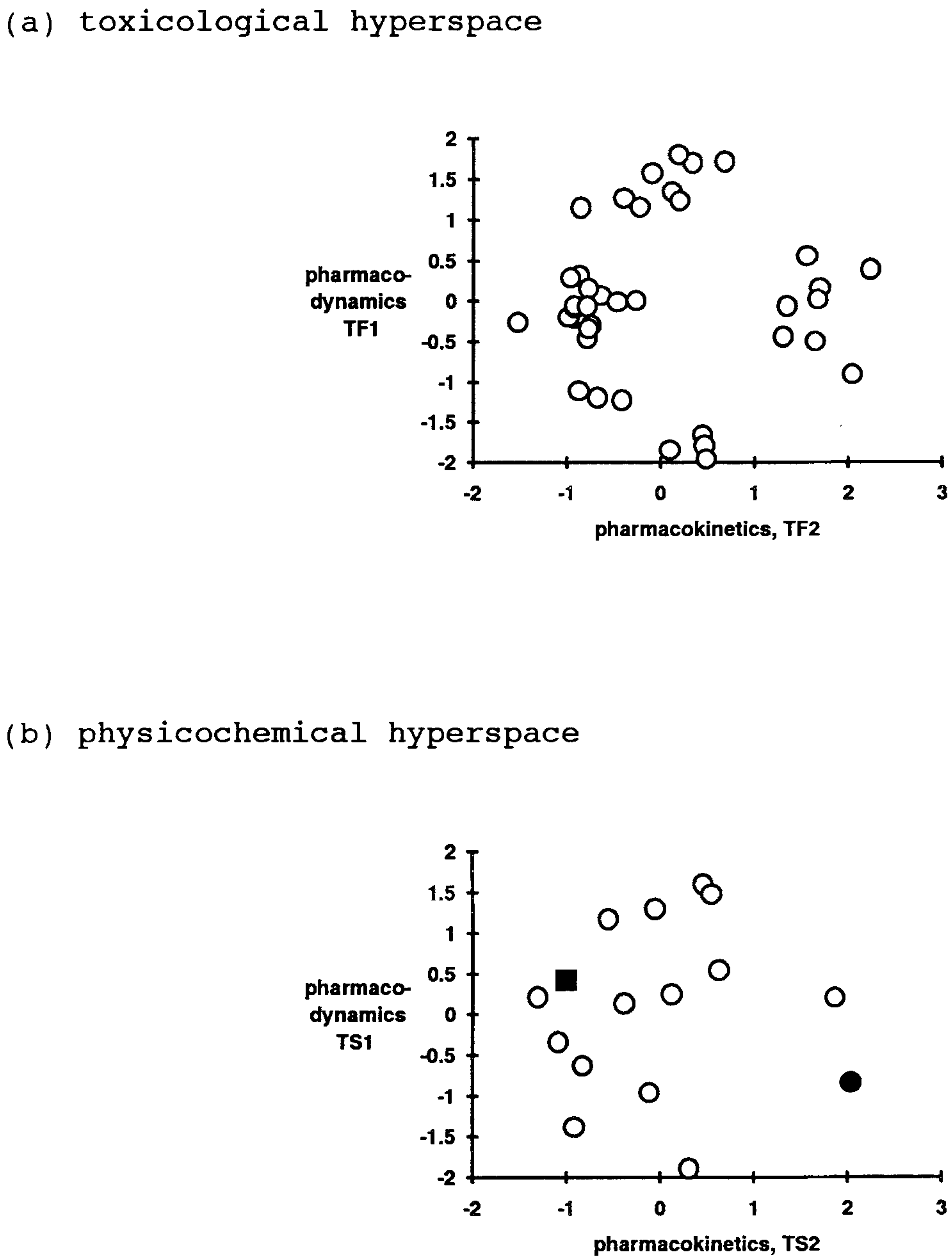


Figure 3. Bivariate maps of the toxicological and physico-chemical hyperspaces, showing the relationships between pharmacodynamics and pharmacokinetics, and a) the toxicity and b) the molecular properties of the aromatic substituted (±)-*cis*/*trans*-benzylcyclopropane carboxylates.

($\log(1/KD_{50}) = -\log KD_{50}$)) and killing ($\log 1/LD_{50} = -\log LD_{50}$) potencies. In contrast to this, high values for $TF2/TS2$ corresponded to small k_e, k_p and λ values (Eqs. (12) and (13)), i.e. slow flow through the insect, and consequently decreased knockdown activity ($-\log KD_{50}$), but increased killing activity ($-\log LD_{50}$) [3, 6]. The maps (Fig. 3) also suggested [3] that the QSAR compound **4** (●), with a low value for $TF1/TS1$ and a high value for $TF2/TS2$, has a relatively good killing activity, but poor knockdown activity as a consequence of its slow penetration and elimination by mustard beetles. QSAR compound **13** (■), which has a low value of $TF2/TS2$, penetrates rapidly, but is quickly eliminated from the insect body. As a result, it is a relatively effective knockdown compound, but only has average insecticidal activity ($TS1 = 0.42$). Thus, although the overall potency of these pyrethroid insecticides is related to $TF1/TS1$, and therefore, $CNVRF1/CNVRS1$ and, hence, to neurotoxicity, the balance between knockdown and killing activity depends on the rate of movement into, elimination from, and distribution within the insect [3, 6].

The bivariate spaces obtained by plotting each variate against the response variables, $\log KD_{50}$ and $\log LD_{50}$, can identify further relationships between molecular properties, pharmacokinetics and toxicity [6].

4.7.5 Useful Features of CCA for the Design of Biologically Active Compounds

Canonical correlation can be regarded as a generalized regression procedure. Unlike linear and multiple regression, it is not limited to problems concerning only univariate y; CCA can consider both multivariate X and Y blocks of data. CCA is susceptible to many of the problems, e.g. assumptions of independance of the X variables, homoschedasticity and normality of the variance of Y, which are associated with the simpler regression procedures, based on univariate y. Furthermore, because two sets of associated variables are under investigation, the strength of these associations, which can appear large when comparisons are made between canonical variates (e.g. $CNVRFi$ with $CNVRSi$), may appear to be quite small when the loadings of the original variables are considered (see Sec. 4.7.3.2). This can be a problem, both for interpretation and prediction. Nevertheless, the technique has advantages and may be useful for investigating the relationship between biological activity and molecular structure.

One obvious benefit is that several biological responses can be considered for the same series of compounds. This is important because it changes the unfortunate emphasis placed in most QSAR studies on large numbers of physico-chemical properties, relative to the biological data. Because of the significantly larger variances in biological data compared with chemical data, application of multiple regression can, therefore, lead to QSAR models which are overdefined and non-general. Such models are of no value for identifying novel compounds of high activity. Techniques such as CCA, which take account of multivariate y, can overcome these limitations by, 1) weighting the Y block variables (the biological activities) relative to the X

block (the molecular properties), and 2) constraining the possible models to constructs, which reflect the associations not only between the X and Y variables, but also those within both the X and Y sets. This latter feature represents a constraint which may reduce the chance of a spurious correlation.

4.7.6 Predicting Biological Activity

There are several ways, in which the results of a CCA can be used to obtain a prediction for the design of a new flavor, drug or agrohemical. One strategy is to regress each of the original biological potencies on the appropriate set of canonical variates which have been constructed. Because each variate is a new, orthogonal variable with a known functional relationship to the original variables, the procedure is straightforward. Each original y variable is regressed on the set of canonical variates (i.e. *CNVRSs*), constructed from the molecular desciptors (the X block), using a standard multiple regression procedure. The assumptions of this procedure are now satisfied and a reliable prediction model should result, so long as other considerations, such as sampling and design criteria have been satisfied. Only those canonical variates which have a significant t-value should be retained in any model, however. Non-significant canonical variates can be removed, without having to re-estimate the regression because the β-coefficients are stable and have been estimated using an orthogonal set of X variables.

A second approach is to base prediction on a method analogous to that used to solve simultaneous equations. The various canonical variates are regarded as i equations in i unknowns, which can be solved analytically; i is the number of variables in the smallest set. This approach, developed at the University of Portsmouth, has considerable potential, particularly if the associated canonical correlation coefficients (R_Cs) are high. Its validity as a drug design strategy, however, requires further evaluation.

4.7.7 The Advantages and Disadvantages of Using CCA in QSAR Studies

Canonical correlation is a procedure for analyzing data, which comprizes more than one set of variables. This is useful in QSAR studies, where biological and chemical properties have both been measured using more than one variable. The advantages of multivariate X and Y blocks in QSAR include a better balanced analysis, which makes use of more structure/activity information, e.g. the within and between set covariances, and a lower chance of spurious correlation, since there are more constraints to model specification. Furthermore, predictions can be made using two strategies, one based on regression and the other on an analytical approach. However, failure to satisfy assumptions about the data will have similar disadvantages to multiple regression under the same conditions. These include multicollinearity, and further study is required to assess the sensitivity of CCA to this effect.

Many of the problems which are likely to be encountered, when using CCA in drug design, can be avoided by adopting the following procedures:

(1) Use the redundancy coefficient of Stewart and Love [1] to check that a substantial amount of variance is shared between the X and Y variable sets.

(2) Ensure that a high correlation does not exist between one of the Y variables and one of the X variables, since this could result in a spuriously high canonical correlation.

(3) Exclude non-significant descriptor/predictor variables by using a backward stepping procedure. This may also reduce the influence of multicollinear variables.

(4) Compare the weights and loadings to identify any discrepancies such as reversal of sign or differences in rank order.

Adoption of these guidelines should facilitate the use of CCA to provide a concise and reliable summary of any multiple associations between two sets of properties, which are observed during a QSAR study.

References

[1] Stewart, D. K., and Love, W. A., *Psychol. Bull.* **70**, 160–163 (1968)

[2] Bartlett, M. S., *Ann. Eugen.* **16**, 199–214 (1951)

[3] Szydlo, R. M., Ford, M. G., Greenwood, R. G., Salt, D. W., The Relationship between the Physicochemical Properties of Substituted Benzyl Cyclopropane-1-carboxylate esters and their Pharmacokinetic, Pharmacodynamic and Toxicological Parameters. In: *Quantitative Approaches to Drug Design*, Dearden, J. C., ed., Elsevier, Amsterdam (1983) p. 203–214

[4] Ford, M. G., Greenwood, R., and Thomas, P. J., *Pestic. Sci.* **12**, 175–198 (1981)

[5] Ford, M. G., Greenwood, R. and Thomas, P. J., *Pestic. Sci.* **12**, 265–284 (1981)

[6] Szydlo, R., M., Ford, M. G., Greenwood, R. G., Salt, D. W., The Use of Multivariate Techniques for the Prediction of Biological Activity. In: *QSAR in Design of Bioactive Compounds*, (Proceedings of the First International Telesymposium on Medicinal Chemistry.) Kuchar, M., ed., J. R. Prous Scientific, Barcelona (1984) p. 301 –320

[7] Szydlo, R. M., *An Investigation of the Pharmacokinetics of Pyrethroid Insecticides in the Adult Mustard Beetle Phaedon cochleariae*, Ph.D. Thesis, CNAA, Portsmouth 1987

4.8 Discriminant Analysis for Activity Prediction

Han van de Waterbeemd

Abbreviations

ALS	Adaptive least squares
ANN	Artificial neural networks
CSA	Cluster significance analysis
KNN	*k*-nearest neighbor
LDA	Linear discriminant analysis
MAO	Monoamine Oxidase
MLR	Multiple linear regression
PCA	Principal component analysis
QSAR	Quantitative structure-activity relationships
SCD	Single class discrimination
SIMCA	Soft independent modeling of class analogy
SPC	Structure-property correlations

Symbols

IC_{50}	Binding affinity (50% inhibition)
ED_{50}	Effective dose (50% effective concentration)
D_{12}	Linear discriminant function
X_1	Chemical or biological data used in classification
$\bar{\chi}_{1k}$	Mean of descriptor k in class 1
$(\sigma_{1k})^2$	Variance of class 1
$S[-3 < EP < +3]$	Measure for the hydropholicity
$\mathscr{F}$	Swain-Lupton field parameter
$(x'x)^{-1}$	Variance-covariance matrix

4.8.1 Theoretical Background

The quality and type of biological data are important factors for selecting the appropriate statistical method to develop quantitative structure-activity relationships. Quite a number of biological tests produce discrete results, e.g. active or inactive, or ++, +, 0, −, −−. Such data are more difficult to deal with than continuous values, such as IC_{50} or ED_{50} values. The problem in this case is how to correlate qualitative biological data to quantitative chemical data [1]. Techniques to handle such cases include adaptive least squares ALS [37], cluster significane analysis CSA (see Chap. 4.9) and linear discriminant analysis LDA. The function of discriminant analysis is to find a linear combination of factors (descriptors) that will best discriminate between two or more groups. Principal component analysis and cluster analysis deal with finding groups among objects, such as chemical compounds, while discriminant analysis deals with objects which are known to belong to different groups. In principle, the number of groups that can be considered can be any number. The maximum number of possible classes equals the total number of compounds. However, in molecular design problems, a rough separation into two groups, active vs inactive, is often considered. Therefore, the problem discussed here is the following. Given a data table with various chemical and/or biological data, X_i, two groups of compounds should be formed. Group classification functions, D_1 and D_2, have two be calculated, such that for compounds 1 to n the discriminant score $D_1 > D_2$, and for the rest of the compounds $D_1 < D_2$.

$$D_1 = a_1X_1 + a_2X_2 + a_3X_3 + \dots \qquad (1)$$

$$D_2 = b_1X_1 + b_2X_2 + b_3X_3 + \dots \qquad (2)$$

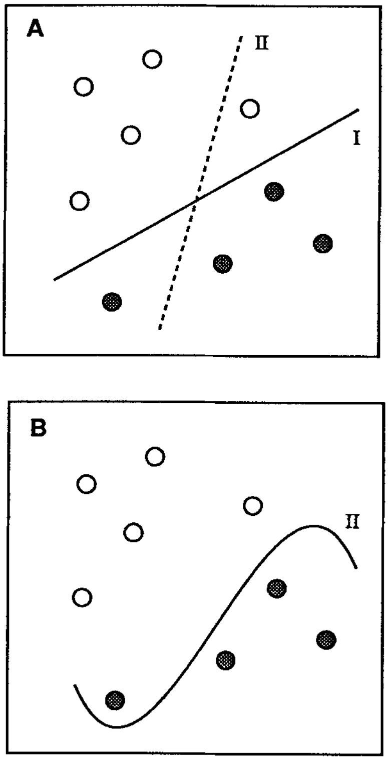


Figure 1. Discriminant functions representing a line or plane (A) or a hypersurface (B). Function I gives correct classifications, while function II has misclassifications.

The coefficients, a_i and b_j, are called discriminant weights and are obtained by a multiple regression procedure. These functions describe a line, plane or, in general, a surface (hyperplane) between the groups (see Fig. 1). The difference between the two group classification functions is called the linear discriminant function, D_{12}.

$$D_{12} = D_1 - D_2 \tag{3}$$

This function defines a hyperplane, separating the two groups of compounds. When the condition for the equality of covariance matrix for the multivariate normal distribution between two observation groups is not fulfilled, a modified procedure may be used [2]. The basic assumptions for linear discriminant analysis (LDA) also include a normal distribution of the descriptor populations and equal covariance matrices [5] for the classes. In principle, also non-linear discriminants, e.g. quadratic, can be used. But these are more complicated to deal with.

The hyperplane represented by a LDA is not unique and may be quite different from a plane calculated by multiple linear regression [31].

LDA works well if the groups of active and inactive compounds are well separated in space. In the case of embedded or asymmetric data (see Fig. 2), other strategies should be preferred, such as SIMCA (see Chap. 4.3) and single class discrimination SCD [37].

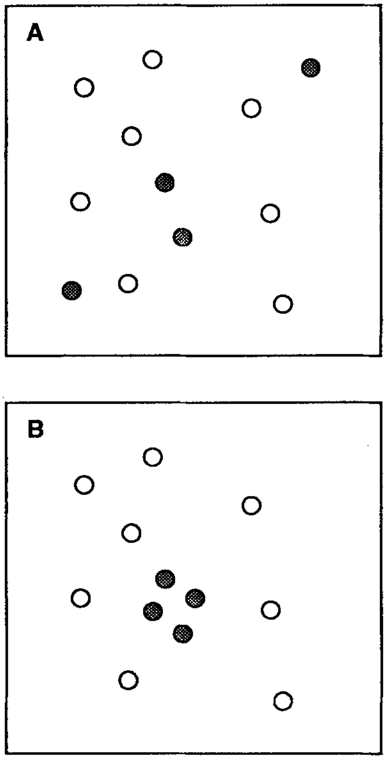


Figure 2. Ovelapping classes (A) and embedded or asymmetric data (B).

Discriminant analysis can also be performed by artifical neural networks, [3, 4, 37] or as PLS discriminant analysis [35]. A further development of linear discriminant analysis is the adaptive least squares (ALS) method [37], which allows the separation of several activity classes by a single discriminant function. LDA is further related to techniques such as Bayesian discriminants [5], non-parametric linear learning machines (LLM) [5] and k-nearest neighbor (kNN) classification. The latter strategy is rather simple, in that one looks for similar compounds in a multidimensional space [37].

4.8.2 Descriptor Selection

To answer the question as to which properties are best suited to separate molecules into classes, we may refer to Sec. 2 of this volume and to Vol. 1 of the present series [6]. No rules can be given here. As in ordinary Hansch analysis, any property which seems to be relevant to the problem may be analyzed. Physico-chemical properties, such as log P, are believed to be more useful in discriminant studies than structural descriptors [7]. A large collection of potential descriptors has been reported [8]. In order to give all descriptors the same weight, autoscaling is often performed before the analysis. In this step, the descriptor is normalized by substracting the mean value of the descriptor and dividing by its standard deviation. Thus, each descriptor has a mean value of zero and a standard deviation of one.

Using feature selection techniques, i.e. elimination of non-significant descriptors, the final discriminant analysis may be more successful. Cluster analysis or principal component analysis are often used for descriptor selection. However, some interesting alternatives may also be attempted. Jurs and coworkers [5] have used a variance method, while Takahashi et al. [9] have used the Fisher ratio. The Fisher ratio is a quantitative estimate of the significance of a given parameter for separating two classes. The Fisher ratio of the descriptor, k, (F_k) is calculated from,

$$F_k = \frac{(\bar{\chi}_{1k} - \bar{\chi}_{2k})^2}{(\sigma_{1k})^2 + (\sigma_{2k})^2} \tag{4}$$

where $\bar{\chi}_{1k}$ and $\bar{\chi}_{2k}$ are the mean values of descriptor k in classes 1 and 2, respectively, and σ_{1k} and σ_{2k} are the standard deviations of those classes.

4.8.3 Chance Correlations with Discriminant Analysis

When data sets having many variables are analyzed, there is the danger of finding chance correlations by fortuitous combinations of variables [10]. Stouch and Jurs [11 – 13] have examined the risk of change correlations in linear discriminant analysis. They concluded that the number of examined variables should be kept to below one half of the number of observations.

4.8.4 Validation

The statistical significance of discriminant functions should be tested by, e.g. a *chi*-square test [14] or by the "jackknifed" leave-one-out technique [15]. According to Kier [16], the quality of the discriminant function may be assessed in three ways: comparison of the *F*-value to tabulated values, determination of the percentage of correctly classified molecules, and prediction of the classification of a test set not included in the original training series. Another approach was followed by Ogino et al. [2]. The best set of discriminant functions was selected, in such a way that 1) a combination of variables, which minimizes the number of misclassified compounds is best, 2) the smallest number of independent variables is used, and 3) the collinearity among the independent variables is minimized.

4.8.5 Examples

Discriminant analysis has been used in various SAR studies, e.g. as in the following:

MAO inhibitors [15, 17]
Antitumor naphthoquinones [18]
Pyrimidine folic acid antagonists [19]
Phenylalkylamines [20]
CNS drugs [21]
Sweet or bitter aldoximes [16]
Antiulcer and antiinflammatory drugs [2]
Mitomycin derivatives and steroids [22]
Carcinogenic aromatic amines [23]
N-nitrosoamines [24]
Perillartine derivatives as sweeteners [9]
Non-narcotic analgetics [25]
Fungicidal 2-antilinopyrimidines [26]
Antiviral *N*-quinolin-4-yl-*N'*-benzylidenehydrazines [27]
Antiinflammatory steroids [7]
Calmodulin inhibitors [28]
Olfactory stimulants [29, 30]
Biodegradation [14]
Genotoxic activity [31]
Anticancer retinoids [36]

In the overview given here, we have selected a few representative examples, which will be discussed below.

4.8.5.1 Mode of Action of Pyrimidine Folic Acid Antagonists

The inhibition of dihydrofolic acid reductase (DHFR) as been subject to many traditional Hansch-type QSAR studies. Part of the differences between the various

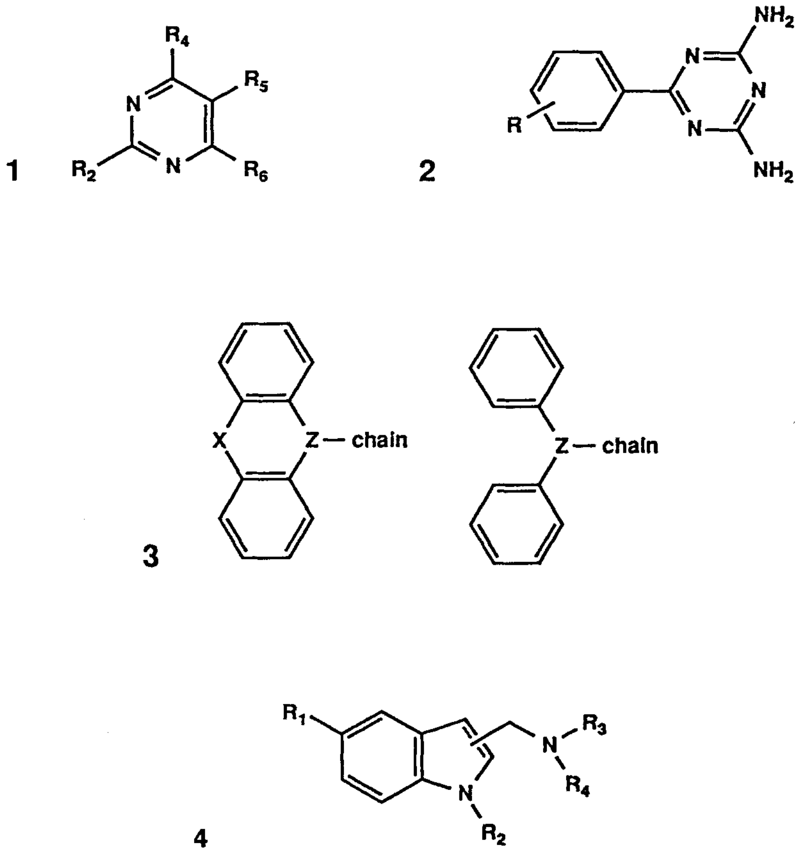


Figure 3. Structures used in discriminant analyzes studies. (1) Antibacterial pyrimidines [19], (2) Antiulcer benzoguanamines [2], (3) Calmodulin inhibitors [28] and (4) Monoamine oxidase inhibitors [15].

compounds are based on their selectivity towards specific of species DHFR, as well as on the differences in cell penetration and metabolism. For a series of 175 pyrimidines (Fig. 3, Structure **1**) studied in an antimalaria program growth, inhibition of *S. faecium* has also been studied [19]. Of these, 155 were classified as reversible or irreversible in their mode of action; the other 20 were inactive. These data were submitted to regression analyzes [32], which gave regression equations describing the structural features responsible for reversible and irreversible inhibition of several bacterial systems (*S. faecium*, *L. casei* and *P. cerevisiae*). However, it was not clear in quantitative terms, which factors were related to the mode of action. Therefore, the same data were subjected to discriminant analysis. Using 123 molecules as the training set, while retaining 32 for prediction purposes, the following classification functions were determined,

$$\text{reversible} = 2.56\pi_2 + 14.21I_2 + 5.96I_5 + 5.40I_{11} + 9.57I_{13} - 7.22 \qquad (5)$$

$$\text{irreversible} = 4.14\pi_2 + 9.34I_2 + 10.57I_5 + 11.51I_{11} + 16.62I_{13} - 8.23 \quad (6)$$

$$D_{12} = -1.57\pi_2 + 4.87I_2 - 4.61I_5 - 6.11I_{11} - 70.5I_{13} + 1.02 \quad (7)$$

where π_2 is the lipophilicity of the substituent at the 2-position, I_2, I_5, I_{11} and I_{13} are indicator values described by Coats et al. [32]. These functions correctly classify 71 of the 78 reversible inhibitors (91%), and 41 of the 45 irreversible inhibitors (91%). For the prediction set, 20 out of 20 reversible inhibitors (100%) and 9 out of 12 irreversible inhibitors (76%) were grouped correctly. The negative coefficient for π_2 in the discriminant function D_{12} indicates that lipophilic substituents at the 2-position of the pyrimidine ring gave rise to irreversible inhibition, with respect to folic acid. The misclassifications could, thus, be rationalized by comparing structural features to the discriminant function.

4.8.5.2 Antiulcer Benzoguanamines

A set of 34 benzoguanamines were tested for their antiulcer activity, expressed as the percent inhibition of the control [2]. Physico-chemical descriptors used in the discriminant analysis include the following: log P values for the unionized form, Hammett constants, σ, and the Swain and Lupton field parameter, $\mathscr{F}$. In discriminant analysis the groups are preestablished, mostly from their natural grouping based on the frequency distribution of the response level. In this particular example, three groups of ca. equal size were formed: the most active, of intermediate actively, and the least active. Three-group and two-group analyzes were performed, and the three-group model gave the following equations:

$$Z(1) = 14.00 \log P + 7.65\Sigma\sigma + 6.82\mathscr{F} - 17.40 \quad (8)$$

$$Z(2) = 11.75 \log P + 5.83\Sigma\sigma + 6.69\mathscr{F} - 12.23 \quad (9)$$

$$Z(3) = 7.01 \log P + 1.09\Sigma\sigma + 7.17\mathscr{F} - 4.56 \quad (10)$$

where $\Sigma\sigma$ is the sum of the Hammett constants of all substituents in the ring. Two-group discriminant functions give similar equations, e.g.,

$$Z(1) - Z(2) = \log P + 0.658\Sigma\sigma + 0.25\mathscr{F} - 2.28 \quad (11)$$

The predictability in the two-group analysis (ca. 80%) appears to be better than by dividing into three groups (60–80%). A certain level of error must always be expected, since better defined categorizations are not possible with biological data. A further improvement in two-group classification was attempted by using "admissible" discriminant functions. These are slight modifications of the usual discriminant method. In this particular case, no improvement was found.

4.8.5.3 Calmodulin Inhibitors

Calmodulin is an intracellular calcium binding protein, involved in the activation of various enzyme systems, such as phosphodiesterase (PDE) and myosin light chain kinase (MLCK). Inhibitors of calmodulin have been classified into four groups [33]. The first three of these groups have been characterized by a discriminant analysis study, in which the following descriptors, among other descriptors, of the structure type **3** were considered [28]:

All molecules consist of a ring with a chain part connected to atom, Z.

Geometrical parameters:

S_r (solvent accessible surface (*SAS*) of the ring),
N_{OH} (number of OH groups in a ring)

Electronic parameters:

Q_Z (atomic charge of atom Z),
EP surface areas (*SAS* divided by the level of electrostatic potential), where $S[-3 < EP < +3]$ is a measure of the hydrophobicity of the molecule

Because most of the parameters chosen here are based on the three-dimensional molecular structure, a low-energy conformer for each compound had to be selected. As far as possible, X-ray structures were used, and others were estimated by MNDO calculations.

Discriminant analysis produced a set of three discriminant functions giving rise to a complete separation of the 22 compounds into three groups:

$$Y(\mathrm{I}) = -4.45S_c[EP > +3] + 43.7S_r + 4.14S_r[-3 < EP < +3] - 90.36 \qquad (12)$$

$$Y(\mathrm{II}) = -6.05S_c[EP > +3] + 44.8S_r + 12.76S_r[-3 < EP < +3] - 111.65 \qquad (13)$$

$$Y(\mathrm{III}) = -1.70S_c[EP > +3] + 28.1S_r + 3.03S_r[-3 < EP < +3] - 39.85 \qquad (14)$$

where the subscripts r and c stand for contributions of the ring and side-chain, respectively. The interpretation of these functions is as follows:

Group II shows a smaller *SAS* area, with the ring having a positive potential and a larger hydrophobic area than Groups I and III; Group III has a larger *SAS* area, with the side-chain having a positive potential and the ring having a smaller total area than Groups I and II. Based on this model, twenty-nine additional inhibitors have been classified. The compounds of Group I have also been studied by a QSAR analysis, using adaptive least squares (ALS), showing that hydrophibicity is important for the ring, but not for the side-chain. The negative potential *SAS* of the side chain is required for activity. In this QSAR analysis, conformation-dependent parameters were used for sets of conformers. Thus, a simultaneous selection of the best set of conformers and the best subset of structural parameters was attained.

4.8.5.4 MAO Inhibitors

Monoamine oxidase (MAO, EC 1.4.3.4) is an enzyme, bound to the mitochondrial membrane, involved in the desamination of biogenic and xenobiotic monoamines, particularly of various neurotransmitters. Two forms of this enzyme. MAO-A and MAO-B, have been characterized. MAO-B inhibitors are of interest for the treatment of Parkinson's disease, while inhibitors for the MAO-A form, such as moclobemide, are used as antidepressants.

Discriminant analysis has been used to develop relationships between physical properties and MAO inhibition by aminotetralins and aminoindans [17, 34]. More recently the selectivity of a series of indole inhibitors of MAO-A and MAO-B was studied [15] (Structure **4** in Fig. 3). The discriminant method LDA was compared to the non-discriminant method, kNN (*k*-nearest neighbors [37]). Using the full data set based on a Free-Wilson matrix, a total of 93.4% correct predictions could be attained. The predictive capability of the method is seen by a "jackknifed" classification, giving 87.9% correct predictions for selective and non-selective compounds. A problem arises in the visualization of the selective and non-selective groups in the 16-dimensional space. The distance between points can be calculated by the Euclidean distance, while the distance between groups can be expressed by the generalized or Mahalanobis distance [38]. A useful graphical representation of the separation of both classes was obtained by plotting the Mahalanobis distance of a compound of the first class mean against the corresponding distance of the same compound to the second class mean. The Mahalanobis distance between the objects i and j is calculated as follows,

$$d^2(i, j) = (x_i - x_j)' \, (x'x)^{-1} \, (x_i - x_j) \tag{15}$$

where $(x'x)^{-1}$ is the variance-covariance (or correlation) matrix. When the variance matrix is the unity matrix, this distance coincides with the Euclidean distance. It was observed further that the non-selective compounds are much more widely distributed. This is to be expected, since non-selectivity may arise from various origins. With the kNN method, 100% of the tightly clustered selective compounds, and 85% of the non-selective are correctly predicted. The kNN method was superior in this case, since it is less sensitive to asymmetrical distribution of the compounds in the variable space.

4.8.6 Conclusions

Discriminant analysis is a simple pattern recognition tool for quickly elucidating structure-property correlations in data sets with categorized biological data [35]. The following steps are involved:

- grouping of biological data
- definition of the groups (usually two or three)
- generation and/or measurement of (physico-)chemical data
- autoscaling of all data to remove unequal weights
- feature selection (= selection of the final descriptor set)
- calculation of the LDA hyperplane
- validation of training and test set by various methods
- prediction of activities of new compounds

The advantage of LDA is that the discriminant functions can be easily understood in terms of the available variables. The disadvantage is that it does not work with embedded or asymmetric data. It is good practice to combine discriminant analysis with other pattern recognition methods, particularly when the number of activity classes is not known or can only be loosely defined, such as agonist-partial agonist-antagonist.

References

[1] McFarland, J. W., and Gans, D. J., Linear Discriminant Analysis and Cluster Significance Analysis. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Hansch, C., Sammes, P. G., and Taylor, J. B., eds., Pergamon Press, Oxford (1990) p. 667–689

[2] Ogino, A., Matsumura, S., and Fujita, T., *J. Med. Chem.* **23**, 437–444 (1980)

[3] Livingstone, D. J., and Manallack, D. T., *J. Med. Chem.* **36**, 1295–1297 (1993)

[4] Salt, D. W., Yildiz, N., Livingstone, D. J., and Tinsley, C. J., *Pestic. Sci.* **36**, 161–170 (1992)

[5] Stuper, A. J., Brügger, W. E., and Jurs, P. C., *Computer-Assisted Studies of Chemical Structure and Biological Function*, Chapter 4, Wiley, New York (1979) p. 95–125

[6] Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches*. Methods and Principles in Medicinal Chemistry, Vol. **1**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1993

[7] Stouch, T. R., and Jurs, P. C., *J. Med. Chem.* **29**, 2125–2135 (1986)

[8] van de Waterbeemd, H., and Testa, B., *Adv. Drug Res.* **16**, 85–225 (1987)

[9] Takahashi, Y., Miyashita, Y., Tanaka, Y., Hayasaka, H., Abe H., and Sasaki, S., *J. Pharm. Sci.* **73**, 737–741 (1984)

[10] Topliss, J. G., and Edwards, R. P., *J. Med. Chem.* **22**, 1238–1244 (1979)

[11] Stouch, T. R., and Jurs, P: C., *J. Chem. Inf., Comput., Sci.* **25**, 45–50 (1985)

[12] Stouch, T. R., and Jurs, P. C., *J. Chem. Inf. Comput. Sci.* **25**, 92–98 (1985)

[13] Stouch, T. R., and Jurs, P. C., *Quant. Struct.-Act. Relat.* **5**, 57–61 (1986)

[14] Gombar, V. K., and Enslein, K., A Structure-Biodegradability Relationship Model by Discriminant Analysis. In: *Applied Multivariate Analysis in SAR and Environmental Studies*, Devillers, J., and Karcher, W., eds., ECSC, Brussels (1991) p. 377–414

[15] Cativiela, C., Garcia, J. I., Fernandez-Alvarez, E., and Elorriaga, C., *Acta Chim. Hung.* **130**, 129–143 (1992)

[16] Kier, L. B., *J. Pharm. Sci.* **69**, 416–419 (1980)

[17] Martin, Y. C., Holland, J. B., Jarboe, C. H., and Plotnikoff, N., *J. Med. Chem.* **17**, 409–413 (1974)

[18] Prakash, G., and Hodnett, E. M., *J. Med. Chem.* **21**, 369–374 (1978)

[19] Smith, C. C., Genther, C. S., and Coats, E. A., *Eur. J. Med. Chem.* **14**, 271–276 (1979)

[20] Bercher, H., Laass, W., Schult, E., Grisk, A., and Franke, R., *Pharmazie* **34**, 336–337 (1979)

[21] Henry, D. R., and Block, J. H., *J. Med. Chem.* **22**, 465–472 (1979)

[22] Moriguchi, I., Komatsu, K., and Matsushita, Y., *Anal. Chim. Acta* **133**, 625–636 (1981)

[23] Yuta, K., and Jurs, P. C., *J. Med. Chem.* **24**, 241–251 (1981)

[24] Rose, S. L., and Jurs, P. C., *J. Med. Chem.* **25**, 769–776 (1982)

[25] Broto, P., Moreau, G., and Vandycke, C., *Eur. J. Med. Chem.* **19**, 79–84 (1984)

[26] Krause, G., Klepel, M., and Franke, R., Stepwise Variation Strategy in the Evaluation of QSAR for Fungicidal 2-Anilinopyrimidines by Means of Discriminant Analysis. In: *QSAR and Strategies in the Design of Bioactive Compounds*, Seydel, J. K., ed., VCH, Weinheim (1985) p. 416–419

[27] Gombar, V. K., *Arzneim.-Forsch.* **35**, 1633–1636 (1985)

[28] Liu, Q., Hirono, S., and Moriguchi, I., *Chem. Pharm. Bull.* **38**, 2184–2189 (1990)

[29] Jurs, P. C., and Edwards, P. A., Computer-Aided Studies of Molecular Structure and Olfactory Properties. In: *Computational Chemical Graph Theory*, Rouvray, D. H., and Randić, M., eds., Nova Press, New York (1990) p. 279–298

[30] Edwards, P. A., Anker, L. S., and Jurs, P. C., *Chem. Sens.* **16**, 447–465 (1991)
[31] Stouch, T. R., and Jurs, P. C., *Environ. Health Perspect.* **61**, 329–343 (1985)
[32] Coats, E. A., Genther, C. S., and Smith, C. C., *Eur. J. Med. Chem.* **14**, 261–270 (1979)
[33] Zimmer, M., and Hofmann, F., *Eur. J. Biochem.* **164**, 411–420 (1987)
[34] Martin, Y. C., *Quantitative Drug Design. A Critical Introduction*, Marcel Dekker, New York, 1978
[35] Dunn, W. J., and Wold, S., Pattern Recognition Techniques in Drug Design. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Hansch, C., Sammes, P. G., and Taylor, J. B., eds., Pergamon Press., Oxford; 691–714 (1990) p. 691–714
[36] Jaeger, E. P., Jurs, P. C., and Stouch, T. R., *Eur. J. Med. Chem.* **28**, 275–290 (1993)
[37] van de Waterbeemd, H., ed., *Advanced Computer-Assisted Techniques in Drug Discovery*, Methods and Principles in Medicinal Chemistry, Vol. **3**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1995
[38] Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., and Kaufman, L., *Chemometrics: a Textbook*, Elsevier, Amsterdam, 1988

4.9 Cluster Significance Analysis

James W. McFarland and Daniel J. Gans

Abbreviations and Symbols

CL	Confidence limits
CSA	Cluster significance analysis
IC_{50}	50% inhibitory concentration
LDA	Linear discriminant analysis
MR	Molar refractivity
MSD	Mean squared distance
π	Hansch-Fujita hydrophobic substituent constant
p	Probability
SARs	Structure-activity relationships

4.9.1 Introduction

Medicinal chemists have diverse interests, but one of the more common is to understand how changes in chemical structure relate to changes in biological activity. For a chemist attempting to discover a better drug, such an understanding would greatly ease his or her task. Unfortunately, such knowledge is hard to obtain. The reason for this is that a single change in the structure leads to many changes in the properties of the compound. For example, substituting a methyl group on a basic nitrogen may alter not only the compound's potency, but also its pK_a, hydrogen bonding capacity, lipophilicity, and extension in space. When many such analogs are considered together, it is difficult to see the structure-activity relationships (SARs) in so many dimensions.

This volume describes a number of statistical methods for detecting such relationships in multivariate space. Cluster Significance Analysis (CSA) [1, 2] is another, but one that can be used in the important case of biological data expressed as one of two responses, for example: "active-inactive" or "agonist-antagonist". While the biological data must be binary, the descriptors can be continuous variables. In this regard, CSA resembles one important aspect of Linear Discriminant Analysis (LDA; see Chap. 4.8). However, it differs from LDA in that it can treat data sets in which the compounds giving the biological response of interest are clustered in the descriptor space, with the non-responders scattered in all directions from this group. Such data distribution has been termed "asymmetric" by Dunn and Wold [3] and is also known as "embedded data" (see Chap. 4.5 in [15]). LDA can not treat situations of this type.

The purpose of CSA is to identify among many possible descriptors those that truly influence activity. CSA can be used to identify such descriptors in asymmetric data and give a general idea about what the optimal descriptor values are; however, it does not furnish a precise classification rule. Nevertheless, CSA can give insights that are valuable when trying to discover SARs.

4.9.2 Theoretical Background to CSA

4.9.2.1 Parameter Focusing

CSA was derived to complement a graphical concept called "parameter focusing", originated by Magee [4]. Fig. 1 illustrates the basic idea of this concept. It presents hypothetical biological test results on six compounds that are characterized by the physical properties, *X* and *Y*. Compounds **1–3** are active (▲), while **4–6** are inactive (○). It appears that the actives are clustered in the midst of the inactives, and can be considered as "focused". From this arrangement, we can judge that *X* and *Y*, or at least one of them, are determinants of biological activity.

How did we decide this? The logic is as follows. If *X* and *Y* had *no* influence on biological activity, we would expect the "actives" to be distributed randomly throughout the graph (the null hypothesis). If, instead, the actives are localized in one region (the alternative hypothesis), so that they are not scattered, then we may infer that these descriptors are related to the biological response. "Focused" clusters indicate non-random, i.e. informative, descriptor patterns. The problem is deciding when a group is indeed focused. How can we tell that the "focused" group did not simply arise by chance? CSA addresses this question.

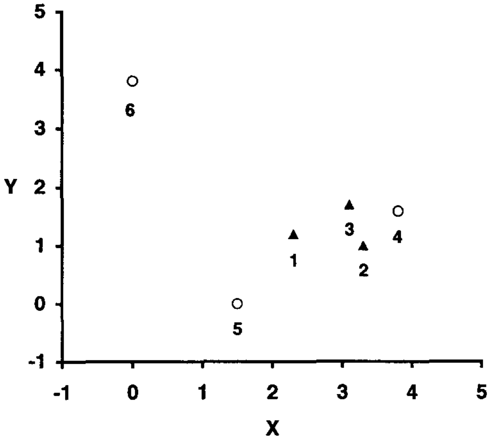


Figure 1. Hypothetical case for CSA. Active (▲) and inactive (○) compounds plotted in the space of the physical properties *X* and *Y*. Reprinted with modification and permission from Ref. [1]; Copyright 1986 American Chemical Society.

4.9.2.2 A Graphical Explanation

Before getting to the mathematics, it would be helpful to illustrate these ideas by continuing with the example of Fig. 1. There are six compounds; of these, three are "active". From the formula for combinations, there are 20 possible sets of six things taken three at a time. If it is chance alone that isoperating, the three active compounds could have appeared as any one of the 20 sets with equal likelihood. That is, the observed placement of the three actives in Fig. 1 could have arisen by accident with a probability of 0.05. However, we would be as or more convinced that the actives were "focused" if they appeared in a set which was as or more compact. Therefore, a significance probability or p-value for testing the null hypothesis of randomness is the ratio formed by the number of all such compact sets divided by the total number of possible sets.

In Fig. 1, how many sets of three are as compact or smaller than the observed active group? This case is so simple that one can easily obtain an answer by inspection. The active group itself, compounds **1**–**3**, is of course one set satisfying the condition, but compounds **2**–**4** form an even smaller set. The Sets **1**, **2** and **4**, and **1**, **3** and **4** are close in size to the active group, but the members are somewhat farther apart from each other. All other groups include compounds **5** and/or **6** and are, therefore, much more loosely associated. Thus, there are only two sets that are as or more compact than the observed group of actives. Therefore, the probability that this degree of clustering would occur by chance alone is:

$$p = 2/20 = 0.10 \tag{1}$$

The p-value of Eq. (1) exceeds the usually accepted maximum of 0.05 for significance. However, with so few compounds in this data set we can at least suspect that the null hypothesis may be false, and that the active compounds are "focused".

4.9.2.3 Calculations

Mean Squared Distances

Cases which are this simple are rare. A more rigorous mathematical treatment, one that can be programmed for a computer, is needed to handle larger and more complex situations. However, the same basic principle remains unchanged: with sets containing the same number of compounds as the observed active set, count those that are as compact or smaller than the observed active set, and divide that count by the total number of allowed possible sets.

We will begin by defining the compactness of a group as the mean squared distance (*MSD*) between compounds as represented by points in a multidimensional space. It is calculated by summing the squared distances between all pairs of points and dividing that total by the number of pairs. Thus, the *MSD* of the active group in Fig. 1 would be computed as follows:

$$\text{total squared distance} = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (x_1 - x_3)^2 + (y_1 - y_3)^2 + (x_2 - x_3)^2 + (y_2 - y_3)^2 \quad (2)$$

$$MSD = (\text{total squared distance})/3 \quad (3)$$

We recommend autoscaling (transforming linearly to unit variance) all descriptors before the distance computations are made. CSA1 and CSA2 are two computer programs (see below) that autoscale automatically.

The FORTRAN Computer Programs CSA1 and CSA2

With this definition of *MSD* in hand, we can proceed to compare the compactness of each allowed set to that of the observed active group, counting those at least as compact, and computing the *p*-value. The program CSA1 does this exhaustively over all allowed sets. However, this process is computationally demanding and can take a long time for some problems. For example, a set of 20 compounds with 9 actives requires 167, 960 *MSD*s to be determined. A VAX computer can handle this in just a few minutes. However, with the addition of more compounds, the computation becomes increasingly time-consuming. A set of 24 compounds with 13 actives entails the calculation of 2, 496, 144 *MSD*s. While this is still possible to calculate with CSA1, the c.p.u. time rapidly increases.

The program CSA2 was created to handle larger data sets. It operates on the same principles as CSA1, but instead of *exhaustively* calculating all possible *MSD*s, it samples at random a predetermined number of allowed sets. The *p*-value is estimated from the number of randomly chosen sets that have *MSD*s equal to or smaller than that of the observed active set. Because this approach is stochastic, there will be some uncertainty in the probabilities estimated in this way. The program also calculates 95% confidence limits for the actual *p*-values.

These two FORTRAN programs are discussed further in Sect. 4.9.3.1.

4.9.2.4 Choosing among Sets of Parameters (Sequential CSA)

Up to this point we have been discussing CSA as if all descriptors under consideration must either all succeed or all fail to be true determinants of activity. Realistically, when there is more than one descriptor, the problem becomes more complex: even if there is genuine clustering, not all the descriptors need be contributing. A feature of CSA is that it can help separate relevant descriptors from irrelevant ones. Sequential CSA is an efficient means to achieve this.

When only a few descriptors are under consideration, they can be evaluated in various combinations in order to arrive readily at a conclusion. However, as descriptors increase, the number of possible combinations expands rapidly. This number is $2^k - 1$, where k is the number of descriptors. With k only 5, there are already 31 combinations; the task of discovering the biologically relevant ones becomes tedious. To reduce the labor involved, we proposed [5] a sequential approach to CSA, one which allows satisfactory conclusions to be drawn, without having to consider all possible combinations.

In the absence of an equivalent to the partial F-test in multiple regression analyses, we suggested that a descriptor's importance may be quantified by considering its effect on the overall p-value, when it is added or deleted. We proposed the following operational rules as guidelines for deciding whether to include a descriptor in the model:

- the addition of a *relevant* descriptor to the model lowers the p-value for the "active" cluster;
- the addition of a *non-relevant* descriptor increases the p-value;
- the subtraction of a *relevant* descriptor increases the p-value;
- the subtraction of a *non-relevant* descriptor lowers the p-value.

These rules are somewhat heuristic, but they are inspired by the notion of "information" or "noise" being added to or subtracted from the data set, and have an intuitive appeal. Thus, relevant descriptors can be identified rapidly by the following steps:

1. Determine the p-value of each descriptor when used alone.
2. Arrange the descriptors, k in number, in a list with increasing p-value. Analyze this list sequentially by CSA, first considering k descriptors. Next, omit the descriptor with the highest p-value, i.e. the last descriptor, and analyze the $k - 1$ remaining descriptors. Continue in this manner until only the first descriptor (the one with the lowest p-value) remains.
3. Using the selection rules above, decide which of these descriptors are contributing positively to the model and which are not. At this point leave out those variables which are not contributing.

If ambiguities remain, repeat the sequence of steps with those descriptors that are still viable possibilities. An example of this process is given in Sec. 4.9.4.3.

4.9.3 Practical Considerations

4.9.3.1 Software Availability

A Commercially Available Program

Oxford Molecular (Oxford, U.K.) offers a version of CSA as part of its software package TSAR (Version 2.1). TSAR requires 16 MB of RAM and 24 MB of disk space on Silicon Graphics, Hewlett-Packard 700 series, or IBM Risc 6000 workstations. The CSA part of the package has a convenient interface for entering data, and for selecting the independent variables for analysis.

Do-It-Yourself

In an earlier publication [1], we presented efficient mathematical algorithms in enough detail to enable users who so wish to write their own computer programs to implement CSA. Many, however, will prefer to start from existing FORTRAN code for CSA1 and CSA2, given in Ref. [2]. From our contacts with many people who have expressed an interest in CSA, we are aware that these programs have been implemented successfully on VAX and IBM mainframes, IBM PCs and their clones, and, as indicated above, various workstation systems. At Pfizer we have even adapted CSA for a Cray supercomputer. Thus, CSA can be used on a variety of hardware and operating systems. FORTRAN is not indispensable: the algorithms in Ref. [1] are independent of language, and CSA1 and CSA2 have been written in PASCAL and RPL [the programming language of RS/1 (BBN, Cambridge MA, USA)]. Thus, CSA can be used on almost any reasonably fast computer system.

Although the information given in Refs. [1] and [2] is complete in all necessary respects, to help those interested in creating their own CSA programs we have written a few pages of additional suggestions on organizing and manipulating data files. We will send this document and, if desired, the FORTRAN codes for CSA1 and CSA2 on request (to J.W.M.).

4.9.3.2 Dividing the Dependent Variable into Two Categories

Biological test results for a series of compounds are sometimes presented qualitatively, such that they can be divided readily into two classes of responses, e.g., "+" for "mutagenic" and "−" for "non-mutagenic". Data of this type present no difficulties.

When the results are given in a graded manner (−, ±, + and + +, for example) those familiar with the test may see a natural division between ± and +, and divide the data into two classes at this point. Other divisions are possible, of course, but they should be decided on the basis of the analytical objective in mind. For instance, one might really only be interested in the factors leading to strong activity. In this case the "+ +" compounds would be the group of interest and all the others would constitute the non-responders.

In some cases the compounds differ only in degree of potency; that is, there are no inactives. One may find that when the compounds are ordered in decreasing potency, there may be a large step in the middle of the list; this then could be a natural division point to generate the two classes necessary. When this does not occur, then an arbitrary division may succeed. An example of this type is given in Sec. 4.9.4.2.

Table 1. The hypothetical example: data used to generate Fig. 1

Compound	Activity[a]	*X*	*Y*
1	1.0	2.3	1.2
2	1.0	3.3	1.0
3	1.0	3.1	1.7
4	0.0	3.8	1.6
5	0.0	1.5	0.0
6	0.0	0.0	3.8

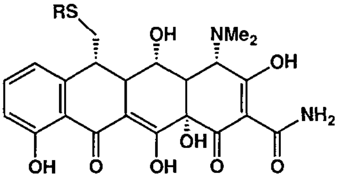
[a] Active = 1.0; inactive = 0.0.

4.9.4 Examples

4.9.4.1 CSA1: The Hypothetical Example

In Sec. 4.9.2.2 we inspected Fig. 1 in order to obtain an indication (with $p = 0.10$) that the "active" compounds are clustered. Because of the small number of compounds involved, we accepted this result as suggesting that at least one of X and Y is a determinant of biological activity. The two dimensional nature of Fig. 1 allows one to see this readily. However, a more difficult question would be: are both X and Y determinants? The answer to this is not so straightforward. If you project the six points in Fig. 1 onto the X and Y axes, it is not clear just by inspection what the respective p-values in each dimension will be. The solution to this problem can be obtained by applying CSA1.

Table 2. Chemical structure, inhibitory potencies and physico-chemical descriptors for arylthio and alkylthio derivatives of methacycline[a]



Compound No.	R	Log $(1/IC_{50})$	L	B_1	B_4	π
1	phenyl	−0.838	6.28	1.71	3.11	2.13
2	4-chlorophenyl	−0.204	7.74	1.80	3.11	2.83
3	4-bromophenyl	−0.146	8.05	1.80	3.11	3.32
4	4-methoxyphenyl	−0.079	8.20	1.80	3.11	2.09
5	benzyl	−0.146	3.63	1.52	6.02	2.63
6	4-chlorobenzyl	−0.322	4.42	1.52	7.44	3.33
7	3,4-dichlorophenyl	−0.716	4.42	1.52	7.44	4.03
10	methyl	−0.898	3.00	1.52	2.04	0.50
11	ethyl	0.222	4.11	1.52	2.97	1.00
12	*n*-propyl	0.222	5.05	1.52	3.49	1.50
13	*i*-propyl	0.222	4.11	2.04	3.16	1.30
14	*n*-butyl	0.155	6.17	1.52	4.42	2.00
15	*i*-butyl	0.255	5.05	1.90	3.49	1.80
16	*t*-butyl	0.301	4.11	2.59	2.97	2.00
17	*n*-hexyl	−0.643	8.22	1.52	5.87	2.50
18	cyclohexyl	0.222	6.17	2.04	3.49	2.50
19	cyclopentyl	0.533	4.97	2.04	3.98	2.14
20	*i*-pentyl	0.533	6.17	1.52	4.42	2.30
21	*n*-decyl	−1.017	12.32	1.52	8.80	5.50
24	2-hydroxyethyl	−0.672	4.79	1.52	3.38	0.39
25	2,3-dihydroxypropyl	−0.755	5.73	1.52	3.38	0.29
27	3-chloropropyl	0.222	6.82	1.52	3.49	2.20

[a] Reprinted with modification and permission from Ref. [6]; Copyright 1993 American Chemical Society.

The data used to generate Fig. 1 are listed in Table 1. Of course, CSA1 gives the p-value of 0.10 when both X and Y are included in the analysis. When we chose just the X parameter, the p-value remained unchanged at 0.10. Thus, the parameter X may be all that is necessary to account for the clustering. When we assessed the contribution of Y alone, however, the p-value was now 0.15, greater than that of X, or the $X-Y$ combination. We conclude that adding Y to X provides no improvement, while adding X to Y does. This suggests that X is a possible determinant of activity, while Y is probably not. The good p-value observed for the $X-Y$ combination appears to be due solely to the contribution of X.

4.9.4.2 CSA2: Inhibition of Tetracycline Efflux Antiport Protein

Recently, Nelson et al. [6] published a set of data that affords an instructive example of CSA, using the random sampling technique (CSA2). The problem concerns a set of 27 arylthio and alkylthio derivatives of methacycline that inhibit a tetracycline efflux antiport protein isolated from a tetracycline-resistant bacterium. Twenty-two of the compounds considered and some of their properties are presented in Table 2. Five of the tetracyclines in the original work were omitted, because they had structural features that did not permit meaningful values of π to be assigned to them. These were beyond the scope of the general structure given at the top of Table 2.

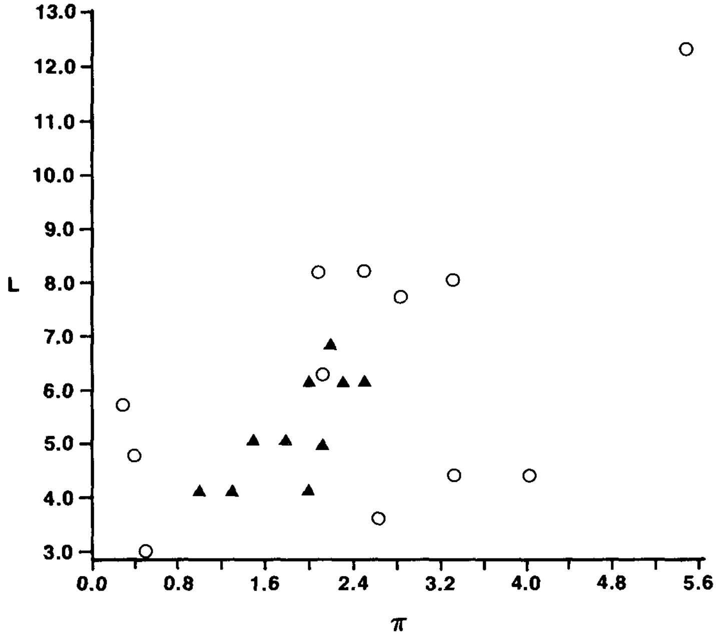


Figure 2. A plot of 22 arylthio and alkylthio derivatives of methacycline in the dimensions of L and π. The more potent compounds (▲) inhibit a tetracycline efflux antiport protein isolated from a tetracycline resistant bacterium. The less potent compounds are designated by (○).

Table 3. CSA2 results for various combinations of descriptors in the series of methacycline derivatives

#[a]	L	B_1	B_4	π	p-value		Number of subsets sampled
					estimate	95% CL[b]	
4	1	1	1	1	0.071900	±0.003580	20000
3	1	1	1	0	0.219050	±0.005732	20000
	1	1	0	1			
	1	0	1	1	0.000018	±0.000012	500000
	0	1	1	1			
2	1	1	0	0			
	1	0	1	0	0.000650	±0.000158	100000
	1	0	0	1	0.000042	±0.000018	500000
	0	1	1	0			
	0	1	0	1			
	0	0	1	1	0.000330	±0.000113	100000
1	1	0	0	0	0.009400	±0.001337	20000
	0	1	0	0	0.977350	±0.002062	20000
	0	0	1	0	0.012740	±0.000983	50000
	0	0	0	1	0.002410	±0.000304	100000

[a] Number of descriptors in set.

[b] Uncertainty (still at the 95% confidence level) in the difference between two estimated p-values is given by the square root of the sum of the squares of the two individual uncertainties (i.e. the values following the "±" symbols).

The biological response of interest is the IC_{50} (μM) and expressed in Table 2 as log $(1/IC_{50})$, which results in ten compounds with positive values. We used these as the group of interest (i.e. the "actives"). The physical properties are the Verloop STERIMOL parameters L, B_1, and B_4 [see Chap. 2.1], and the Hansch-Fujita hydrophobic substituent constant π [see also Chap. 2.1]. Fig. 2 indicates that the data are asymmetric in the dimensions of L and π; in similar plots the actives appear clustered in B_4 but not in B_1 space.

Table 3 shows the 15 possible combinations of the four descriptors. Each row represents one potential CSA2 run. A "1" in a physical property column indicates the presence of that variable in the combination. The estimated p-value with its 95% confidence limits (95% CL) and the number of random subsets used are also given. As the results illustrate, it is not necessary to run all the combinations for a satisfactory picture to emerge. The last four rows give the p-values for each of the descriptors separately. Clearly, B_1 is not a likely determinant of activity. However, the other three are good candidates, with π being perhaps the most important, followed by L, and then lastly, B_4.

The first row shows that when all four variables are included, the p-value is just short of being statistically significant. However, when π is removed from the set, the p-value becomes considerably higher. From the rules given in Sec. 4.9.2.4, we conclude that π is a relevant determinant. On the other hand, when B_1 is removed from the set of four parameters, there is a dramatic improvement in the p-value.

Table 4. Chemical structures, sweetness and descriptor values for the R group of carbosulfamates[a]

$R-NH-SO_3^-Na^+$

	Taste[b]	*MR*	*L*	B_1	B_0	B_r	B_l	σ^*
n-propyl	1.0	15.0	4.92	1.52	3.49	1.91	1.91	−0.12
n-butyl	1.0	19.6	6.17	1.52	4.43	1.92	1.90	−0.13
2-Me-butyl	1.0	24.2	6.17	1.52	4.43	3.16	1.90	−0.16
isopentyl	1.0	24.3	6.17	1.52	4.43	3.15	1.92	−0.16
isobutyl	1.0	19.6	4.92	1.52	4.21	3.16	1.90	−0.13
neopentyl	1.0	24.3	4.92	1.52	4.22	3.16	3.15	−0.17
c-hexyl	1.0	26.7	6.06	1.91	3.24	3.59	2.81	−0.15
2-Me-*c*-hexyl	1.0	31.3	6.06	1.91	3.24	3.77	3.59	−0.15
3-Me-*c*-hexyl	1.0	31.3	6.15	1.91	4.47	3.59	3.35	−0.15
c-pentyl	1.0	21.5	4.90	1.90	4.06	3.42	2.58	−0.20
2-Me-*c*-pentyl	1.0	26.1	4.90	1.90	4.06	3.62	3.42	−0.20
3-Me-*c*-pentyl	1.0	26.1	5.91	1.90	4.06	3.42	2.58	−0.20
c-pentylmethyl	1.0	26.1	6.05	1.52	4.11	2.86	2.86	−0.13
phenyl	1.0	25.4	6.28	1.71	1.71	3.11	3.11	0.60
ethyl	0.0	10.3	4.11	1.52	2.98	1.91	1.90	−0.10
n-pentyl	0.0	24.3	6.97	1.52	4.94	1.92	1.90	−0.16
isohexyl	0.0	28.9	6.97	1.52	5.66	3.17	1.90	−0.16
2,3-Me_2-*c*-hexyl	0.0	35.9	6.15	1.91	4.47	3.77	3.59	−0.15
2,5-Me_2-*c*-hexyl	0.0	35.9	6.06	2.14	3.56	4.42	4.29	−0.15
2,6-Me_2-*c*-hexyl	0.0	35.9	6.06	1.91	3.24	4.50	3.77	−0.15
3,3,5-Me_3-*c*-hexyl	0.0	40.6	6.26	1.91	4.47	4.50	3.44	−0.15
2-Et-*c*-hexyl	0.0	40.6	6.06	1.91	4.29	4.66	3.59	−0.15
4-*t*-Bu-*c*-hexyl	0.0	40.6	8.20	1.91	4.59	3.52	2.72	−0.15
4-*t*-pentyl-*c*-hexyl	0.0	49.9	8.98	1.91	4.59	3.52	2.72	−0.15
c-butyl	0.0	17.9	4.77	1.77	3.18	3.20	2.04	−0.15
c-hexylmethyl	0.0	31.3	6.12	1.52	5.30	3.26	3.14	−0.13
4-vinyl-*c*-hexyl	0.0	35.4	8.28	1.91	4.01	3.59	2.81	−0.15
benzyl	0.0	30.0	4.62	1.52	6.02	3.13	3.10	0.22
1-adamantyl	0.0	40.6	6.17	3.16	3.16	3.49	3.49	−0.26
methyl	0.0	5.7	2.87	1.52	2.04	1.90	1.90	0.00
n-hexyl	0.0	28.9	8.22	1.52	5.88	1.92	1.90	−0.17
n-heptyl	0.0	33.6	9.03	1.52	6.39	1.92	1.90	−0.17
n-octyl	0.0	38.2	10.27	1.52	7.33	1.93	1.90	−0.15
isopropyl	0.0	15.0	4.11	1.91	3.16	2.98	2.76	−0.12
1-Me-propyl	0.0	19.6	4.92	1.91	3.16	3.49	2.76	−0.21
1-Me-butyl	0.0	24.3	6.17	1.91	4.39	3.54	2.99	−0.23
1-Me-pentyl	0.0	28.9	6.97	1.90	4.94	3.16	2.76	−0.26
1-Me-hexyl	0.0	33.6	8.22	1.91	5.63	4.30	2.99	−0.27
1,2-Me_2-hexyl	0.0	28.9	6.17	1.91	4,39	3.74	2.99	−0.23
1,3-Me_2-hexyl	0.0	28.9	6.17	1.91	4.42	3.49	2.98	−0.26
1,4-Me_2-hexyl	0.0	28.9	6.97	1.91	4.39	4.50	2.99	−0.26
1,2,2-Me_3-propyl	0.0	33.5	4.92	1.91	4.40	3.74	2.99	−0.29
t-butyl	0.0	19.6	4.11	2.77	2.98	3.16	3.15	−0.30
1,1-Me_2-propyl	0.0	24.2	4.92	2.77	3.49	3.16	3.15	−0.31
1,1,3,3-Me_4-butyl	0.0	38.2	6.17	2.60	4.22	3.16	3.15	−0.36
c-propyl	0.0	13.5	4.14	1.55	3.08	3.24	1.81	−0.15
1-Me-*c*-pentyl	0.0	26.1	4.90	2.66	3.24	4.09	3.17	−0.30
1-Me-*c*-hexyl	0.0	31.3	6.06	2.73	3.48	3.30	3.16	−0.44
phenethyl	0.0	34.6	8.33	1.52	3.16	3.12	3.11	0.08
3-phenylpropyl	0.0	39.3	6.67	1.52	7.47	3.14	3.10	0.02

[a] Reprinted with modification and permission from Ref. [7]; Copyright 1986 Elsevier Science Publishers, BV.

[b] Sweet = 1.0; not sweet = 0.0.

Thus, our first impression was confirmed, and B_1 could be eliminated as a determinant. When the remaining three descriptors are considered in pairs, the p-value is, in each case, statistically greater than when all three descriptors are treated together. Hence, the conclusion is that π, L, and B_4 appear to be genuine determinants of tetracycline efflux antiport protein inhibition. From Fig. 2, we can by inspection obtain the approximate ranges in π and L, where the most potent activity will be found; with a similar plot using B_4, the range for good activity in that dimension can also be estimated.

4.9.4.3 Sequential CSA: Sulfamate Sweetening Agents

In the previous example there were only 15 possible combinations of the four descriptors. By first considering the descriptors one-at-a-time we were able to select combinations such that we arrived at a satisfactory answer by using only ten of the 15; in fact this task could have been accomplished with nine had we followed a strict sequence of events. We really did not need to exclude just π; this was done merely to show the effect on the p-value when a relevant descriptor was removed.

To illustrate the sequential approach in more detail, let us consider a set of sulfamate sweetening agents. Miyashita and coworkers [7] first introduced this problem. Seven physical descriptors were considered: molar refractivity (MR), Taft's σ^*, and the STERIMOL parameters L, B_1, B_0, B_r and B_l (see Chap. 2.1). According to the discussion in Sec. 4.9.2.4, there are 127 possible combinations. By plotting the data in the space of their first two principle components, Miyashita et al. [7] showed that the sweet compounds were clustered in the midst of the their non-sweet congeners. However, it was not possible to relate sweetness directly to the original properties. We recently published an analysis of these same data using CSA [5]. The following is a summarized discussion of our previous work.

Table 4 presents the Miyashita data. It consists of properties for 50 sulfamic acids substituted on the nitrogen atom with various alkyl, cycloalkyl, and phenylalkyl groups, and a phenyl group. Sweetness is the biological response of interest; 14 compounds fell into this category. Because there are nearly 10^{12} combinations of 50 things taken 14 at a time, we used the CSA2 program.

We began by evaluating the descriptors one-at-a-time; the results are shown at the bottom of Table 5. From this we found that L is the most likely determinant of sweetness among these sulfamates. After L in decending order of importance are: MR, B_1, B_0, B_r, B_l, and σ^*. Of these, only L, MR, and B_1 are each statistically significant. When all of the descriptors are considered together (top row of Table 5), the cluster of actives is statistically significant. However, when the least significant variable (σ^*) is omitted, the p-value becomes dramatically smaller. Hence, σ^* need not be considered further. We then eliminated the next least significant descriptor (B_l) and found a still further decrease in p-value. Proceeding in this manner, we arrived at a p-value of only 0.00002 ± 0.000003, with just those descriptors found on an independent basis: L, MR, and B_1.

As a final step, as with the Nelson problem, we deleted each descriptor in turn from this collection of three descriptors. Thus, we investigated whether further

Table 5. CSA2 results for various combinations of descriptors of carbosulfamate sweetening agents[a]

#[b]	L	MR	B_1	B_0	B_r	B_t	σ^*	p-value		Number of subsets sampled
								estimate	95% CL[c]	
7	1	1	1	1	1	1	1	0.039600	±0.005400	5000
6	1	1	1	1	1	1	0	0.000600	±0.000215	50000
5	1	1	1	1	1	0	0	0.000050	±0.000031	200000
4	1	1	1	1	0	0	0	0.000010	±0.000006	1000000
3	1	1	1	0	0	0	0	0.000002	±0.000003	1000000
2	1	1	0	0	0	0	0	0.000027	±0.000010	1000000
	1	0	1	0	0	0	0	0.000058	±0.000021	500000
	0	1	1	0	0	0	0	0.000115	±0.000047	200000
1	1	0	0	0	0	0	0	0.000900	±0.000131	200000
	0	1	0	0	0	0	0	0.001260	±0.000311	50000
	0	0	1	0	0	0	0	0.045400	±0.005770	5000
	0	0	0	1	0	0	0	0.061400	±0.006654	5000
	0	0	0	0	1	0	0	0.125600	±0.009186	5000
	0	0	0	0	0	1	0	0.578600	±0.013687	5000
	0	0	0	0	0	0	1	0.720200	±0.012443	5000

[a] Reprinted with modification and permission from Ref. [5]; Copyright 1990 Drug Information Association.

[b] Number of descriptors in set.

[c] Uncertainty (still at the 95% confidence level) in the difference between two estimated p-values is given by the square root of the sum of the squares of the two individual uncertainties (i.e. the values following the "±" symbols).

improvement could be found in combinations taken two-at-a-time. The next entries in Table 5 show that none of these subcombinations of two descriptors is better than all three descriptors together. Thus, by considering only 15 out of 127 possible descriptor combinations, we arrived at a reasonable identification of the determinants of sweetness among sulfamates.

4.9.4.4 Literature Examples

Other examples of CSA may be found in the literature, and Table 6 gives reasonably complete list of such examples.

Table 6. Literature references of other examples of CSA

Ref.	Host (system)	Activity	Compound class	Determinants found
[1]	mouse	monoamine oxidase inhibition	aminotetralins	E_s (steric parameter), Π
[1]	(Ames test)	mutagenicity	aminoacridines	log K, R_m, pK_a, μ
[8]	(in vitro)	antibacterial	lasalocid derivatives	log P
[8]	duck	antimalarial	naphthaquinones	LUMO energy,
[8]	rodent	carcinogenicity	polycyclic aromatic hydrocarbons	excited energy states E_s and $E_{s,t}$
[9]	rat	antihyperptensive	prazosin analogs	π, σ
[9]	chicken	anticoccidial	acridinediones	Swain-Lupton F
[10]	rodent	solid tumor inhibition	diarylsulfonylureas	π, volume
[11]	(not stated)	β-adrenergic: agonists vs antagonists	phenethylamines	hydrogen bond descriptors
[12]	mouse	antitrematodal, anticestodal	pyrazinoisoquinolines	${}^1\chi^V$, μ
[13]	adult female filaria	cidal	antimycin analogs	*ATCH5*, *ATCH4*, *DIPV_X*
[14]	(steroid-binding globulins)	receptor binding	corticosteroids, testosterones	similarity indices

References

[1] McFarland, J. W., and Gans, D. J., *J. Med. Chem.* **29**, 505–514 (1986)

[2] McFarland, J. W., and Gans, D. J., Linear Discriminant Analysis and Cluster Significance Analysis. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Ramsden, C. A., ed., Pergamon Press, Oxford (1990) p. 667–689

[3] Dunn III, W. J., and Wold, S., *J. Med. Chem.* **21**, 1001–1007 (1978)

[4] Magee, P. S., Parameter Focusing — A New QSAR Technique. In: *IUPAC Pesticide Chemistry: Human Welfare and the Environment*, Miyamoto, J., and Kearney, P. C., eds., Pergamon Press, Oxford (1983) p. 251–260

[5] McFarland, J. W., and Gans, D. J., *Drug Information J.* **24**, 705–711 (1990)

[6] Nelson, M. L., Park, B. H., Andrews, J. S., Georgian, V. A., Thomas, R. C., and Levy, S. B., *J. Med. Chem.* **36**, 370–377 (1993)

[7] Miyashita, Y., Takahashi, Y., Takayama, C., Ohkubo, T., Funatsu, K., and Sasaki, S., *Anal. Chim. Acta* **164**, 143–149 (1986)

[8] McFarland, J. W., and Gans, D. J., *J. Med. Chem.* **30**, 46–49 (1987)

[9] McFarland, J. W., and Gans, D. J., Discovering Activity Determinants: Graphics and a Related Probability. In: *QSAR: Quantitative Structure-Activity Relationships in Drug Design*, Fauchère, J. L., ed., Alan R. Liss, New York (1989) p. 199–202

[10] Howbert, J. J., Grossman, C. S., Crowell, T. A., Rieder, B. J., Harper, R. W., Kramer, K. E., Tao, E. V., Aikins, J., Poore, G. A., Rinzel, S. M., Grindey, G. B., Shaw, W. N., and Todd, G. C., *J. Med. Chem.* **33**, 2393–2407 (1990)

[11] Raevsky, O. A., Grigor'ev, V. Y., Kireev, D. B., and Zefirov, N. S., *J. Chim. Phys.* **89**, 1747–1753 (1992)

[12] Ordorica, M. A., Velázquez, M. L., Ordorica, J. G., Escobar, J. L., and Lehmann, P. A., *Quant. Struct.-Act. Relat.* **12**, 246–250 (1993)

[13] McFarland, J. W., and Gans, D. J., *Quant. Struct-Act. Relat.*, **13**, 11–17 (1994)

[14] Anonymous. *TSAR User's Guide*, Issue 3, Oxford Molecular, Oxford, 1993, p. 2–45 to 2–50

[15] van de Waterbeemd, H., ed., *Advanced Computer-Assisted Techniques in Drug Discovery*, Methods and Principles in Medicinal Chemistry, Vol. **3**, R. Mannhold, P. Krogsgaard-Larsen, H. Timmerman, eds., VCH, Weinheim, 1995

Chemometric Methods in Molecular Design
edited by Han van de Waterbeemd
© VCH Verlagsgesellschaft mbH, 1995

5 Statistical Validation of QSAR Results

5.1 Validation Tools

Svante Wold and Lennart Eriksson

Abbreviations and Symbols

CV	Cross-validation
G	Number of CV groups
IC_{50}	Concentration needed to lower cell viability by 50%
LOO	Leave-one-out
LSO	Leave-several-out
MLR	Multiple linear regression
NBP	4-nitrobenzylpyridine
PCA	Principal components analysis
PLS	Projection to latent structures
PRESS	Prediction error sum of squares or Predictive residual sum of squares
Q^2	Amount of predicted (CV) variance
R^2	Amount of modelled sum of squares (variance)
RSD	Residual standard deviation
SS_y	The sum of squares of the response values
X matrix	Table of *N* compounds × *K* structure descriptors

5.1.1 Introduction

The procedure for establishing reliable quantitative structure-activity relationships (QSAR) involves a number of important steps that are closely related. Notably, the most significant of these steps are: (a) the selection of representative compounds with which to calibrate and validate the QSAR (i.e the training set and the validation set), (b) the multivariate chemical characterization of these sets, (c) the biological profiling of these compounds, (d) the QSAR modeling, and (e) the validation of the resulting QSAR model. Some of these steps are usually considered, whereas others are largely neglected, and, in particular, model validation [1 – 3]. This is unfortunate, because a practical consequence is that a QSAR model can not be taken seriously, until its performance in a real situation has been adequately checked.

Any QSAR model needs to be properly validated prior to its use for interpreting and predicting biological responses of non-investigated compounds. But the question arises, how do we assure ourselves that a specific QSAR is valid, and what do we mean by model quality? There exists a number of ways of expressing the performance of a model. The conventional approach adopted in QSAR analysis, based on multiple linear regression (MLR), is to consider R^2, the "explained variance" or (multiple) correlation coefficient, and, s, the residual standard deviation (*RSD*). The former quantity varies between 0 and 1, where 1 means a perfect model, explaining 100% of the response data (Y), and 0 a model without any explanatory power at all. Thus, a high R^2 (close to 1) and a low *RSD* are necessary conditions for model validity. However, excellent values of R^2 and *RSD* are not sufficient indicators of model validity. This depends on the property of regression models (including PLS and PCR) to give a closer fit – the better the replicative capability of the model to the data, the more parameters and terms are incorporated to the model.

Furthermore, if we have many chemical and structural descriptors (X) to choose from, we may construct a QSAR which produces an apparently good relation between calculated and observed response data, even with few descriptor variables, provided that these are selected from the larger set according to their apparent contribution to the fit. Remarkably, this can be achieved, even when a set of descriptor variables has been altogether constructed by means of random numbers, and have no correspondence whatsoever to the biological problem under scrutiny [4,5]. This risk of coincidental correlations is one main reason why stepwise MLR is not to be recommended for data sets composed of a multitude of descriptor variables. Other methods of model fitting, such as PLS, [6–8] should then be used, and this is discussed in other parts of this volume.

Since a high R^2 and a small *RSD* are not sufficient as model validity indicators, alternatives must be provided. In principle, two reasonable approaches of validation can be envisaged, one based on predictions and the other based on the fit of the predictor variables to randomized rearranged response variables. Ideally, of course, the best option would be a comprehensive validation set of representative compounds, which enables predicted values to be compared to the actual observed values, and which allows a reliable estimate of Q^2, the "predicted variance", to be calculated (see below). Obviously because of time and resources, however, adequate validation sets are not common. In the light of this fact, other techniques have been devized, and the objective of this chapter is to outline these techniques.

In essence, four tools of assessing the validity of QSAR models can be differentiated. These are: (i) randomization of the response data into reordered response vectors, (ii) cross-validation, (iii) splitting of the chemical compounds into a training and a validation set, and (iv) confirmation using an independent external validation set. Without the luxury of an independent validation set, which is regarded as the most reliable of these tools, the soundness of the model may, thus, be checked by either of the other three procedures.

5.1.2 Examples

In order to illustrate the four methods of assessing QSAR reliability, we shall consider two examples from the literature. The first example is taken from environmental chemistry and concerns a series of 15 epoxides and a model of their chemical reactivity, which strongly influences the mutagenicity of these compounds [9]. For six of the epoxides, chemical reactivities are available from two different chemical model systems, involving the standard nucleophile 4-nitrobenzylpyridine (NBP). Thus, simultaneous QSAR modeling of two dependent variables (Y) is possible. In the second example, statistical experimental design has been used to create representative training and validation sets among a series of halogenated aliphatic hydrocarbons. Ten chemicals were allocated to the training set, and six to the validation set, and these compounds were subsequently investigated for their cytotoxicity to human cells (expressed as IC_{50} values). Computational and other experimental details of this example have been described by Sjöström et al. [10]

The QSAR models for these two data sets will be reexamined by means of the randomization technique and by cross-validation. Moreover, the epoxide data set gives the possibility of experimental validation using split data sets, whereas the haloalkanes provide the opportunity of demonstrating external validation using a designed validation set. We have also noted that the data analysis was carried out using PLS [6].

5.1.3 Four Tools for Model Validation

5.1.3.1 Tool 1: Randomization of the Responses into an Array of Reordered Variables

The first of the four tools is based on repetitive randomizations of the response data (Y) of N compounds in the training set. Thus, a random number generator is used to allocate the integers between 1 and N to sequences of N numbers. In each cycle, the resulting arrangement of random integers is employed in order to reorder the Y data – leaving the X data intact – and then the full data analysis is carried out on these scrambled data. Every run will yield estimates of R^2 and Q^2, which are recorded. If in each case the scrambled data give much lower R^2 and Q^2 values than the original data, then one can feel confident about the relevance of the "real" QSAR model. Randomization of the Y data a number of times (at least ten) gives a fairly good idea of the significance of the real QSAR, but in order to enhance the precision of the probability level, some hundreds of runs of rerandomized data are usually required. When hundreds of trials have been performed, histograms of R^2 and Q^2 can provide a precise estimate of the significance level of the real QSAR model.

We realize, however, that hundreds of repetitions of the QSAR calculations might be tedious and time-consuming. Fortunately, our experience shows that already at around ten trials, the essential features of the R^2 and Q^2 histograms are already discernible. Moreover, sometimes rather high R^2 and Q^2 are to be expected, because the randomized response variable may be highly correlated to the parent response variable. Thus, it is recommended that one always keeps a track on the interrelationships among the original and reordered data, so that misinterpretations can be avoided.

5.1.3.2 Tool 2: Cross-Validation

In contrast to the previous method, cross-validation (CV) is based on predictions [11, 12]. CV operates by making a number (*G*) of slightly reduced modifications to the parent data set, estimating parameters from each of these modified data sets, and then calculating the precision of the predictions by each of the resulting models. Thus, CV creates *G* modified data sets by taking away one or a small group of compounds from the data in such a way that each observation (here: compound) is taken away once, and only once, over the total number of CV cycles, *G*. The model is then fitted to the data, devoid of the omitted part, and is then used to compute predictions based on the response data of the left out compounds. This is repeated for each modified data set, whereupon the squared differences between predicted and actual response values are summarized to form *PRESS* (Predictive REsidual Sum of Squares, or, alternatively, PRediction Error Sum of Squares). In the end, *PRESS* will contain one contribution from each observation and is, thus, a good indicator of the real predictive capability of the model. Next, *PRESS* is compared to the sum of squares of the response values (SS_Y), and if the former is smaller than the latter, the QSAR predicts better than chance and can be regarded as "statistically significant" [13]. This because the best "estimate" for the activity of each compound is $\bar{y}$, giving the "estimate error" equal to SS_Y. The predictive performance of the QSAR model can be reexpressed as Q^2, the predicted or cross-validated variance, which is (1-$PRESS/SS_Y$) and accompanies the parameter R^2. For more discussion on CV in the context of PLS, we refer to Wold [6] in this volume.

In certain situations CV may not work as one wishes. The first is when the compounds are grouped considerably and, hence, are not independent. This may, for instance, occur when two or more different types of compounds are incorporated into the same model, and the activity of these compounds differs only according to this grouping. Any model will then primarily account for this difference in activity between the groups, and CV as well as any other significance test (except randomization), will identify this triviality. Another situation where CV is misleading occurs when CV is applied *after* variable selection in stepwise MLR. Here, the problem is that the final model is based on descriptors that have been selected according to their correlation with the response values, and the resulting apparent correlations are also, in retrospect, stable in CV. Finally, CV may yield *too conservative* results if the *X* matrix is generated from an orthogonal statistical experimental design, but this is exceptionally rare in QSAR.

Leave One or Several Out?

Intuitively, we may feel that CV gives better precision, the larger the number of CV groups and, hence, *G* cycles. This has led most users to have one compound in each CV group, which gives, of course, *N* groups. The CV procedure is then often called "leave-one-out" (LOO). For computational reasons, however, CV with multivariate models (PLS, PCA, etc) has usually been performed with much fewer groups, typically between five and ten. Interestingly, Shao [14] has recently shown that both theoretically and practically, that this "leave-several-out" (LSO) approach is preferable to LOO. This result can be understood when we consider what happens when the number of compounds, N, increases. The LSO technique always leaves out a certain portion of the data, thus, creating a constant perturbation in the data structure. The LOO approach perturbs the data structure by removing $1/N$th in each CV round, thus, accomplishing an increasingly smaller perturbation with increasing N. Hence, in the limit, the Q^2 of LOO approaches R^2, which is highly unsatisfactory. In short, we recommend setting G around 7 with CV.

5.1.3.3 Tool 3: Splitting of Parent Data Set into Training and Validation Sets

Cross-validation provides a reasonable approximation of the ability with which the QSAR predicts the activity values of new compounds. Usually, this is termed *internal* validation because all the considered chemicals belong to the same data set. However, should the number of available compounds be large enough, they can be divided to form a separate training set and a separate validation set, thus, enabling *external* validation. This subdivision of the data set can be accomplished in many ways, but it is desirable that the two series of compounds span approximately similar ranges of the biological responses and the structural properties.

5.1.3.4 Tool 4: External Validation Using a Designed Validation Set

An often overlooked stage in QSAR is the selection of appropriate training and validation sets, i.e. how to select the sets to meet the fundamental statistical criterion of representativity. The training set and validation set compounds must be representative for the class of compounds from which they originate, which means that they must be chosen in such a manner that they adequately span the chemical and structural properties of the compounds considered. One practical way of attaining such sound sets of chemicals is to use statistical experimental design, which has already been discussed by Sjöström et al. [10] in this book.

The use of statistical experimental design to generate well-balanced training and validation sets of representative compounds is infrequent in QSAR [10]. However, in the case where this has been done, a validation set will exist that spans the entire X-space evenly and is independent of the training set. Provided that analogies and relationships prevail between the chemical and structural properties and the biological responses of these chemicals, this type of high quality validation set would enable the QSAR to be experimentally validated across the entire range of biological activity.

5.1.4 Results

The QSAR models, calculated in the following section were all obtained by PLS using the SIMCA package [15] with cross-validation. Thus, in every model described below, values of R^2 and Q^2 from CV will be quoted.

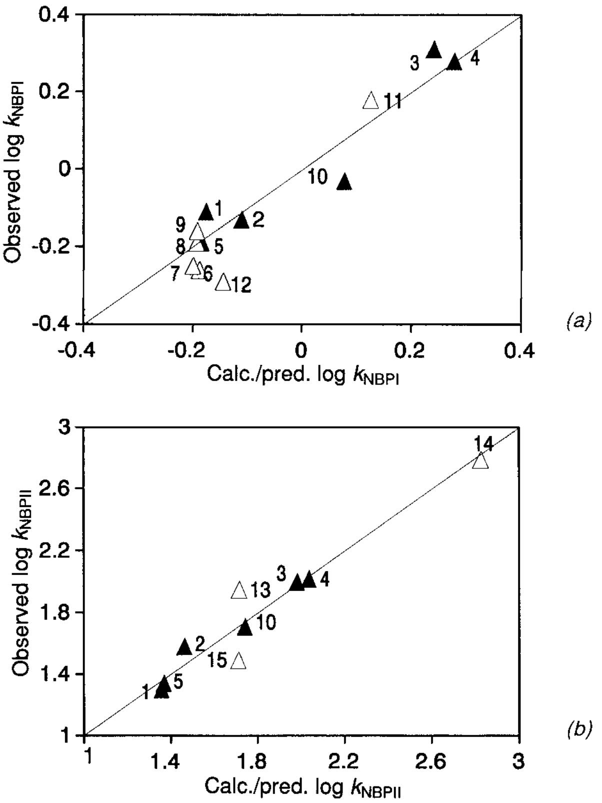


Figure 1. Scatter plots for the observed and calculated (training set, solid triangles)/predicted (validation set, open triangles) values of a) log k_{NPBI} and b) log k_{NBPII}. The epoxides are: propylene oxide (1), glycidol (2), epichlorohydrin (3), epibromohydrin (4), 1,2-epoxybutane (5), 1,2-epoxyhexane (6), 1,2-epoxyoctane (7), 1,2-epoxydecane (8), 1,2-epoxydodecane (9), styrene oxide (10), butadienediepoxide (11), 1,2,7,8-diepoxyoctane (12), epifluorohydrin (13), 3,3,3-trichloropropylene oxide (14), butadiene monoxide (15).

5.1.4.1 The Epoxide Example

The 15 epoxides studied were chemically and structurally characterized using nine theoretical quantum chemical descriptors, such as bond orders, atomic charges, delocalizabilities, electronegativities, and so forth [9]. For six of the compounds, chemical reactivities originating from two related chemical model systems, log k_{NBPI} and log k_{NBPII}, were accessible. The PLS modeling, using these six compounds as the training set, gave the "real" QSAR with $R^2 = 0.94$ and $Q^2 = 0.92$ (obtained with Tool 2), which was further validated using Tool 3. Fig. 1 shows the relationships between the observed and calculated/predicted chemical reactivities for the two endpoints. Evidently, the QSAR is able to adequately forecast the chemical reactivities of the epoxides in the validation sets. What is particularly remarkable is the prediction of log k_{NBPII} of epoxide 14, which corresponds to an extrapolation of nearly 100% outside the range of reactivity of the training set.

Table 1. Observed R^2s (unadjusted) and Q^2s using Tool 1

Epoxides			Haloalkanes		
Trial	R^2	Q^2	Trial	R^2	Q^2
"Real"	0.94	0.92	"Real"	0.90	0.88
1	0.58	0.42	1	0.23	0.07
2	0.14	−0.86	2	0.12	−0.01
3	0.27	−0.92	3	0.60	0.53
4	0.27	−1.15	4	0.24	0.07
5	0.09	−0.99	5	0.06	−0.14
6	0.11	−0.58	6	0.40	0.32
7	0.39	−0.62	7	0.11	−0.03
8	0.05	−1.15	8	0.28	0.14
9	0.65	0.54	9	0.20	−0.08
10	0.06	−0.90	10	0.14	0.04

Next, the validity of this QSAR was tested using Tool 1. Thus, ten randomized and reordered pairs of chemical reactivity variables were constructed and were modeled pairwise. Table 1 lists the resulting explained and predicted variances, which are also shown in Fig. 2. As can be seen, the majority of the R^2s lie in the range of 0.0 to 0.4, and in only two cases (1 and 9 with 0.58 and 0.65, respectively) is this interval exceeded. Similarly, the Q^2s of trials 1 and 9 are the only ones of notable quality, whereas the other values are negative and indicate nonsense models. Moreover, it is of relevance to explore why cases 1 and 9 have such comparatively good values. The explanation is that the two artificial response variables are rather strongly correlated with the response of the parent model, with correlation coefficients of 0.89 ($Y1$, case 1), 0.92 ($Y1$, case 9), 0.65 ($Y2$, case 1) and 0.73 ($Y2$, case 9). Thus, the data structure of the synthetic variables 1 and 9 considerably resemble the systematic variation in the observed log ks, and consequently, the nine structural descriptors utilized are able to encode the dominant features in the random variables. In summary, the result of Tools 1−3 is, thus, compelling evidence that the real QSAR was well founded.

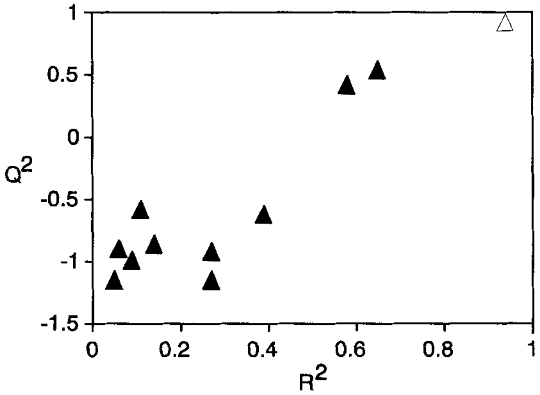


Figure 2. Scatter plot of the recorded Q^2s and R^2s for the epoxide example. The open triangle corresponds to the "real" QSAR model.

5.1.4.2 The Haloalkane Example

In contrast to the epoxide example, this illustration is a case in which both the training set and the validation set have been generated by means of statistical experimental design. All the details pertaining to this example can be found elsewhere in this volume [10]. In this instance, we will only show that five chemical descriptors reflecting hydrophobicity and molecular size, were used to parameterize the properties of the 16 tested halogenated aliphatics. These chemicals were tested for their cytotoxic potential and the endpoint determined was the inhibitory concentration, which lowered cell viability by 50% (IC_{50}).

The PLS modeling based on the ten training set compounds and with the aim to establish the parent QSAR, gave a model with R^2 equal to 0.90 and Q^2 equal to 0.88 (according to CV, Tool 2). In order to externally validate (Tool 4) the predictive behavior of this QSAR, the cytotoxicities of the six validation set compounds were predicted and compared with the experimentally determined IC_{50} values. In Fig. 3, a scatter plot, representing the agreement between the observed and calculated/predicted cytotoxicities, is shown. Obviously, Tool 4 shows that the QSAR is able to predict the biological activities of the validation set compounds in a sound way.

Furthermore, the quality of this QSAR was tested using Tool 1. Analogously to the above example, ten repetitive randomizations of the response data were carried out. The PLS modeling, treating the simulated response variables one by one, yielded the explained and predicted variances printed in Table 1 and those plotted in Fig. 4. Interestingly, two cases (3 and 6) occur where there were rather high scores of R^2s and Q^2s. This can be understood differently when considering the underlying correlation structure between the original response and the two artificial constructs. The correlation coefficients in question amount to 0.6 (Case 3) and 0.4 (Case 6),

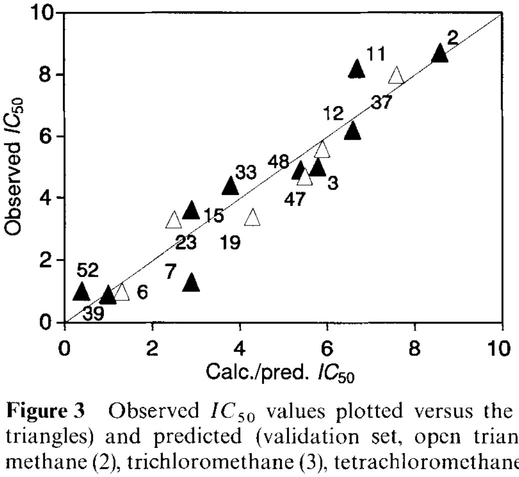


Figure 3 Observed IC_{50} values plotted versus the corresponding calculated (training set, solid triangles) and predicted (validation set, open triangles) values. The compounds are: dichloromethane (2), trichloromethane (3), tetrachloromethane (6), fluoro-trichloromethane (7), 1,2-dichloroethane (11), 1-bromo – 2-chloroethane (12), 1,1,2,2-tetrachloroethane (15), 1,2-dibromoethane (19), 1,2,3-trichloropropane (23), bromo-ethane (30), 1,1-dibromoethane (33), bromochloromethane (37), fluorotribromo-methane (39), 1-chloropropane (47), 2-chloropropane (48), 1-bromobutane (52).

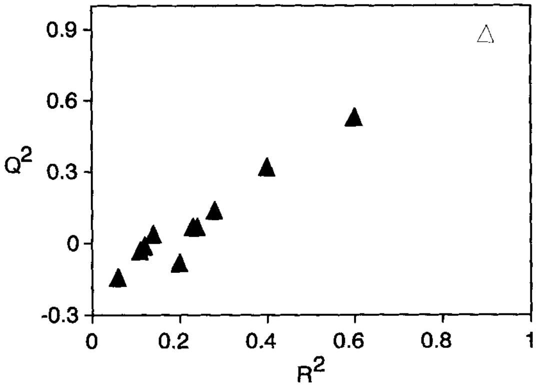


Figure 4 Correlation plot of the recorded Q^2s and R^2s for the haloalkane example. The open triangle corresponds to the "real" QSAR model.

and hence Cases 3 and 6 have comparatively much in common with the measured response variable. Apart from these two cases, no other trial has produced R^2s and Q^2s in the proximity of the corresponding values of the real QSAR. Thus, the conclusion is that Tools 1, 2 and 4 all point to the same conclusion and indicate the sound predictive capability of the haloalkane QSAR.

5.1.5 Concluding Remarks

A QSAR model should, in general, be viewed with caution until its validity and predictive power has been properly assessed. As discussed and illustrated above, a number of alternative procedures exist for such purposes. Ideally, the four tools mentioned should not be used in isolation, but rather in combination with each other, due to their complementary character. The absolute minimum requirement, when developing QSAR, is to test the validity with the randomization technique (Tool 1), but CV (Tool 2) ought to be carried out as well. However, the external validation (Tool 4), with a designed validation set, clearly produces the most trustworthy result.

To some people our emphasis on model validation may seem to be an overly cautious attitude, but in QSAR modeling, this is far better than being totally accepting of the results. Usually, it is objected that the purpose of developing a QSAR is to lay the ground for a better understanding of the mechanisms of biological action, and not prediction or optimization. However, a QSAR model that cannot predict better than chance is a poor basis for comprehending relationships between chemical and biological properties.

References

[1] Turner, L., Choplin, F., Dugard, P., Hermens, J., Jaeckh, R., Marsmann, M., and Roberts, D., *Toxicol. In Vitro* **1**, 143–171 (1987)

[2] Hermens, J. L. M. , *Quantitative Structure-Activity Relationships of Environmental Pollutants.* In: *The Handbook of Environmental Chemistry*, Hutzinger, O., ed., Springer Verlag, Berlin (1989) p. 111–162

[3] Blum, D. J. W., and Speece, R. E., *Environ. Sci. Technol.* **24**, 284–293 (1990)

[4] Topliss, J. G., and Edwards, R. P., *J. Med. Chem.* **22**, 1238–1244 (1979)

[5] Wold, S., and Dunn III, W. J., *J. Chem. Inf. Comp. Sci.* **23**, 6–13 (1983)

[6] Wold, S., *PLS for Multivariate Linear Modelling*. Chap. 4.4, this volume

[7] Cramer, R. D., Bunce, J. D., Paterson, D. E., and Frank, I. E., *Quant. Struct.-Act. Relat.* **7**, 18–25 (1988)

[8] Clark, M., and Cramer, R. D., *Quant.Struct.-Act. Relat* **12**, 137–145 (1993)

[9] Eriksson, L., Verhaar, H. J. M., and Hermens, J. L. M., *Environ. Toxicol. Chem.* **13**, 683–691 (1994)

[10] Sjöström, M., and Eriksson, L., *Applications of Statistical Experimental Design and PLS modeling in QSAR*. Chap. 3.2, this volume

[11] Stone, M., *J. Roy. Stat. Soc.* **B36**, 111–147 (1974)

[12] Wold, S., *Technometrics* **20**, 397–405 (1978)

[13] Wold, S., *Quant. Struct.-Act. Relat.* **10**, 191–193 (1991)

[14] Shao, J., *J. Amer. Stat. Assoc.* **88**, 486–494 (1993)

[15] *SIMCA*, Umetri AB, Umeå, Sweden

5.2 How to Choose the Proper Statistical Method

Sergio Clementi and Svante Wold

Abbreviations

ACC	Auto and cross covariance
ACE	Alternating conditional expectations
ALS	Adaptive least squares
CFA	Correspondence factor analysis
CoMFA	Comparative molecular field analysis
DOF	Degrees of freedom
FD	Factorial design
FFD	Fractional factorial design
GA	Genetic algorithm
GOLPE	Generating optimal linear PLS estimations
IVS	Interactive variable selection
LDA	Linear discriminant analysis
LOO	Leave-one-out
LOT	Level of triviality
MLR	Multiple linear regression
NLM	Non-Linear mapping
NN	Neural networks
OLS	Ordinary least squares
PCA	Principal components analysis
PCR	Principal components regression
PLS	Partial least squares
PPs	Principal properties
PRESS	Predictive residual sum of squares
RR	Ridge regression
RSD	Residual standard deviation
QPLS	Quadratic partial least squares
QSAR	Quantitative structure-activity relationship
SAMPLS	Sample distance partial least squares
SDEP	Standard deviation of error of predictions
SIMCA	Soft independent modeling of class analogy
SMA	Spectral mapping analysis
SPLS	Spline partial least squares
SSY	Sum of squares of response value
VIP	Variable influence on the predictions
VSS	Variable subset selection

5.2.1 Introduction

We all remember that, having got lost in the wonderland, little Alice met the Cat. She asked him: "Would you tell me, please, which way I ought to go from here?" „That depends a good deal on where you want to got to", said the Cat. "I don't much care where – ", said Alice. "Then it doesn't matter which way you go", said the Cat. " – so long as I get somewhere", Alice added as an explanation. „Oh, you are sure to do that if you only walk long enough", concluded the Cat [1]. This little excerpt stresses the point that, in order to answer properly the question which is the title of this chapter, one should clearly specify the goals of a QSAR study, otherwise any kind of statistics could be used.

The purpose of developing a QSAR for a given problem is that it gives us information on how changes in the structure of the actual compounds influence their biological activity. This, in turn, allows us to modify the structure in order to improve the biological response and to improve our understanding of the actual biological mechanism [2]. In other words, there are two main objectives in QSAR studies: interpretation, i.e. understanding which structural features affect the response, and prediction, i.e. estimating the activity of new compounds before they become available. The requirements of a chemometric tool, aimed at meeting these objectives, have been described and updated several times over the past ten years [2–8].

The variety of chemometric methods reviewed in this, and in the other books of the series [9, 10] may be daunting for the QSAR enduser or the newcomer, who wishes to select the proper method for his or her own problem. However, we still wish to encourage such people: the statistical method used is not the most important step in solving a complex problem. A correct problem formulation and the fact that the collected data do contain information relevant to the problem itself are, by far, more important.

The idea that in QSAR research "statistics frequently become merely a tool to confirm hypotheses, and is not used as a language to describe phenomena" was also pointed out by Benigni and Giuliani [11]. They also claimed that often "researchers apply statistics in an exclusively procedural way, as a set of formalized rules aimed at obtaining a reliable result, which generates a real cult for statistical indices. Many researchers feel very comfortable with statistical software packages specifically designed to obtain quick and reliable correlations, while the deep involvement of statistics with QSAR requires that researchers use it in a very active and conscious way".

We agree entirely with these statements and also with their conclusion that "a greater importance should be given to descriptive analysis", namely principal components analysis (PCA), "instead of dwelling almost entirely on statistical significance". However, we cannot agree with their other conclusion that "the present use of statistics does not help to understand the role of chemical parameters in biological activity". If this were true, we could attribute possible drawbacks, not to the statistical tool, but to an inappropriate problem formulation.

In fact there are two souls in chemometrics, which can be referred to as the multimethod and monomethod philosophies [12]. The first one was developed mainly by analytical chemists and led to software packages which contain a variety of multivariate statistical methods: it is recommended that several methods are used in order to find confirmation of the findings. The second one is based on a physical organic chemistry background and uses, within a unique framework, projection methods, such as PCA and PLS, plus design criteria: the SIMCA (soft independent modeling of class analogy) philosophy. In principle, choosing a philosophy is like choosing a religion as one is usually happy with the religion that he was born into. Similarly, one is usually content with the most familiar method. However, we believe that working with a single method is simpler for the newcomer and permits a better understanding of this unique chemometric tool when known in great detail.

Furthermore, all the methods of multivariate statistics are based on one of the two similarity criteria, either the Euclidean distance or the mathematical models [13]. The former can only be used to look at molecules in the light of their descriptors, but they are not aimed at discovering any cause-effect relationship; consequently these methods, particularly LDA (linear discriminant analysis), SMA (spectral mapping analysis), NLM (non-linear mapping) or CFA (correspondence factor analysis), which is similar to PCA, can only be aimed at classification studies. On the contrary, QSAR studies need a chemometric method aimed at finding the quantitative relation between activity and structural descriptors: these are called regression methods and among these we prefer PLS. Mixing together methods coming from different criteria, e.g. using principal components scores for a discriminant analysis, leads to a logical stepwise modification in the problem formulation. The combined methods, PCA/PLS, seem to be particularly appropriate, both for the exploratory analysis of the structural data and for establishing the quantitative relationship in the same descriptor space.

5.2.2 Problem Formulation

In the QSAR literature, the statistical models which are generally reported, express the biological response in terms of a few structural parameters, usually the traditional substituent constants. The apparent goodness of the model is derived thereafter by checking the goodness of fit of the simple linear regression between calculated and experimental y values. The QSAR equations usually contain only some of the available analogy factors and have the form of an ordinary multiple regression model, either linear or with some squared terms. However, the way in which these equations have been derived is seldom discussed.

Therefore, it seems appropriate to illustrate briefly the questions that should be answered by step, with a definition of the procedure which leads to informationally sound QSAR models.

5.2.2.1 Parameter Selection

The most important question in a QSAR study, is "which parameters, i.e. which structural factors, do really affect the response?" To answer this question, we should first define which parameters are to be used. The way of detecting the important ones will be dealt with in the next section.

In the traditional QSAR approach (Hansch analysis, see vol. 1 [9]), one uses substituent parameters which measure the substituent behaviour in some reference reactions or systems. Therefore, they are "analogy" parameters and their use implies the implicit assumption that the substituent effect is somehow proportional in the system under investigation. Since the number of such parameters, proposed so far, is fairly large, and continues to increase [14], how can we select the parameters to be used?

Moreover, since QSAR models are likely to be non-linear, most equations contain some descriptors expressed to the second power, and even reciprocal values have been used. However, the cross-terms, indicating the interaction between two distinct effects, seem to be far less popular. Therefore, how many terms should be taken into account initially in the regression equation before starting any variable choice? The underlying assumptions involve our desire to describe the activity data either in terms of a linear model, or in terms of a response surface.

Although traditional descriptors are numerous, PCA provides a tool for grouping systematic patterns of behavior into a few orthogonal scales. In fact, it is possible to apply the strategy of experimental design [15] also to discrete systems, provided that these are multivariately characterized by a principal components analysis of some selected data. The latent variables obtained as statistical scores are called principal properties (PPs) and represent in an appropriate way each system by few (usually three) "constants", which condense the systematic behavior of the original data. These have been applied to describe amino acids (AAs) in peptides, [16–17] or aromatic substituents in general organic series [14, 18]. A chapter in this series is devoted entirely to this topic [19].

A QSAR table is then prepared, describing each amino acid in a peptide sequence, or each substituent in a polysubstituted organic skeleton, by their PPs in triplets or pairs: the descriptor matrix. This table is then submitted to the chemometric analysis, in order to find out the relationship between the y vector, or the Y matrix, and the descriptor, X matrix.

5.2.2.2 Regression Methods

When talking about regression methods, one should first divide them into two main groups depending on the underlying assumptions. On the one hand, one should group together multiple linear regression (MLR), also called ordinary least squares (OLS), adaptive least squares (ALS), ridge regression (RR), variable subset selection (VSS) and other stepwise methods, where the underlying assumption is the independence between variables. Accordingly, they are only appropriate when there are few descriptors, many compounds, and no variable selection is attempted. On the other hand there are the projection methods, namely partial least squares (PLS), also called projection to latent structures, and principal components regression (PCR), where the underlying assumption is that there are few "underlying" latent variables.

Among the methods used in QSAR, [4, 7, 8, 20] we shall restrict our selves here to the two more commonly used methods: MLR and PLS. It was only a few years ago [3] that the conditions for the applicability of statistical methods to QSAR were first reviewed. The most important condition to be fulfilled was based on the simple fact that for a model to have predictive value, the number of estimated parameters, P, should be appreciably smaller than the number of degrees of freedom in the data set (DOF). Accordingly, one can define a level of triviality (LOT) as the point when P = DOF. At this point and beyond, ($P >$ DOF), the predictions of the model, with the calculated parameters, are no better than random, even if the fit of the model to the training set looks good. Thus, such results are spurious, trivial and fortuitous.

The comparison between MLR and PLS has been reported several times [4, 7, 8, 21]. There are at least three weak points in using MLR in the area of QSAR: (a) the number of objects should be much larger than the number of variables; (b) MLR is based on the assumption that each variable is important for the problem, in other words, the model dimensionality is fixed a priori; (c) the regression coefficients become unreliable if there are significant correlations among the descriptor variables, multicollinearity. On the contrary, in PLS the ratio between variables and objects is not limited, the relevance of individual variables results from the analysis and their correlations are just used to find out the numerical solution. Consequently, if the question is "which parameters do really affect the response?", the most appropriate answer seems to be obtainable by PLS.

Furthermore, the results of PLS are usually presented as plots, so that groupings, if any, are easily detected, whereas in MLR, in which the results are usually given only numerically, this is never observed. Since substituents are grouped [22], MLR results often appear to be deceptively excellent, as a result of the dramatic decrease in the real degrees of freedom due to the groupings not being taken into account. On the other hand, the traditional stepwise procedure [23] cannot strictly speaking be considered as a multivariate approach, as it does not take into account the interactions among substituents.

Finally, it is appropriate to point out that a recent statistical report claims ridge regression (RR) is the method which gives the best predictions [20]. However, it seems reasonable to presume that the assumptions underlying PLS are much closer to the requirements of a real problem formulation in QSAR [8]. Also, a simple refinement of the PLS method gives as good, or better predictions than RR [65].

The last question, regarding regression methods, takes us back to the problem left open in the previous section, i.e. the choice of the parameters to be used in the equation. Should we start from all possible descriptors, or should we consider only a few of them, and in this latter case, how should we choose them?

A further point should be made in that the obvious way for finding out the "best subset of variables", giving the best model in MLR, is to test all possible variable combinations. However, this may easily become impractical due to the increase in the number of variables. Consequently, two alternative procedures are commonly used, namely, forward selection and backwards elimination [24, 25]. However, we have examples [26], which shows that the selected variables are different depending upon the method chosen. In other words, there might be several equations which give almost the same predictivity, based on quite different groupings of structural descriptors. Since different variables, which are retained in the model, have a different interpretation, it is clear that such methods cannot be accepted as a reliable tool in order to understand/interpret QSAR equations. Unfortunately, it seems that quite a number of scientists, involved in QSAR studies, find their best equation as end-users of some "general-purpose" statistical package, which is numerically based only in terms of goodness of fit, and pays no attention to the problem formulation, model predictivity, interpretation of the results, plots showing homogeneity of the data set, and in general, to any method derived for understanding why things happen.

Accordingly, since PLS has no limitations concerning the number of variables to be used from the beginning, we would suggest that all possible variables should be taken into account, and then to allow the chemometric method to select only the important variables. In this context, the description using PPs seems to be particularly suitable. Proposals on how to carry out variable selection are given in Sect. 5.2.3.2. However, it seems that most equations reported in the literature, which contain only few traditional descriptors, have been simply derived by forward selection.

5.2.2.3 Model Evaluation

Only if a QSAR model were valid, may we use the model with its parameter values to predict what would happen, when the factors were changed. However, how do we judge that a QSAR model is valid? Using ordinary regression or PLS regression, we can calculate values of the variable parameters, coefficients or loadings, in such a way that the residuals are small. A measure of the size of the residuals is given by the residual standard deviation (s or RSD). Likewise, R^2, the multiple correlation coefficient, measures the "explained" y variance.

Therefore, the first necessary condition for model validity is that R^2 is close to 1.0, and s is small. However, a large R^2 and a small s are not sufficient for model validity due to the unfortunate property of regression models to give a closer fit, the larger the number of parameters and terms in the model. And, what is even worse, if we have many structure descriptor variables to select from, we can make a model fit data very closely, even with few terms, provided that they are selected according to their apparent contribution to the fit. This is true, even if the variables we choose from, are completely random and have nothing whatsoever to do with the problem being investigated! This is one reason why stepwise regression is impractical with data sets containing many collinear predictor variables.

Although the risk for these chance correlations with variable selection has been pointed out [3, 27], it seems that this risk is not sufficiently recognized by the chemical and biological communities. The big problem with chance correlations is that predictions for new compounds of such models are very poor: the model fits the training set data well, but is useless for predicting and understanding.

In order to evaluate the validity of a model, the best approach would be to have a fairly large and representative validation set of compounds, for which the predicted activity values can be compared with the actual values. In the absence of a real validation set, we can use a simulated one, since recent developments in statistics provide us with a new interesting set of measures of validity that are based on simulating the self-consistent predictive power of a model. Nowadays, cross-validation and bootstrapping [28, 29] constitute the basis of the modern statistical philosophy of "replacing standard assumptions about the data with massive calculations", for assessing the generality of a relationship found from a sample data set [30]. These tools operate by creating a number of slight modifications of the original data set, estimating parameters from each of these modified data sets, and then calculating the variability of the predictions by each of the resulting models.

In cross-validation, the data set is divided into a number of groups. The model, of a given complexity, is fitted to the data set, reduced by one of the groups. Predictions are calculated with the fitted model for the deleted data and the sum of squares of predicted minus observed values for the deleted data is formed. Then, in the second round, the same procedure is repeated, but with the second group left out. Then a third round is performed, etc., until each data point has been left out once only. The total sum of squares of predictions minus observations then contains one term for each point. The sum, abbreviated *PRESS*, is a measure of the predictive power of the model of a given complexity for the given data set. In the end *PRESS* (Predictive REsidual Sum of Squares or PRedicition Error Sum of Squares) will contain one contribution from each observation.

PRESS is a good estimate of the real prediction error of the model, provided that the observations (compounds) are independent. If *PRESS* is smaller than the sum of squares of the response values (*SSY*), the model predicts better than chance and can be considered to be statistically significant. In a reasonable QSAR model, *PRESS*/*SSY* should be smaller that 0.4, whereas a value smaller than 0.1 for this ratio indicates an excellent model. If the *PRESS* value is transformed in a dimensionless term by relating it to the initial sum of squares, one obtains Q^2, i.e. the complement to the fraction of unexplained variance over the total variance ($Q^2 = 1 - PRESS/SSY$). *PRESS* and Q^2 have good properties which render them appropriate for statistical testing with critical distributions.

To sum up, a model can be considered reliable when (a) *PRESS* has been calculated and *PRESS*/*SSY* is lower than 0.4, (b) there are plots of the data patterns, and (c) a clear description is given of the candidate set of variables and of the variable selection procedure, if applied. Someone may object in that the purpose of developing a QSAR is to achieve a better understanding, not for prediction or optimization. However, a model that cannot predict better than chance is a really poor basis for understanding chemical-biological interactions.

5.2.3 The SIMCA Philosophy

The SIMCA philosophy is based on three main methods: PCA, PLS and design, and all these three topics have been covered in more detail in previous chapters. Its peculiar characteristics, which make it particularly suitable for QSAR modeling, will only be briefly illustrated here in order to suggest an overall chemometric strategy for molecular design studies. This stategy relies on two major steps: design in latent variables and PLS modeling, the latter being refined by taking into account validation and variable selection. The PLS algorithm, already discussed in Sec. 5.2.2.2, is the most appropriate tool for establishing quantitative relationships between a biological activity vector and a matrix of structural descriptors, and it has illustrated its capability in detecting the structural features, which affect the biological activity as well as in providing reliable predictions [4, 31].

PCA is only used for classification purposes, and gives reliable results in terms of confidence values, also in QSAR studies, where the objective of discriminating active from inactive compounds cannot, in principle, be obtained by methods, based on the Euclidean distance. In fact, while active compounds can be described by a statistical model and, therefore, constitute a homogeneous class, inactivity may have originated due to the lack of any of several different structural features and, therefore, cannot form another separated homogeneous class. On the contrary inactives are spread all over the descriptor space, thus, defining an asymmetric problem [32]. Even the use of PLS as discriminant function, which is sometimes used here [33], is not to be recommended in this context.

5.2.3.1 Factorial and D-Optimal Designs in PPs

In order to develop sound QSARs, it is essential that the chemical compounds, on which the model is to be based, are selected by a design technique. Only when such requirements are fulfilled, will QSARs permit sound predictions for other molecules. Design means a computer-assisted strategy, which is able to span the operational space in the best possible way. The operational space is the space containing the object under control, and is described by numerical values. These object descriptors define the operational space, which is usually called variable space.

The importance of design has not yet been fully recognized: new structures are usually derived on changing one substituent at a time for each substitution site. Sets of molecules, obtained in this way, do not contain enough information for ranking the importance of individual features which affect biological activity and for providing stable models to be used in predictions. The message that there are strategies and tools to handle complex data has not yet reached all research teams. Reliable models can be obtained only by a designed training set, or with available data, by a well balanced data set, containing structures selected by a design strategy in the latent variables space, which has been derived from raw data, contained structural descriptors for all the available compounds [16, 34].

Experimental designs provide a strategy for selecting the few most informative molecular structures in a homologous series. In fact, it is also possible to apply the strategy of experimental design to discrete systems after a multivariate characterization of some selected data, which generate PPs. The strategy of fractional factorial designs can be applied afterwards by (a) using blocks of three PPs for defining each item at each site to be varied and (b) selecting a representative item for each position of the PP space.

Factorial designs (FDs) and fractional factorial designs (FFDs) are simple, straightforward, and, therefore, good in facilitating an understanding of the concept of design by spanning the variable space. The set of possible substituents is divided into subsets according to their relative position in the PP space. A substituent, representing each subspace, (quadrants or octants) is selected thereafter, and labeled by the pair or triplet of signs corresponding to that subspace. Each "experiment" of a factorial design matrix can be transformed into a molecule, if we assign a pair of triplet of columns of the sign matrix to define the substituent corresponding to each site.

However, with this approach a polysubstituted molecule should bear as many substituents as many substitution sites. Therefore, the FD approach might not be easy to apply when a synthetic chemists wants to keep under control a number of different substitution sites at the same time. It is clear that FDs give a synthetic plan in which the most informative compounds are difficult to synthesize, because they contain too many substituents. Because of these reasons we later investigated [6, 35] the effect of using D-optimal designs instead of FDs with PPs in QSAR, since D-optimal designs can be used as a general alternative to FDs in constrained situations, e.g. when some regions of the variable space are excluded, or when the data set is discrete, as with molecular structures.

The D-optimality criterion to evaluate the goodness of experimental designs has been dealt with by several authors [15]. It consists of determining the n experiments which minimize the volume of the ellipsoid of the conficence intervals of the estimated parameters for the coefficients in a multiple linear regression equation. An experimental design with n-points is D-optimal if the value of its determinant is maximum compared with all the other possible designs with n-points, which can be constructed in the experimental domain. Since the number of coefficients to be computed is equal to the number of the variables plus one, the experimental design should contain an equal or greater number of points. In particular, the Mitchell algorithm [36], which we selected [35], works by starting from an arbitrary initial design, and adding one point of the experimental domain to the starting design in such a way as to increase as much as possible the value of the determinant.

FDs are D-optimal when each substitution site is controlled by a single parameter. On using two PPs, the goddness of D-optimal and factorial designs is comparable. However, on increasing the number of PPs for describing each substitution site, the efficiency of D-optimal designs increases much more. Suppose that the problem formulation is one of 6 variables generated by controlling 2 sites by triplets of PPs, the total number of possible molecules, allowing our weight selected substituents for each site, is $8^2 = 64$. The FFD approach would first generate a design matrix with 8 rows and 6 columns and then assign substituents to the two triplets of signs according to the subspace codes.

On the contrary, the D-optimality approach works in a 6-dimensional space with the actual values of the PPs. The D-optimality criterion then results in the selection of the seven or more points out of the 64, which meet the requirement of maximizing the quoted determinant, i.e. roughly by spanning the domain in the 6-dimensional space in the best possible way. It is not appropriate to include all possible substituents in the D-optimality search, because of the large number of possible candidates.

Consequently, the use of the substituents, which are representative for each octant according to FD theory, is highly recommended for reducing the number of candidates to eight to the power of the number of sites.

The advantages of using D-optimal designs instead of fractional factorial designs in principal properties can be summarized as follows. It is possible (a) to reduce the number of required structures; (b) to reduce polysubstitution, and even controlling several sites; (c) to exclude molecules, which are too difficult to synthesize; (d) to include molecules, which are already available and/or have been tested.

5.2.3.2 Validation and Variable Selection

In sect. 5.2.2.3 the importance of validating regression models according to their predictivity has been illustrated, and a whole chapter of this volume is devoted to this topic [37]. When the relevance of individual variables has been derived by models, obtained in the usual way (fitting), the physical meaning of the results might be misleading, if the model happens to be anchored to points of much higher or much lower activity. Only when a variable has been proved to be useful in increasing model predictivity, can it also be judged to be relevant to the response and, therefore, to be used for interpreting the relationship from a chemical viewpoint.

In cross-validation, *PRESS* values are calculated for different subgroups of the training set, until each object has been withdrawn and predicted once, and the total *PRESS* is formed by summing all partial *PRESS* values [2]. Nevertheless, for practical reasons in various branches of chemistry, the use of the square root of *PRESS/N* seems to be more directly related to the uncertainty of the predictions, since it has the same units as the actual y values. Accordingly, we suggested [31, 38] that the term *SDEP* (Standard Deviation of Error of Predictions: $SDEP = (PRESS/N)^{1/2}$) be used.

This equation, however, has yet to define a unique way of computing the parameter *SDEP*, since the way the predictions are made should also be selected. For example, in defining the cross-validation procedure, the data set should be divided into a number of groups, but one can also increase the number of groups until it equals the number of data points, thus obtaining the leave-one-out (LOO) procedure.

LOO should theoretically be the best approach provided data are randomly distributed or designed, but LOO gives *SDEP* values lower than the approach using groups, when data are clustered. Since in QSAR the descriptor variables usually generate grouped data, owing to the discrete nature of substituents at the various substitution sites, the prediction capability of a model should be evaluated in a non-favorable cross-validation technique, i.e. by the formation of the lowest reasonable number of groups.

Moreover, one should not just be satisfied with using a cross-validation technique that forms groups in one particular way, and we computed *SDEP* several times on groups formed in a random way [38]. This definition of *SDEP* places it halfway between cross-validation and bootstrapping. In fact, the computation was repeated several times, as in bootstrapping, but each point was excluded just once in each run, as in cross-validation. We showed that the higher the number of random pathways of forming groups, the more stable is the *SDEP* value, which is to be regarded as the mean value of the individual "*sdep*" values obtained by each computation.

The *SDEP* parameter can be logically associated with the uncertainty of any new prediction made by that model. However, being dependent on the parameter scale, it is obvious that the "absolute" prediction capability of the model should be evaluated by Q^2.

Procedures for variable selection have long been used with ordinary least squares regression methods [24]. However, almost all previous work in variable selection was undertaken exclusively for describing data sets (fitting), and it was shown in Sec. 5.2.2.3 that all regression models increase their fitting capability with increasing number of variables. In order to evaluate the relevance of individual variables in validated regression models, investigation of model predictivity is, therefore, needed.

By means of the *SDEP* parameter, it was possible to compare the prediction capability of different regression methods [38], or to select groups of variables, capable of giving the best prediction capability of a single model [39]. In the latter, we suggested a preliminary outline of a procedure called GOLPE (Generating Optimal Linear PLS Estimations), aimed at obtaining the best predictive PLS models, which allowed us to show that: (a) in PLS modeling all variables are relevant for fitting but some of them may be detrimental to predictivity; (b) the GOLPE procedure is a method for detecting variables, which increase predictivity; (c) the PLS models, obtained by using only variables selected by GOLPE, are more predictive than the PLS model obtained by using all variables; and (d) PLS models with variable selection are more predictive than similar models [26] obtained by ordinary least squares.

The procedure was based on statistical designs, as design matrices used in fractional factorial designs (FFDs) are a suitable tool for finding an efficient way of selecting the best combination of variables [15]. The strategy was developed by using combinations of variables according to a FFD, where each of the two levels (1, −1) corresponded to the presence and absence of the variable. The design matrix, including only the "plus" and excluding the "minus" variables, suggested that only the prediction capability of these reduced models should be tested. Each model had a different combination of variables and was all in all a good presentation of all the possible combinations. For each such combination, the prediction capability of the corresponding PLS model could be evaluated by means of *SDEP*. Accordingly, a response vector was obtained, indicating the model predictivity for each combination of variables as the lowest *SDEP* value corresponding to the dimensionality for which *SDEP* assumes the minimum value.

Variable selection procedures can find apparent good models that do not give reliable predictions. This means that the presence of chance correlations may mask the true effects of individual variables [2, 40]. In fact, when the objects/variables ratio is far smaller than unity, the biased regression methods may also fail in terms of predictions. In order to avoid the risk of chance correlations, one should provide some general rules for obtaining a reliable variable selection. These criteria should take into account the ratio between variables and objects, the existence of some structure in the data, and the presence of some initial predictivity, e.g. *PRESS* being smaller that *SSY*. Only when at least one of these criteria, namely the latter, is met, should the variable selection procedure be allowed and its results should give much better predictions.

In order to estimate as precisely as possible the significance of a single variable effect on predictivity, the GOLPE procedure was later refined by introducing a number of dummy variables into the design matrix [41]. These dummy variables were not actual numbers and we labelled some columns in the design matrix (say one out of four) as dummy (or ghost) variables, which were inserted among the true ones. These dummy variables were not involved in the variable combinations which evaluate the predictivity of each row of the design matrix. They are only used to compute the apparent effects on predictivity given by a non existent variable by means of the Yates algorithm, so that a decision on the positive or the negative effects of individual true variables can be taken on the basis of a Student-*t* tailoring. Variables with a positive effect on predictivity can be fixed within the variable combinations, while variables with a negative effect on predictivity can always be excluded from the variable combinations. If variable selection proceeds in an iterative manner, it increases the stability of the results, thus furnishing the complete list of selected variables.

The GOLPE procedure appears, therefore, to be a powerful and efficient tool for variable selection. However, we should note that it can only be properly applied provided that the regression model on the whole data set have at least some initial predictive ability (Q^2 greater than $0.1 - 0.3$). In such cases GOLPE can take away the noise and improve considerably the Q^2 value. If this is not the case, variable selection can still be allowed, provided that there is some structure in the X data, e.g. design has been used, implying that the dimensionality of the problem is lower than the number of variables.

An independent measure of the relative importance of the x variables can be calculated as *VIP* (variable influence on the predictions). *VIP* is derived from the PLS weights, taking into account the fraction of variance explained in each model dimension [7]. In addition to *VIP*, the regression coefficients are also useful for assessing the importance of x variables: only those with b values larger than about half the maximum b value are seen to be important. A further development of VIP led to the proposal of the Interactive Variable Selection (IVS) method for PLS studies [65]. There is still no comparison between the selection of important x variables by GOLPE and by *VIP* and b values or by IVS.

Selecting variables according to both their *VIP* and b values gives good results, provided that some caution is taken, since selecting variables is difficult and risky. In order to avoid pruning, the elimination of variables should be undertaken to simplify the model, and not be influenced by the degree of fit or the prediction error. The latter usually leads to a partly spurious model that overfits the data considerably.

According to a chapter in one of the previous volumes in this series [42], variable selection was performed by an iterative procedure, based on the cross-validated R^2 of PLS models. At each step, the amount of information carried by each variable was assessed by its standardized regression coefficient, and the elimination of the variable with the lowest coefficient improved the model. This improvement was again quantitatively estimated by cross-validation, and was shown to go through a maximum, after which any further elimination caused a decrease in R^2_{cv}, so that the iterative procedure was stopped. The reliability of this procedure sounds a little doubtful, both because of the risk of pruning (cross-validation is made in leave-one-out (LOO), and because the relative ranking of the regression coefficients is considerably dependent upon the number of significant dimensions of the PLS model.

Alternative strategies, which have been suggested for variable selection, include genetic algorithms [43, 44] and forward/backward methods [25]. However, we should warn against evaluating predictive performance by LOO. It has been claimed that ordinary regression, using reduced models, obtained by such selection techniques, behaves better than the biased regression methods, RR and PLS [26]. Here it is not difficult to illustrate that PLS behaves even better with selected variables [39]. It is striking that the two methods of variable selection, forward selection and backwards elimination, selected totally different groups of variables, thus casting serious doubts on the reliability of the interpretation of the final model. Once again, the attention given to predictivity overshadowed other aspects of the regression analysis.

5.2.4 Other PLS Codes

Non-linear variants of PLS modeling have been developed that are very similar to ordinary PLS models, except that they have a curved inner relation. Thus the y scores are modeled as a quadratic (QPLS) [45], or cubic polynomial or spline (SPLS) [46] in the corresponding x scores.

Alternatively, the use of neural networks (NN) [47] has been advocated for multivariate non-linear modeling. It is clear, however, that NNs, with their non-linear regression-like formalism, do not work with many variables and few cases. Therefore, some kind of variable reduction, preferably by projections, is appropriate and a PLS-projected version of NNs, which is very similar to non-linear PLS, was recently developed [48].

Other non-linear methods, suggested over the last few years, are alternating conditional expectations (ACE) [49–51] and some genetic algorithms (GA) [52]. However, the experience with these non-linear models in QSAR is still limited, and the benefits of non-linearity may not always compensate the drawbacks of more complicated estimation and interpretation.

In addition, non-linear models seem to have considerable problems with overfitting. In fact, non-linear models are likely to give a better fit to the training set data, but are unable to give better predictions. This was shown, at least for QPLS and ACE, with five QSAR data sets [31]: In principle, it should be obvious that the smoother the algorithm, the closer the model fits the available points, and optimizing predictions is not a main objective in this context. Accordingly, the interpretation, which is already somewhat obscure, is also greatly dependent upon the representativity of the training set compounds.

A few years ago a geologist, at a workshop on chemometrics in geochemistry, proposed the use of a method called Similarity Correspondence Analysis, which he wanted to call SIMCA: the time of cloning was upon us [53]! Nowadays we can see that PLS, having already taken some ten years to be recognized, has become popular and is accepted as a new, and perhaps better, regression method as compared to ordinary least squares. As a consequence, several people have written their own codes and started to use PLS for their QSAR studies. However, it seems to us that quite often the poor experience of such people in handling PLS models may lead to some misleading problem formulation, generating, in turn, some misunderstanding in the interpretation.

A notable example of such a behavior is given in the recently published paper on SAMPLS [54]. Here the authors, Bush and Nachbar, wrote their own new code in such a way that the PLS implementation was sample-based instead of property-based. SAMPLS reduced all explanatory data to the pairwise distances among samples (molecules), that can be subsequently used to fit the PLS components under cross-validation (LOO) conditions. They showed that SAMPLS exactly reproduced conventional PLS analyses, being by far faster.

Even if they are numerically correct, we should point out that transforming a molecular descriptor matrix into an intermolecular distance matrix is not appropriate at all, since all the information needed for interpretation is lost completely. In a distance matrix, rows and columns are equal and this makes the problem formulation less clear. Projections methods highlight the differences between samples because of the variables: if samples and variables are the same, there can be no way of formulating a relation from a chemical viewpoint, even if the equation is correct and execution of the model fast. Deliberately "SAMPLS does not calculate any statistical quantities related to the explanatory properties" [54], which is the real goal of QSAR.

It is not surprising that Bush and Nachbar [54] stated that using groups in cross-validation, instead of using a full LOO procedure, is a short cut needed for saving computational time: their method cannot work with groups. Furthermore, they quote another disparaging claim made by the authors of the chemometric system SPECTRE [55]: "the main disadvantage of the PLS method is that the latent variables are abstract and difficult to interpret". Therefore SAMPLS was claimed to be "ideally suited to structure-activity analysis based on CoMFA fields", that "expresses its predictions in terms of displacements between real chemical groups, e.g. halfway between cyclohexyl and phenyl" in order to avoid the use of latent variables, that are so abstract and complicated. We really have to admit how surprising it is that PLS is nowadays so widely and successfully used, even despite the fact that it is not thoroughly understood [20]. Much better than SAMPLS are the kernel algorithms for PLS, which were proposed both for Tall [56] and wide [65] x-matrices, since they involve the y-vector. The Kernel algorithm is fast and memory saving and still retains the total information carried by the variables, describing the structural features [56].

5.2.5 3D QSAR

In a CoMFA study [57], or on applying PLS to the energy field computed by GRID [58], the rigorous procedure suggested by GOLPE may be impractical, since the variables in 3D QSAR are in the order of hundreds or thousands. Therefore, a strategy providing a reduced number of variables from the beginning and, therefore, a reduced number of combinations, would be highly desirable.

In Sect. 5.2.3.1, we showed that D-optimal designs are more efficient than FFDs in constrained problems. Accordingly, in 3D QSAR, the D-optimality criterion may be used for a preliminary selection of variables in the loading space according to a D-optimal design. In fact, with so many variables, the information is largely redundant, and D-optimality is an appropriate criterion to select variables in such a way as to retain almost all the information by a much smaller number of variables, which are spread as much as possible in the principal components space.

This preselection ends up by taking away redundancy without destroying collinearities, since it is recommended that the D-optimality criterion is used to retain not less than a half of the variables at a time, in an iterative manner, and stops as soon as the model predictivity begins to change. Of course, this may not be the only, or the best way of reducing redundancy, but at present it seems to work quite satisfactorily. However, after cleaning the x-matrix by removing all small values, all variables at two levels with skewed distribution, and all variables with a small variance, the number of active variables becomes in the order of hundreds and the D-optimal preselection is not needed any more [66].

However, while we are trying to develop better chemometric tools and procedures for handling 3D QSARs, we can see several examples of poorly formulated problems. Typically, on the one hand, there is the risk of obtaining trivial results, as shown by a series of papers by Kim [59], who implemented PLS to determine obvious dependencies between CoMFA or GRID fields and the traditional analogy constants. 3D QSARs represent a highly appealing area for researchers, where one is dealing with really important problems, to which one can apply the most advanced computational tools. Their special features are (a) updating current QSAR studies in keeping with the Hansch trandition and (b) combining chemometrics and modeling techniques in order to develop a procedure that we would like to call $(mc)^2$, an acronym for Modeling and Chemometrics in Medicinal Chemistry.

There are very few examples of $(mc)^2$ outside of the 3D QSAR area. We are pleased to announce that the procedure undertaken by the research group of Pitea and Todeschini [42] represents another good example of this kind, where chemometric methods and information, derived from modeling, are used in logical sequence to solve real problems, although strictly speaking, it is not a QSAR method. Furthermore, although it seems to us that there are a couple of slightly weak points in the procedure, both in using a classification tool at the beginning in a fashion similar to the active analog approach for a case that ought to be asymmetric, and in the variable selection strategy, that we discussed earlier, we are pleased to see that the procedure has been successfully applied to solve several real problems.

The real drawbacks in 3D QSAR are different, as they are strictly linked to the continuity and congruency requirements of such models. Auto- and cross-correlation and covariance (ACC) transforms are suitable tools for recognizing the information contained in the 3D fields, generated by CoMFA in such a way that they appear to be more appropriate for 3D QSAR. This rearrangement provides new data that have two favorable properties: they take into account neighbor effects, and, therefore, the required continuity between grid nodes, and they are independent of alignment within the grid lattice [60].

Indeed, present day 3D QSAR models depend almost exclusively on the alignment criterion used, so that it is sometimes customary to realign molecules after the first analysis in order to improve the model. On the contrary, we would like to find a molecular description independent of the alignment, and this could be obtained, if we succeed in describing the molecule in a way that is independent of its location within the grid space. Using our ACC transformation, data descriptions derived by CoMFA or GRID can be modeled by PLS without any need for alignment, thus meeting both the congruency and continuity requirement. The 3D ACC transforms, developed so far, allow a unique and congruent description of "degenerate" numbering of molecules [60]. At present, we can only satisfactorily deal with planar molecules, but cannot yet properly describe flexible molecules. The improvements under investigation will hopefully help us to reach our other objectives.

Another aspect which we would like draw attention to is the relative importance that should be expected (or given) to the different fields in CoMFA, or to the different probe energies in GRID. In principle, we agreed [7, 41] that the row data should be blockscaled in order to give the same initial importance to each "fundamental" effect. Their relative importance is then derived from the results, often depending upon the absolute values of the fields, and discussed in order to interpret the biological mechanism and to design new parent molecules.

It might be the case that we would have to change our problem formulation, if we could dissect a ligand-receptor interaction into sequential steps, each depending upon specific properties: (a) first, the capability of crossing a membrane, presumably linked to some molecular hydrophobicity parameter, which produces the actual concentration in the cell; (b) second, the molecular recognition phase, which is presumably an electrostatic interaction across large distances, and driven, therefore, by molecular electrostatic potentials; (c) finally, the real binding, which is namely due to H-bonding and steric/lipophilic interactions. If this were true, assigning the same importance to all the aspects in the PLS analysis might not be the best choice.

5.2.6 Conclusions

We have tried to give an overall view of the problems concerning the title of this chapter, "how to choose the proper statistical method", highlighting either some philosophical aspects and some good or bad examples of applications of chemometric methods in molecular design.

Obviously, as we have already remarked, the SIMCA philosophy provides a unique framework of multivariate tools that seem to be particularly suitable for QSAR studies. Anyway, we focused on the importance of an appropriate problem formulation with respect to the statistical method used. Nevertheless, whatever the method, it should depend on design, validation, variable selection and inspection of plots, in order to obtain informationally sound QSAR models.

Design has already been discussed thoroughly in this volume [19, 61]. We would just like to emphasize once more that the common strategy of constructing a training set of compounds by changing one structural feature at a time does not work well, mainly because it does not provide information about the combined influence of all the varied structural elements which affect the biological activity. In contrast, multivariate statistical designs allow the selection of a training set of compounds that is informationally sound, i.e. that gives data with good predictive power. Furthermore, it seems appropriate that a QSAR study is carried out in two phases: the first one for a preliminary screening, typically by fractional factorials, and a second one with a finer control on substituents, which lie within the good subspaces, for the final response surface model, by means of composite design or D-optimal design. Quadratic models can be derived either by the CARSO procedure [62] or by QPLS [45].

PLS is based on the projection of the structural descriptor variables (X) down onto a low dimensional subspace simultaneously with the projection of the biological activity variables onto the same subspace. PLS is not based on assumptions of independence or exactness or relevance of the X variables, and is, therefore, suitable for the analysis of the typical QSAR data set with many variables in both X (the structural description) and Y (the biological activity), even when the number of investigated compounds is fairly small.

When cross-validation is used to estimate the prediction errors of a model, the cross-validation must start anew with each deleted group or deleted compound. Thus, it is wrong to use a stepwise selection of variables to develop a model for the whole training set and thereafter delete one compound at a time and reestimate the model with the reduced set of variables. The correct way is to leave one group of compounds out from the beginning, apply the variable selection and model development, based on the remaining data, and thereafter predict the left out compounds. Then, a second group of compounds is left out, a new variable selection and model development is undertaken, the left out compounds predicted, and so on. If it is not done in this way, the cross-validation gives a much too optimistic view of the predictive power of the final model.

It is appropriate to recall that some QSAR endusers have raised some doubts about the efficacy of cross-validation techniques. It is our pleasure to state that professional statisticians do believe that cross-validation is, nowadays, a totally reliable tool [20]. Moreover, it was recently claimed that better theoretical and practical results could be obtained with cross validation, when several samples are deleted together groupwise instead of one at a time [63]. There are only two situations when cross-validation does not work well. The first is when compounds are strongly grouped and, hence, not independent. The second situation occurs when cross-validation is applied after variable selection in stepwise multiple regression.

The final point was raised in the question about the capability of the PLS algorithm to determine the few variables which are really related to the response among a large number of noisy ones [64]. Although this was probably true for the original PLS algorithm, it is not true any more, if an appropriate validated procedure of variable selection is used [Clementi, S., unpublished; Wold, S., unpublished results]. Consequently, little Alice can really be helped to proceed securely in the "wonderland" of QSAR modeling, possibly with the aid of the SIMCA philosophy.

Acknowledgements

We would like to warmly thank the members of our research teams in Perugia (G. Cruciani, M. Baroni, D. Bonelli, G. Costantino, D. Riganelli and R. Valigi) and in Umeå (N. Kettaneh-Wold, L. Eriksson, E. Johansson, J. Jonsson, F. Lindgren, S. Rännar and M. Sjöström) for discussion and advice, whereas we gratefully acknowledge our national research agencies (MURST and CNR in Italy, NFR and CFM in Sweden) for the research grants which supported the work outlined in this chapter.

References

[1] Carrol, L., *Alice's Adventures in Wonderland*, 1865

[2] Wold, S., *Quant. Struct.-Act. Relat.* **10**, 191–193 (1991)

[3] Wold, S., and Dunn III, W. J., *J. Chem. Inf. Comput. Sci.* **23**, 6–13 (1983)

[4] Dunn III, W. J., and Wold, S., Pattern Recognition Techniques in Drug Design. In: *Comprehensive Medicinal Chemistry*, Vol. **4**, Hansch, C., Sammes, P. G., Taylor, J. B., and Ramsden, C. A., eds., Pergamon Press, Oxford (1990), p. 691–714

[5] Wold, S., Berntsson, P., Eriksson, L., Geladi, P., Hellberg, S., Johansson, E., Jonsson, J., Kettaneh-Wold, N., Lindgren, F., Rännar, S., Sandberg, M., and Sjöström, M., *Pharmacochem. Lib.* **16**, 15–24 (1991)

[6] Clementi, S., Cruciani, G., Baroni, M., and Skagerberg, B., *Pharmacochem. Lib.* **16**, 217–226 (1991)

[7] Wold, S., Johansson, E., and Cocchi, M., Partial Least Squares Projections to Latent Structures. In: *3D QSAR in Drug Design: Theory, Methods and Applications*, Kubinyi, H., ed., ESCOM, Leiden, 1993

[8] Wold, S., *Technometrics* **35**, 136–139 (1993)

[9] Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches*. Methods and Principles in Medicinal Chemistry, Vol. **1**, Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., eds., VCH, Weinheim, 1993

[10] van de Waterbeemd, H., ed., *Advanced Computer – Assisted Techniques in Drug Discovery*. In: Methods and Principles in Medicinal Chemistry, Vol. **3**, Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., eds., VCH, Weinheim, 1995

[11] Benigni, R., and Giuliani, A., *Quant. Struct.-Act. Relat.* **10**, 99–100 (1991)

[12] Esbensen, K. H., *Chemom. Intell. Lab. Syst.* **7**, 199–202 (1990)

[13] Mardia, H. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis*, Academic Press, New York, 1979

[14] van de Waterbeemd, H., El Tayar, N., Carrupt, P. A., and Testa, B., *J. Comput.-Aided Mol. Des.* **3**, 111–132 (1989); van de Waterbeemd, H., Carrupt, P. A., El Tayar, N., Testa, B., and Kier, L. B., Multivariate Data Modeling of New Steric, Topological and CoMFA-Derived Substituent Parameters. In: *Trends in QSAR and Molecular Modeling '92*. Wermuth, C. G., ed., ESCOM, Leiden (1993), p. 69–75

[15] Box, G. E. P., Hunter, W. G., and Hunter, J. S., *Statistics for Experimenters*, Wiley, New York, 1978

[16] Hellberg, S., Sjöström, M., Skagerberg, B., and Wold, S., *J. Med. Chem.* **30**, 1127–1135 (1987)

[17] Eriksson, L., Jonsson, J., Sjöström, M., and Wold, S., *Quant. Struct.-Act. Relat.* **8**, 204–209 (1989)

[18] Skagerberg, B., Bonelli, D., Clementi, S., Cruciani, G., and Ebert, C., *Quant. Struct.–Act. Relat.* **8**, 32–38 (1989)

[19] Sjöström, M., and Eriksson, L., Application of Statistical Experimental Design and PLS Modeling in QSAR Chap. 3.2, this volume

[20] Frank, I. E., and Friedman, J. H., *Technometrics* **35**, 109–148 (1993)

[21] Clementi, S., Coata, G., Ebert, C., Lassiani, L., Linda, P., Hellberg, S., Sjöström, M., and Wold, S., *Pharmacochem. Lib.* **10**, 19–23 (1987)

[22] Alunni, S., Clementi, S., Edlund, U., Johnels, D., Hellberg, S., Sjöström, M., and Wold, S., *Acta Chem. Scand.* **37**, 47–53 (1983)

[23] Fujita, T., The Role of QSAR in Drug Design. In: *Drug Design: Fact or Fantasy*, Jolles, G., and Wooldridge, K. R. H., eds., Academic Press., London (1984), p. 19–33

[24] Draper, N. R., and Smith, H., *Applied Regression Analysis*, Wiley, New York, 1981, Chap. 6

[25] Lanteri, S., *Chemom, Intell. Lab. Syst.* **15**, 159–169 (1992)

[26] Kowalski, K. G., *Chemom. Intell. Lab. Syst.* **9**, 177–184 (1990)

[27] Topliss, J. G., and Edwards, R. P., *J. Med. Chem.* **22**, 1238–1244 (1979)

[28] Diaconis, P., and Efron, B., *Scientific American*, 96–108 (1983)

[29] Efron, B., *J. Amer. Statist. Assoc.* **78**, 316–331 (1983)

[30] Cramer, D. R., Bruce, J. D., Patterson, D. E., and Frank, I. E., *Quant. Struct.-Act. Relat.* **7**, 18–25 (1988)

[31] Cruciani, G., Baroni, M., Bonelli, D., Clementi, S., Ebert, C., and Skagerberg, B., *Quant. Struct.-Act. Relat.* **9**, 101–107 (1990)

[32] Dunn III, W. J., and Wold, S., *J. Med. Chem.* **23**, 595–599 (1980)

[33] Hasegawa, K., Miyashita, Y., Sasaki, S. I., Sonoki, H., and Shigyou, H., *Chemom. Intell. Lab. Syst.* **16**, 69–75 (1992)

[34] Cecchetti, V., Fravolini, A., Bonelli, D., Clementi, S., and Cruciani, G., *Pharmacochem. Lib.* **16**, 393–396 (1991)

[35] Baroni, M., Clementi, S., Cruciani, G., Kettaneh-Wold, N., and Wold, S., *Quant. Struct.-Act. Relat.* **12**, 225–231 (1993)

[36] Mitchell, T. J., *Technometrics* **16**, 203–210 (1974)

[37] Wold, S., and Eriksson, L., Validation Tools. Chap. 5.1, this volume

[38] Cruciani, G., Baroni, M., Clementi, S., Costantino, G., Riganelli, D., and Skagerberg, B., *J. Chemometrics* **6**, 335–346 (1992)

[39] Baroni, M., Clementi, S., Cruciani, G., Costantino, G., Riganelli, D., and Oberrauch, E., *J. Chemometrics* **6**, 347–356 (1992)

[40] Cramer, R. D., Clark, M., Simeroth, P., and Patterson, D. E., *Pharmacochem. Lib.* **16**, 239 to 242 (1991)

[41] Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R., and Clementi, S., *Quant. Struct.-Act. Relat.* **12**, 9–20 (1993)

[42] Pitea, D., Cosentino, U., Moro, G., Bonati, L., Fraschini, E., Lasagni, M., Todeschini, R., Chemometrics and Molecular Modeling. In: van de Waterbeemd, H., ed., *Advanced Computer-Assisted Techniques in Drug Discovery*. (Methods and Principles in Medicinal Chemistry, Vol. **3**, Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., eds., VCH, Weinheim, 1995

[43] Kubinyi, H., personal communication to S.C.

[44] Leardi, R., Boggia, R., and Terrile, M., *J. Chemometrics* **6**, 267–281 (1992)

[45] Wold, S., Kettaneh-Wold, N., and Skagerberg, B., *Chemom. Intell. Lab. Syst.* **7**, 53–65 (1989)

[46] Wold, S., *Chemom. Intell. Lab. Syst.* **14**, 71–84 (1992)

[47] Manallack, D. T., Livingstone, D. J., Neural Networks – a Tool for Drug Design. In: van de Waterbeemd, H., *Advanced Computer-Assisted Techniques in Drug Discovery*. Methods and Principles in Medicinal Chemistry, Vol. **3**, VCH, Weinheim, 1995

[48] Qin, S. J., and McAvoy, T. J., *Comput. Chem. Eng.* **16**, 379–391 (1992)

[49] Breiman, L., and Friedman, J. H., *J. Amer. Stat. Assoc.* **80**, 380–619 (1985)

[50] Frank, I. E., and Lanteri, S., *Chemom. Intell. Lab. Syst.* **3**, 301–313 (1988)

[51] Clare, B. W., Alternating Conditional Expectations in QSAR. In: van de Waterbeemd, H., *Advanced Computer-Assistet Techniques in Drug Discovery* Methods and Principles in Medicinal Chemistry, Vol. **3**, Mannhold, R., Krogsgaard-Larsen, P., Timmerman, H., eds., VCH, Weinheim, 1995

[52] Davis, L., *Genetic Algorithms and Simulated Annealing*, Pitman, London, 1987

[53] Esbensen, K. H., personal communication to S.C.

[54] Bush, B. L., and Nachbar, R. B. Jr., *J. Comp. Aid. Mol. Des.* 7, 587–619 (1993)

[55] Katsumi, H., Yoshida, M., Kikuzono, Y., Takayama, C., and Marsili, M., Analyt. Sci. 7, 719 (1991)

[56] Lindgren, F., Geladi, P., and Wold, S., *J. Chemometrics* 7, 45–49 (1993)

[57] Cramer III, R. D., Patterson, D. E., and Bunce, J. D., *J. Am. Chem. Soc.* **110**, 5959–5967 (1988)

[58] Goodford, P. J., *J. Med. Chem.* **28**, 849–857 (1985)

[59] Kim, K. H., *Quant. Struct. Act. Relat.* **11**, 309–317 (1992)

[60] Clementi, S., Cruciani, G., Riganelli, D., Valigi, R., Costantino, G., Baroni, M., and Wold, S., *Pharm. Pharmacol. Lett.* **3**, 5–8 (1993)

[61] Austel, V., *Experimental Design*. Chap. 3.1, this volume

[62] Clementi, S., Cruciani, G., Curti, G., and Skagerberg, B., *J. Chemometrics* **3**, 499–509 (1989)

[63] Shao, J., *J. Amer. Stat. Assoc.* **88**, 486–494 (1993)

[64] Kubinyi, H., and Abraham, U., Practical Problems in PLS Analyses. In: *3D QSAR in Drug Design: Theory, Methods and Applications*, Appendix B., Kubinyi, H., ed., ESCOM, Leiden, 1993

[65] Lindgren, F., *Third generation PLS*, Umeå University, ISBN 91-7174-911-X; Rännar, S., Lindgren, F., Geladi, P., and Wold, S., *J. Chemometrics* **8**, 111–126 (1994)

[66] Clementi, S., Cruciani, G., Riganelli, D., Valigi, R., GOLPE: merits and darwbacks in 3D-QSAR, In: *Trends in QSAR and Molecular Modeling '94*, Sanz, F., ed., Prous, J. R., Science, ed., Barcelona, in press.

Chemometric Methods in Molecular Design
edited by Han van de Waterbeemd

Index