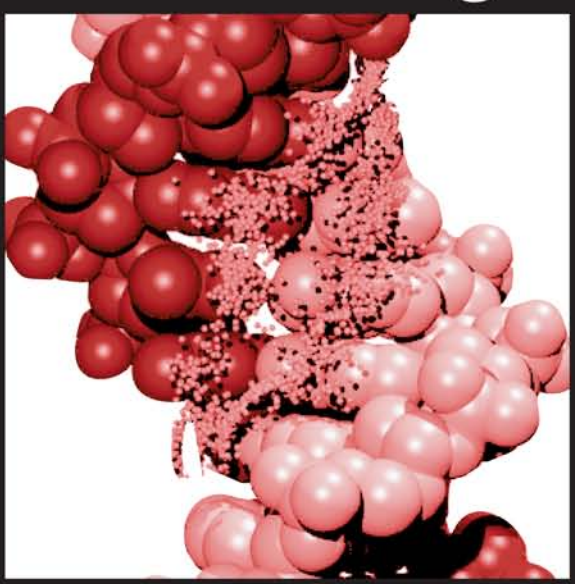
Quantitative Structure-Activity Relationship

(QSAR)

Models of Mutagens and Carcinogens



Edited by
Romualdo Benigni

CRC PRESS

Quantitative Structure-Activity Relationship **(QSAR)** Models of Mutagens and Carcinogens

Quantitative Structure-Activity Relationship **(QSAR)** Models of Mutagens and Carcinogens

Edited by
Romualdo Benigni

CRC PRESS

Boca Raton London New York Washington, D.C.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20131029

International Standard Book Number-13: 978-0-203-01082-2 (eBook - PDF)

Preface

In recent decades, the "big" science, the science that has generated thousands of specialized publications and occupied the front pages of newspapers, has been the science of *life*. The discoveries of molecular biology and of the Human Genome Project have been made under the watchful eyes of everyone. Whereas genetics and, in general, modern biology were developed on a strong quantitative basis (just remember the research of the father of genetics, Mendel), sometime in the 1970s, the life sciences started to rely more and more on simply qualitative approaches, and quantitative methods all but disappeared from the curricula of investigators. However, in recent years, investigators in biomedical research have recognized that the amount of data being generated, particularly with newer genomics technologies, cannot be easily managed, and further progress will be possible only if a strong quantitative (computational) dimension is added to the area. This is the essence of the so-called *bioinformatics revolution*.

The science of quantitative structure–activity relationships (QSARs), at the interface between chemistry and biology, is an exception; it is one of the few fields of biomedical research where a systematic quantitative character has been maintained since its beginnings in the 1960s. QSAR was initiated by the pioneering work of Corwin Hansch and other researchers, who found the way to combine two areas that seemed to be far apart: physical chemistry and biology. The tool that permitted such an operation was mathematical modeling: "By science is meant mathematical descriptions using a relatively small number of well tested parameters and graphics to make the connections" (C. Hansch).

QSAR analysis, permitting the quantitative study of the interaction between chemicals and life, has been applied with success in many different areas. The use of QSARs has become very popular in the field of rational design of drugs and pesticides because it supports faster and more efficient design. Many books have presented the principles of QSARs and applications primarily to rational drug design. This is the first book devoted, in a comprehensive way, to QSAR studies on chemical mutagens and carcinogens. Mutagenicity and carcinogenicity are chronic toxic effects of primary importance to human health. Cancer is the second leading cause of mortality in the Western countries, after cardiovascular diseases. Mutations are involved in the causation of many cancers and are at the origin of heritable diseases as well. A considerable amount of cancer and mutation is provoked by chemicals (e.g., environmental pollutants, professional exposure, food constituents, tobacco smoking). QSAR methods can contribute to elucidation of mechanisms, identification of toxic chemicals solely on the basis of the chemical structure, design of safer chemicals, and reduction of animal studies.

This book provides information for both the newcomer and the expert and is intended to be useful to both biologists and chemists. The book provides background

information on the principles of QSAR modeling, as well as on the biological mechanisms of action of toxic chemicals, and includes extensive surveys of existing QSAR models focusing on individual classes of chemical mutagens and carcinogens. It also provides information on web-based resources of carcinogenicity and mutagenicity data and issues pertaining to the use of these data in QSAR study. A number of well-characterized QSAR applications are presented in specific chapters. To offer a wider perspective, a comparison is made between QSAR models for mutagenicity and carcinogenicity and those for the environmental toxicity of the chemicals. Finally, the potential and limitations of QSAR models as supporting tools for risk assessment are treated extensively.

The Editor

Romualdo Benigni received his education in chemistry at the University of Rome La Sapienza. He then joined the Istituto Superiore di Sanita' (Italian National Institute of Health), where he assumed a permanent position in 1977 and remained, except for two sabbaticals, one at the New York University in 1988, the other at the Jawaharlal Nehru University in New Delhi in 2000. He has worked experimentally in the field of molecular biology and environmental chemical mutagenesis. In the 1980s, he turned his attention to the statistical analysis and modeling of toxicological data and to the study of the relationships between the structure of organic compounds and their toxicological properties (primarily mutagenesis and carcinogenesis). Dr. Benigni has published over 100 journal articles and book chapters based on his applications of a wide variety of quantitative analysis techniques, including QSAR, to the examination of chemical toxicity information. Dr. Benigni's work stands out in terms of its breadth, objectivity, and creativity and in terms of his ability to offer clear and insightful quantitative analysis of toxicological information.

Contributors

Subhash C. Basak
University of Minnesota, Duluth
Duluth, Minnesota

Romualdo Benigni
Istituto Superiore di Sanita'
Rome, Italy

Mark T.D. Cronin
John Moores University
Liverpool, England

Rainer Franke
Consulting in Drug Design GbR
Basdorf, Germany

Alessandro Giuliani
Istituto Superiore di Sanita'
Rome, Italy

Andreas Gruska
Consulting in Drug Design GbR
Basdorf, Germany

Brian D. Gute
University of Minnesota, Duluth
Duluth, Minnesota

Douglas M. Hawkins
University of Minnesota
Minneapolis, Minnesota

David Y. Lai
U.S. Environmental Protection Agency
Washington, D.C.

Denise Mills
University of Minnesota, Duluth
Duluth, Minnesota

Laura Passerini
Istituto Superiore di Sanita'
Rome, Italy

Ann M. Richard
U.S. Environmental Protection Agency
Research Triangle Park, North Carolina

Herbert S. Rosenkranz
Florida Atlantic University
Boca Raton, Florida

ClarLynda R. Williams
U.S. Environmental Protection Agency
Durham, North Carolina

Yin-Tak Woo
U.S. Environmental Protection Agency
Washington, D.C.

Abstract

One of the salient characteristics of the scientific life in recent years has been the explosion of the so-called *bioinformatics revolution*. Investigators in biomedical research have recognized that further progress will be possible only if a strong quantitative dimension is added. The science of quantitative structure–activity relationships (QSARs), at the interface between chemistry and biology, has been one of the few fields of biomedical research where a systematic quantitative character has been present for decades. A number of books have presented its principles and applications to the design of pharmaceutical drugs and pesticides. This is the first book devoted, in a comprehensive way, to QSAR studies on chemical mutagens and carcinogens.

Mutagenicity and carcinogenicity are chronic toxic effects of primary importance to human health. QSAR methods can contribute to elucidation of mechanisms, identification of toxic chemicals solely on the basis of chemical structure, design of safer chemicals, and reduction of animal studies. This book provides information for the newcomer and the expert and is intended to be useful to biologists and chemists. It provides background information on the principles of QSAR modeling, as well as on the biological mechanisms of action of toxic chemicals. The book includes extensive surveys of existing QSAR models focusing on individual classes of chemical mutagens and carcinogens. It also provides information on web-based resources of carcinogenicity and mutagenicity data and issues pertaining to the use of these data in QSAR study. A number of well-characterized QSAR approaches are presented in specific chapters. To offer a wider perspective, a comparison is made between QSAR models for mutagenicity and carcinogenicity and those for the environmental toxicity of chemicals. Finally, the potential and limitations of the QSAR models as supporting tools for risk assessment are treated extensively.

Table of Contents

1 General Introduction to QSAR

Rainer Franke and Andreas Gruska

CONTENTS

1.1 INTRODUCTION

Classical chemometric QSAR methods for the analysis of quantitative structure–activity relationships (QSARs) are sometimes regarded to be out of fashion when compared with the rapid development of molecular modeling, structure-based design, and protein crystallography. In addition, an equation is more difficult to understand than a colored three-dimensional picture generated by computer graphics. However, classical QSAR methods still play an important role and will continue to be a useful tool in modern drug design.[1–3] They have contributed greatly to the development of science in medicinal chemistry (QSAR "know how"), and thousands of documented QSARs and success stories of QSAR predictions and QSAR-guided

0-8493-1507-7/03/$0.00+$1.50

drug design attest to their versatility. In particular, the quantitative description of pharmacokinetic processes remains the domain of classical QSAR techniques. This aspect and QSAR-based concepts such as "drug likeness" are gaining in importance in connection with high throughput screening (HTS) for hit to lead decisions in order to avoid the selection of compounds with unfavorable adsorption/distribution/metabolism/excretion (ADME) properties. Another important issue is the design of safe and selective compounds and a better understanding of toxic, carcinogenic, or mutagenic effects.

This chapter presents a condensed introduction to the most important classical QSAR methods with the main emphasis on Free–Wilson and Hansch analyses. Only references absolutely essential for the understanding of the text will be presented with no attempt for completeness in the sense of a review. For a follow-up, the reader is referred to a number of monographs[2–21] on various aspects of the QSAR field, to the proceedings of the European QSAR conferences (see References 22 to 25 for the last four meetings), and to the journal *Quantitative Structure–Activity Relationships*, which provides an excellent and exhaustive abstract service.

1.2 SOME BASIC PRINCIPLES

Probably the first general formulation of a quantitative structure–activity relationship was presented by Crum-Brown and Fraser in 1868 who assumed that biological activity is a function of chemical structure ("constitution"):

$$\Phi = f(C) \tag{1.1}$$

From this general formulation to the development of true QSARs was still a long way to go because it was necessary to define proper measures of F, suitable mathematical formalisms for the function *f*, and methods to quantitatively describe chemical structure *C*. Modern QSAR technology started in 1964 with publications by Hansch and Fujita[26] and Free and Wilson.[27] The first publication led to development of the well-known *Hansch analysis*, the most widely-used QSAR method also known as the *extrathermodynamic* or *linear free-energy-related* approach. The second paper resulted in development of the so-called Free–Wilson analysis, which supplements Hansch analysis and has turned out to be a very useful method for certain types of structural modifications. Both methods use multiple regression analysis as the mathematical method (*f* in Equation (1.1)) but differ in the description of chemical properties. In Hansch analysis, substituent constants and other physicochemical descriptors are used, while Free–Wilson analysis is based on chemical fragments directly derived from the two-dimensional structure of compounds.

Today, a large variety of mathematical methods is available to express the *f* in Equation (1.1). To name just a few, the most frequently used methods are multiple regression analysis, principal component and factor analysis, principal component regression analysis, partial least squares (PLS), discriminant analysis and other classification methods, and neuronal nets. The variety of mathematical methods is accompanied by a huge number of chemical descriptors to characterize chemical structure; an impressive encyclopedic guide to such descriptors has been presented

by Todeschini and Consonni in their *Handbook of Molecular Descriptors*.[28] Not all of these descriptors have proven to be useful. Broadly speaking, they may be categorized as experimental quantities, such as log *P*, pKa (these quantities can also be computed; see below), and spectroscopic data; substituent constants (electronic, hydrophobic, and steric); parameters derived from molecular modeling and quantum chemical computations; graph theoretical indices; and variables describing the presence or the number of occurrences of certain substructures.

Typical measures of biological activity are the molar concentration C of a compound producing a certain effect derived from a dose–response curve (e.g., ED_{50} or IC_{50}); binding, association, or inhibition constants; and rate constants. In order to obtain larger values for more active compounds, reciprocal values are usually considered for dissociation constants and the molar-concentration-based quantities. Based on thermodynamic or kinetic reasoning, such parameters can be turned into free-energy-related quantities by logarithmic transformation, which is required for the formalism of Hansch analysis (for a detailed discussion, see Franke[7]). Thus, typical expressions for F in Equation (1.1) are $pC = -\log C = \log 1/C$ (examples: pED_{50} or pIC_{50}), $\log K$ (where K is a binding, inhibition, or rate constant), and $\log 1/K_d$ (where K_d is a dissociation constant). By convention, the logarithmic transformation of biological measurement is used not only in Hansch analysis (or other methods based on linear free energy relationships) but in all QSAR approaches applied to quantitative (continuous) biological measurements. One of the reasons is that the results are better comparable. Sometimes, biological measurements result in %effect data measured at a single dose. Strictly speaking, such data are not suitable for Hansch-type and related QSAR approaches. Experience has shown, however, that such data can still lead to meaningful QSARs after logarithmic transformation, provided that the entire range from a few percent values to values close to 100% is covered. A good alternative for such values is a logit transformation according to:

$$\Phi = \log\left(\%\,\text{effect}/(100 - \%\,\text{effect})\right) \tag{1.2}$$

Another alternative is to translate %effect data into a classification scheme that can then be analyzed by classification methods. Such methods are also necessary if biological measurements only allow a scoring of biological potency. In the following text, the logarithmically transformed activity values will be designated as log *BR* (BR = biological response).

1.3 FREE–WILSON ANALYSIS

The Free–Wilson analysis can be applied to series of compounds where the compounds consist of a common (constant) parent structure and variable fragments (usually substituents) (see Figure 1.1). The basic assumptions of Free–Wilson analysis are:

- The parent structure and each variable fragment contribute an additive increment to the logarithm of biological response.

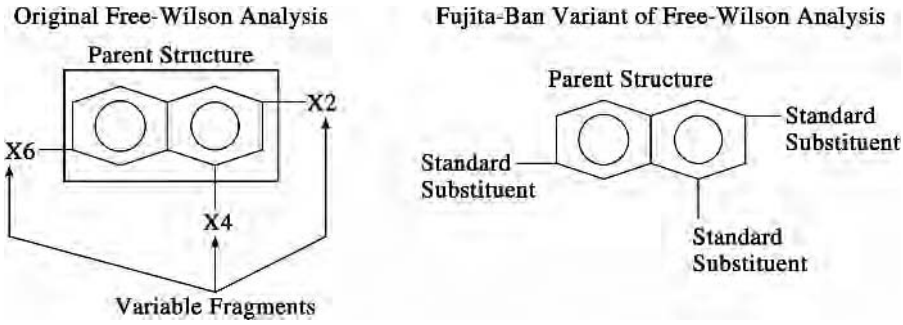


FIGURE 1.1 Schematic presentation of the parent structure according to the original Free–Wilson formalism and in the Fujita–Ban variant of Free–Wilson analysis.

- The increment of a given fragment is constant and independent from structural variations in other positions (no interaction between the fragments).

For each molecule of the series, the following relation is then valid:

$$\log BR_i = \mu + \Sigma b_{ijk} z_{jk} \tag{1.3}$$

BR_i = biological response of the ith molecule
μ = activity contribution of the parent structure
z_{jk} = activity contribution of the jth substituent in the kth position of substitution
b_{ijk} = indicates the occurrence of substituents in each compound
= 1 for compound i, if the jth substituent in the kth position of substitution occurs in this compound

The z_{jk} are derived by multiple regression analysis. Input is the so-called Free–Wilson matrix:

- *Rows:* compounds
- *Columns:* biological potency and variable fragments
- *Elements:* b_{ijk}

Free and Wilson[27] considered the compounds shown in Figure 1.2, and the Free–Wilson matrix is presented in Table 1.1. Each row represents one molecule according to (terms in brackets represent the activity contributions of the respective constituents):

$$[R\text{–H}] + [X\text{–NO}_2] + [Y\text{–NO}_2] + \mu = 1.78$$

$$[R\text{–H}] + [X\text{–Cl}] + [Y\text{–NO}_2] + \mu = 1.32$$

$$\vdots$$

$$[R\text{–Me}] + [X\text{–Br}] + [Y\text{–MeCONH}] + \mu = 1.88$$

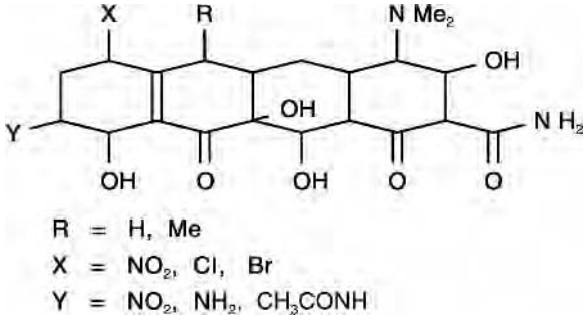


FIGURE 1.2 Compounds considered by Free and Wilson.

Thus, a system of linear equations is obtained from which the activity contributions of the parent structure and of the substituents can be computed by regression analysis; however, the equations are linearly dependent so this problem is not solvable. Two approaches overcome this difficulty:

1. *Introduction of so-called symmetry conditions.* The sum of activity contributions for each position of substitution is set equal to zero (original Free–Wilson analysis).
2. *Fujita–Ban variant of Free–Wilson analysis.* A standard substituent is defined for each position of substitution, and the activity contributions of these standards are set equal to zero. The parent structure is now defined as basic skeleton + standard substituents (see Figure 1.1), and all activity contributions of the nonstandard substituents are computed relative to those of the standards.

Today, the Fujita–Ban variant of Free–Wilson analysis is used because it is much simpler. In addition, the activity contributions from this form of Free–Wilson analysis

TABLE 1.1
Free–Wilson Matrix for the Compounds in Figure 1.2

	b_{ijk}								
i	***R*–H**	***R*–Me**	***X*–NO_2**	***X*–Cl**	***X*–Br**	***Y*–NO_2**	***Y*–NH_2**	***Y*–MeCONH**	**log 1/*C***
1	1	0	1	0	0	1	0	0	1.78
2	1	0	0	1	0	1	0	0	1.32
3	1	0	0	0	1	1	0	0	1.18
4	1	0	0	1	0	0	1	0	2.72
5	1	0	0	0	1	0	1	0	2.51
6	1	0	1	0	0	0	1	0	2.44
7	0	1	1	0	0	0	1	0	2.20
8	0	1	1	0	0	0	0	1	1.18
9	0	1	0	0	1	0	1	0	2.15
10	0	1	0	0	1	0	0	1	1.88

are directly related to substituent constants terms in Hansch analysis which allows both approaches to be mixed (see below). If, in the above example, the substituents R = H, X = NO_2, and Y = NO_2 are selected as standards, the corresponding columns have to be removed from the Free–Wilson matrix, resulting in a modified system of equations. In these equations, the activity contributions of the standard substituents no longer occur as they are zero by definition. The following activity contributions are then obtained:

m = 1.40
[R–H] = 0 (per definition)
[R–Me] = –0.36
[X–NO_2] = 0 (per definition)
[X–Cl] = 0.06
[X–Br] = 0.03
[Y–NO_2] = 0 (per definition)
[Y–NH_2] = 1.13
[Y–MeCONH] = 0.48

It can be seen that variation of substituents in Y has the strongest effects on biological potency with an outstanding positive activity contribution for Y–NH_2. This example was selected for historical reasons. Statistically, the result is significant only at P = 90%.

With the help of the activity contributions, the potency of new analogs (new combinations of substituents) can be estimated. If, in the series considered, N molecules are present where $n_1 - n_4$ substituents are varied in positions 1 to 4, the number of combinatorially possible molecules equals $N_{total} = n_1 ¥ n_2 ¥ n_3 ¥ n_4$ so that the number of possible predictions amounts to $N_{prediction} = N_{total} - N$. Predictions can only be valid as long as the new substituent combinations are compatible with the model assumptions of Free–Wilson analysis (no interactions between substituents).

In many cases not all substituents make significant contributions to log BR. Such substituents should be removed from the analysis. A real problem are substituents that occur only once (unique substituents). Activity contributions for such substituents will contain the full error of measurement of the respective compounds. In addition, unique substituents tend to improve statistics in an unrealistic way as they will always be fit exactly to the regression line. Sometimes, certain substituents always occur together. This will make the corresponding columns in the Free–Wilson matrix linearly dependent. The only possibility in such a case is either to delete columns (which implies that compounds are eliminated) or to combine the substituents in question into a new fictitious substituent that then represents the sum of the activity contributions of the individual substituents.

Intramolecular interactions between variable fragments violate the basic assumption of Free–Wilson analysis; the activity contributions are no longer constant and independent from the presence or absence of other fragments. In such cases, Free–Wilson analysis may still be applied if proper corrections or modifications are introduced. The most commonly used approaches are:

1. Introduction of cross-products to account for interactions (e.g., Bocek–Kopecky model; see Franke[7])
2. Interacting substituents are combined into a fictitious new fragment
3. Hypotheses about interactions are translated into artificial fragments

If applicable, Free–Wilson analysis usually is a good first move. It can effectively aid decision making in experimental work and may also help to develop starting hypotheses for subsequent molecular modeling or other drug-design studies. It is simple (for not too large or complex datasets) and does not require physicochemical descriptors. In cases where a small number of substituents are varied in many positions, Free–Wilson analysis is the method of choice; Hansch analysis and related methods are not applicable in such cases. Experience has shown that the concept of Free–Wilson analysis is not restricted to series where substituents are varied at a constant chemical skeleton. The concept of parent structure and variable fragments with additive and constitutive contributions to biological potency can be stretched fairly far. Thus, variable fragments may include, for example, the exchange of carbon against various heteroatoms in rings or different bridges between certain chemical entities. The most important limitations of Free–Wilson analysis are that predictions can be made only for new combinations of the substituents already present in the series investigated and that mechanistic interpretability is very limited.

1.4 HANSCH ANALYSIS

1.4.1 Basic Assumptions

Hansch analysis is based on the following assumptions:

1. The logarithm of a suitable biological response parameter (*BR*) can be considered to be related to the free energy of binding to the biological target and can thus be described by the same formalisms used in physical organic chemistry to describe equilibrium or rate constants.
2. In congeneric series, substituents make additive and independent contributions to log *BR* (same assumption as in Free–Wilson analysis).
3. These contributions can be factored into hydrophobic, electronic, and steric components that can be described by a linear combination of hydrophobic (x_h), electronic (x_e), and steric (x_s) parameters derived from well-defined chemical standard reactions or from theoretical computations (a_h, a_e, and a_s .. coefficients):

$$\log BR = a_h x_h + a_e x_e + a_s x_s + \text{const.} \tag{1.4}$$

4. If *transport processes* to the site of action are involved, these can be described by a bilinear or parabolic function of log *P* (where *P* is the partition coefficient in the system *n*-octanol/water; but see below). With the parabolic function, the following general expression results:

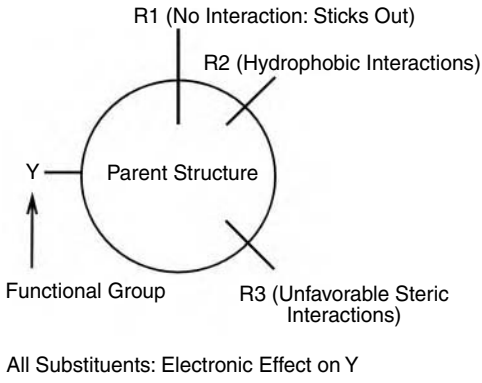


FIGURE 1.3 Hypothetical case of drug–receptor interactions.

$$\log BR = a_h x_h + a_e x_e + a_s x_s - a_1(\log P)^2 + a_2 \log P + \text{const.} \tag{1.5}$$

5. The concrete form of a Hansch equation for a given problem depends on the drug-biosystem interactions. If the hypothetical case of drug–receptor interactions shown in Figure 1.3 is considered, Equation (1.5) would take the following general form (including transport):

$$\log BR = a_h x_h(R2) - a_s x_s(R3) + a_e[x_e(R1) + x_e(R2) + x_e(R3)] - a_1(\log P)^2 + a_2 \log P + \text{const.} \tag{1.6}$$

In this equation, $x_h(\mathrm{R2})$ characterizes the hydrophobicity of substituents in $R2$; $x_s(R3)$ measures steric properties (e.g., size) of substituents in $R3$; and electronic properties of substituents in $R1$, $R2$, and $R3$ are expressed by $x_e(R1)$, $x_e(R2)$, $x_e(R3)$. Clearly, once a Hansch equation is known, an interpretation is possible, allowing conclusions as to the mechanism of action.

1.4.2 Parameters

The huge and ever-increasing number of parameters used in QSAR work during the last decades renders any attempt at a complete discussion an impossible task within this brief QSAR introduction (for an exhaustive review, see Todeschini and Consonni[28]). Thus, only the most commonly used parameters will be presented.

1.4.2.1 Electronic Parameters

The most important relationship to express electronic effects in Hansch type QSARs is the famous Hammett equation which describes the electronic influence of *meta*- and *para*-substituents on reactions occurring at a functional group in substituted benzene derivatives (see Figure 1.4):

$$\log k(X) = \rho\sigma + \log k(\mathrm{H}) = \rho\sigma + \text{const.} \tag{1.7}$$

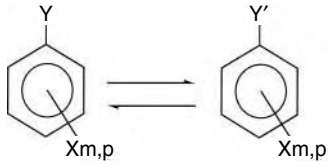


FIGURE 1.4 Structures to which the Hammett equation applies: electronic substituent effects influence the transformation of some functional group *Y* into *Y*¢ (or vice versa).

$$\log K(X) = \rho\sigma + \log K(H) = \rho\sigma + \text{const.} \tag{1.8}$$

In these equations, $k(X)$ and $K(X)$ are the rate and the equilibrium constants, respectively, for a side-chain reaction (Figure 1.4) in a benzene derivative with substituent *X*, and $k(H)$ and $K(H)$ are the corresponding constants for the unsubstituted compound ($X = H$). Called the *electronic substituent constant* or *Hammett constant*, the quantity σ is characteristic of the electronic properties of substituent *X* independent from the type of reaction considered: strongly electron-attracting substituents have high positive values, high negative values indicate electron release, and substituents with small electronic effect have values close to zero. The susceptibility of a given reaction to electronic substituent effects is reflected by the so-called reaction constant ρ, which is independent from substituent properties and characteristic of the reaction considered. Positive (negative) values of ρ indicate that the reaction is enhanced by electron-attracting (electron-releasing) substituents. The scale of ρ values was determined by the dissociation of benzoic acids at 25°C as the reference reaction, where ρ is set equal to unity by definition.

Tabulated values of σ are available for many substituents (e.g., see Hansch and Leo[10] and Hansch et al.[29]). It has been shown that electronic substituent constants can be applied not only to side-chain reactions in benzene derivatives but also to higher condensed aromatic systems, heterocyclic compounds, for the exchange of carbon for heteroatoms in aromatic rings and even for unsaturated aliphatic compounds. In many cases, they can also describe electronic substituent effects on several physical properties such as, for example, ultraviolet, infrared, and nuclear magnetic resonance spectra; polarographic half-wave potentials; ionization potentials; dipole moments; and group dipole moments.

The electronic effects of substituents consist of at least two components: the *inductive effect* (I-effect), which is due to successive polarizations of bonds and electrical through-space effects, and the *mesomeric effect* (M-effect), which results in a change in the overlap of the p_z-orbitals of the electronic system (resonance effect). The relative strength of these components is different in the *meta*- and *para*-positions. In the *meta*-position, the mesomeric effect is small (the I-effect dominates), while a pronounced mesomeric effect operates in the *para*-position. For this reason, σ_{meta} and σ_{para} do not have the same value. Ideally, σ_{meta} and σ_{para} are additive so that in the case of multisubstitution Equations (1.7) and (1.8) become:

$$\log k(X) = \rho\Sigma\sigma + \text{const.} \tag{1.9}$$

$$\log K(X) = \rho\sigma + \text{const.} \tag{1.10}$$

In the context of Hansch analysis, σ values describe substituent effects on the electron density at centers in the molecule that are involved in drug biosystem interactions; they provide a measure of the electron-attracting power of substituents relative to hydrogen. Thus, one possibility to express x_e in Equations (1.4) and (1.5) is σ or $\Sigma\sigma$. As compared to the total number of known QSARs, there are relatively few cases where biological potency depends only on electronic substituent effects. One example is the inhibition of *p*-hydroxybenzoat hydrolase by *para*-substituted benzoic acids: [30]

$$\log 1/C = -1.47(\pm 0.43)\sigma_{para} + 4.30 \tag{1.11}$$
$$n = 6,\ r = 0.978, \text{ and } s \text{ not given}^*$$

Equation (1.11) simply means that enzyme inhibitory potency increases with the electron-releasing power of the substituents.

In complex molecules, a position dependence of the electronic effect may occur requiring different values of ρ for different positions of substitutions. This is particularly true for *ortho*-substituents, where the electronic effect is influenced by steric factors. Several attempts have been made to overcome this problem, including the definition of special constants for *ortho*-substituents, but with only limited success. A common practice in QSAR work is to use σ_{para} values for *ortho*-substituents, allowing for a different regression coefficient (ρ value different from the value for *meta*- and *para*-substituents).

The simple Hammett constant is only valid if, in the series considered, the relative weight of the inductive and mesomeric effects is constant. For this and some other reasons, many modifications of the Hammett equations have been made, resulting in more than 40 different scales of electronic substituent constants. A detailed discussion is far outside the scope of this brief introduction; for an overview in the context of QSAR work and references referring to the evaluation of electronic substituent constants, see, for example, Franke,[7] Hansch and Leo,[10] and Todeschini and Consonni.[28]

In an attempt to simplify this very complicated situation, Swain and Lupton (see Hansch and Leo[10]) introduced two new constants where one, the field constant ¡, is supposed to reflect the inductive substituent effect, while the other, the resonance constant ¬, was attributed to the resonance effect. According to Swain and Lupton, a linear combination of ¡ and ¬ can reproduce any σ scale. Although some of the assumptions underlying the concept of ¡ and ¬ have been criticized, these values have found wide application in the QSAR field; mostly, a modified version according to Hansch and Leo[10] is used. The ¡ scale can be regarded as well established: ¡ equals the so-called inductive substituent constant σ_I which can be estimated from the dissociation of 4-substituted bicyclo[2.2.2]octane-1-carboxylic acids (various

* In this and following equations, *n* is the number of compounds in the series, *r* is the correlation coefficient (measure of goodness), and *s* is the standard deviation (also measure of goodness); for more information, see Section 1.4.3.

other definitions of the inductive constant also exist). ¡ and ¬ are related to s_m and s_p (Hansch and Leo[10]):

$$¡ \int s_I = 1.297s_m - 0.385s_p + 0.033 \tag{1.12}$$

$$¬ = s_p - 0.921 \tag{1.13}$$

The ¬ scale, however, is not of universal validity as it is not independent of the reaction center. This is of particular importance in compounds where direct resonance interactions between substituents and the reaction center can occur (through resonance). This is possible when (1) an electron-donating substituent (e.g., NH_2), is present in the *para*-position, while the reaction center carries a positive charge or has an electron deficiency (positive resonance); or (2) an electron-attracting substituent in the *para*-position (e.g., NO_2) has an electron-donating reaction center as its counterpart (negative resonance). For these situations, the so-called enhanced substituent constants were introduced: s^+ (positive resonance; defined by the solvolysis of *t*-cumyl chlorides) and s^- (negative resonance; defined by the ionization of phenols or anilines in water). With these quantities, enhanced values of ¬ can be defined as:

$$\Re^+ = \sigma_p^+ - \Im \tag{1.14}$$

$$\Re^- = \sigma_p^- - \Im \tag{1.15}$$

For aliphatic compounds, the polar substituent constant s^* according to Taft can be used. This constant is derived from the acid- and base-catalyzed hydrolysis of aliphatic esters XCOOR, with $X = CH_3$ as the standard substituent ($s^*(CH_3) = 0$). An alternative is Charton's inductive substituent constant s_I, which is based on the dissociation of substituted acetic acids in water.

Instead of electronic substituent constants, experimental quantities such as, for example, pKa values, spectroscopic data, or polarographic half-wave potentials can also be used to express electronic properties in Hansch analysis. The disadvantage is that such values are usually only available for those compounds already synthesized (software for the calculation of pKa values is available).

Another alternative is the use of quantum-chemical parameters. From among the large variety of such parameters, the following have most widely been used in the framework of Hansch analysis:

- Energy of the highest occupied and the lowest unoccupied molecular orbital (E_{HOMO} and E_{LUMO})
- Charges at selected atoms
- Dipol moments

1.4.2.2 Hydrophobic Parameters

Hydrophobicity (also called *lipophilicity*) is of central importance for biological potency as it plays a role not only in the interaction of drugs with many targets but

also in pharmacokinetic processes (for an excellent review, see, for example, Taylor[31]). Hydrophobicity characterizes the tendency of molecules (or parts of molecules) to escape contact with water and to move into a lipophilic environment. In QSAR work, the basic quantity to measure hydrophobicity is the logarithm of the partition coefficient in the system *n*-octanol/water, log *P*, which was introduced by Hansch. Its use is based on the Collander equation relating partition coefficients from different solvent/water systems with the tacit assumption that lipophilic biophases behave like organic solvents (where P_I is the partition coefficient in the system solvent I/water, and P_{II} is the partition coefficient in the system solvent II/water):

$$\log P_{II} = a \log P_I + b \tag{1.16}$$

This seems to be true in many cases, as hundreds of examples of Collander-type relationships between biological data and *n*-octanol/water partition coefficients can be found in the literature. An example is the partitioning between red cell ghosts and water of alcohols, phenols, and ethyl carbamate (taken from Hansch and Leo[10]):

$$\log P_{Ghosts} = 0.83(\pm 0.10) \log P - 0.34(\pm 0.26) \tag{1.17}$$
$$n = 11,\ r = 0.987, \text{ and } s = 0.175$$

The Collander equation is only valid as long as the solute–solvent interactions in the two solvents are sufficiently similar. Principal component analysis has shown[32] that log *P* is mainly determined by two solute properties: bulk with a polarity component and hydrogen bonding. If, for example, hydrogen bonding in two organic solvents is different, the Collander equation will break down unless a correction for hydrogen bonding is introduced. This can be difficult if the difference is large and the compounds considered have a very strong capability to form hydrogen bonds. A case in point is penetration of the blood–brain barrier by a set of very polar H_2-antihistaminic drugs possessing several hydrogen-bond acceptor and donor sites. No correlation with log *P* could be found, but a strong dependence on hydrogen bonding as expressed by Seiler's Dlog *P* values was observed:[33]

$$\log (C_{brain}/C_{blood}) = -0.48(\pm 0.16)\ \mathrm{D}\log P + 0.89(\pm 0.50) \tag{1.18}$$
$$n = 20,\ r = 0.83, \text{ and } s = 0.44$$

The hydrogen-bonding ability, Dlog *P*, according to Seiler,[24] is defined as the difference between cyclohexane/water and *n*-octanol/water partition coefficients:

$$\mathrm{D}\log P = \log P_{oct} - \log P_{cyclohexane} = \mathrm{S}I_H - 0.16 \tag{1.19}$$

The I_H values characterize the hydrogen-bonding ability of different functional groups.

Even though log *P* (or quantities derived from log *P*; see below) have been shown to be valid hydrophobicity descriptors in the majority of cases, examples such as that presented in Equation (1.18) and the awareness of the complexity of drug–

membrane interactions[35] have led to intensive investigations of the properties of log *P* and its use in QSAR work (for reviews, see Pliska et al.,[17] Testa et al.,[21] and Taylor[31]), and alternative approaches to describe hydrophobicity in QSAR work have been suggested. Leahy and co-workers[36] proposed that partition coefficients from four solvent/water systems with different hydrogen-bonding behavior of the solvents are required in order to describe the properties of biological membranes. Another strategy is to dissect log *P* into its components and to describe hydrophobicity by a linear combination of bulk/polarity and hydrogen-bonding parameters. An example is Equation (1.20) for blood–brain permeation derived by Abraham and Chadha[37] based on the theory of linear solvation energy relationships:

$$\begin{aligned}\log(C_{\text{brain}}/C_{\text{blood}}) = {} & 1.03(\pm 0.10)V_x - 0.54(\pm 0.10)\mathsf{p}_2^H - 0.61(\pm 0.13)\mathsf{Sa}_2^H \\ & - 0.71(\pm 0.11)\mathsf{Sb}_2^H - 0.08(\pm 0.06) \qquad (1.20)\\ & n = 57,\ r = 0.948, \text{ and } s = 0.202\end{aligned}$$

In this equation, V_x is the McGowan characteristic volume, p_2^H is the so-called solute dipolarity/polarizability, and Sa_2^H and Sb_2^H are the solute overall hydrogen-bond acidity and overall hydrogen-bond basicity, respectively. The problem with this type of approach is that values of the descriptors p_2^H, Sa_2^H, and Sb_2^H are not easily available. A similar approach also using hydrogen-bond-donor and -acceptor descriptors computed from a collection of thermodynamic data has been used by Raevsky.[38] Hydrogen-bonding capability may also be expressed by the polar surface area, a quantity used in several recent ADME studies (see, for example, van de Waterbeemd[39] and Stenberg et al.[40]). Unfortunately, a general scale for hydrogen-bonding strength does not exist. If the biophase differs from a model solvent in its ability to accommodate a solute (bulk effect) or with respect to the formation of hydrogen bonding, combinations of bulk or hydrogen-bonding parameters with log *P* are also possible (for some examples, see Hansch and Leo,[10] Österberg and Norinder,[41] and Feher et al.[42]).

In spite of the limitations that log *P* obviously has in certain situations, it is still the most widely used hydrophobicity parameter. For ionizable compounds, the distribution coefficient must be considered in many cases instead of the partition coefficient, or suitable corrections for the degree of ionization must be introduced.[2,4,7] By 2002, the QSAR database of the Pomona College[30,43–45] contained more than 5400 examples of QSARs involving log *P* or p (see below). One advantage of log *P* is its straightforward computation from chemical structure. To this end, a variety of different methods have been developed and are available as commercial software (for overviews, see, for example, Leo[46] and Duban et al.[47]). The most widely used method is the Clog *P* algorithm[48] (Biobyte Corp., Claremont, CA), which is based on the hydrophobic fragmental constant of Leo and Hansch derived from very accurate measurement of log *P* values of simple compounds (*constructionist* approach). The fragmental method of calculating log *P* from structure was originally introduced by Rekker,[49] who derived hydrophobic fragmental constants from a large number of available log *P* values via regression analysis (*reductionist* approach). Rekker's system is used in the PrologP software (CompuDrug International, Inc., San Francisco, CA).

The quantity log P characterizes whole molecules. In the QSAR context, this is sufficient for many unspecific endpoints or processes where transport to the site of action or concentration in a certain tissue is the critical factor. If, however, interactions with a receptor or an enzyme are to be analyzed, position dependence comes into play (see Figure 1.3). In such cases, it is necessary to describe the hydrophobicity of the variable parts of the molecules (usually substituents) separately. The most widely used quantity for this purpose is the hydrophobic substituent constant (tabulation in Hansch and Leo[10]), defined for aromatic substituents as:

$$\pi_X = \log P_X - \log P_H \tag{1.21}$$

where P_X is the partition coefficient of a derivative with substituent X, and P_H is that of the unsubstituted parent compound in the system *n*-octanol/water. The substituent constant π represents the hydrophobic analog of the electronic Hammett constant and characterizes the hydrophobicity of substituent X relative to hydrogen. It has become common practice to use π values derived from mono-substituted benzenes. This, however, is not correct, as functional groups may influence π values of substituents via electronic interactions. Thus, π values from different series with different functional groups are interrelated via electronic corrections which become more important as the respective functional groups differ more in their electronic properties. An example is Equation (1.22) relating π values of a set of substituents derived from mono-substituted benzenes to the corresponding values derived from benzoic acids: [50]

$$\begin{gathered}\pi_X(\text{benzene}) = 1.05(\pm 0.07)\pi_X(\text{benzoic acid}) - 0.18(\pm 0.15)\sigma_X \\ - 0.12(\pm 0.06) \\ n = 27,\ r = 0.986, \text{ and } s = 0.105\end{gathered} \tag{1.22}$$

As a consequence, the use of π values from the benzene system can produce electronic terms in QSARs that are solely electronic corrections for π and not indicative of electronic interactions with the biological target.

1.4.2.3 Steric Parameters

Steric effects are not easy to describe; for a review covering the following parameters and original references for their definition, see, for example, Franke,[7] Hansch and Leo,[10] and Todeschini and Consonni,[28] as well as the tabulations in Hansch and Leo.[10]

The first steric parameter used in QSAR work is the steric substituent constant E_S due to Taft defined by the acid-catalyzed hydrolysis of RCOOR¢ in relation to the methyl substituted parent, $CH_3COOR¢$. E_S characterizes substituent width and is highly correlated with the van der Waals radius of substituents. Originally designed to characterize intermolecular steric effects, it also turned out to be helpful for intramolecular steric interactions; today, this parameter (which also has some modifications) is no longer much used in QSAR investigations. A quantity highly correlated with E_S is Charton's ν-value defined as:

$$\nu_X = r_{v,X} - r_{v,H} = r_{v,X} - 1.20 \tag{1.23}$$

where r_{rx} is the minimum van der Waals radius of a substituent x. A frequently used steric parameter in QSAR work is molar refractivity (*MR*), usually scaled by the factor 10^{-1}. *MR* is related to molar volume (*V*) and the refractive index (*n*) according to

$$MR = [(n^2 - 1)/(n^2 + 2)]V \tag{1.24}$$

Molar refractivity is, in the first place, a measure of bulk due to its relation with molar volume, but it also contains a polarizability component expressed by the refractive index terms. Because MR is an additive and constitutive quantity, its calculation from chemical structure is straightforward on the basis of available fragment values for both whole molecules as well as substituents. Other bulk parameters occasionally used are the molar volume, the parachor, and the molecular weight in connection with diffusion controlled processes.

The above-mentioned parameters have the disadvantage that they do not take into account molecular shape or, in other words, the directionality of steric interactions typical of the binding of drug molecules to specific biological targets. This led Verloop[51] to introduce his STERIMOL parameters. Originally, five parameters were suggested to describe steric properties of a substituent, but it then turned out that three parameters are sufficient: *B*1, *B*5, and *L*. *L* is a measure of substituent length, *B*1 is the smallest substituent width, and *B*5 is the largest width orthogonal to *L* (a measure of the effective substituent volume).

A very large group of parameters are topological indices (see Todeschini and Consonni[28]) based on graph theoretical considerations. They can directly be computed from the two-dimensional structure of any compound, as, for example, the Kier–Hall connectivity index c. The use of such indices in QSAR work has been extensively investigated by Kier and Hall (for reviews and computational procedures, see Kier and Hall[52,53]). They are related to many physicochemical properties, including hydrophobicity, and, thus, are no pure steric quantities. As a consequence, the interpretation of QSARs containing such parameters is very difficult if not impossible. In addition, the many different topological indices calculable from chemical structure are usually highly interrelated. Therefore, an uncritical combination of such indices into one QSAR model exposes the danger of chance correlations and will lead to severe colinearity problems (see below). For these reasons, the use of such indices in typical Hansch analysis problems is to be considered with caution. If, however, large sets of structurally diverse compounds are to be investigated with the primary purpose of data description, then such indices can be very helpful.

In series of sufficiently similar compounds, steric descriptors are frequently correlated with hydrophobicity parameters, creating a problem for interpretation (see below). Their relation with biological potency can be linear (positive slope, favorablestericinteractions;negativeslope,sterichindrance)orparabolic(optimum for steric fit).

1.4.2.4 Indicator Variables

Indicator variables can be used to combine QSARs for subgroups of compounds belonging to the same series but differing in a certain feature into one common

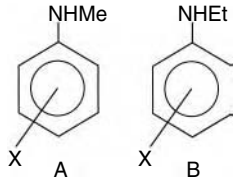


FIGURE 1.5 Structures *A* and *B* (see Equations (1.25) to (1.27)).

equation. Usually, an indicator variable is assigned a value of one for compounds possessing this feature, and a value of zero is assigned to the other compounds. Consider, for example, the structures presented in Figure 1.5 and let us assume that for the two subgroups *A* and *B* the following QSARs for some biological responses are valid:

$$\log BR_A = 0.5\pi + 1.5 \tag{1.25}$$

$$\log BR_B = 0.5\pi + 3.0 \tag{1.26}$$

The two subgroups show the same dependence on π and differ only with respect to the constant term in a plot of log *BR* vs. π. This difference must be due to the different substitution of the amino group which can be accounted for by introducing an indicator variable *I* with $I = 1$ for compounds of subgroup *B* and $I = 0$ for compounds of subgroup *A* (the definition can, of course, also be reversed). Introduction of this indicator as an additional variable into an analysis for the subgroups *A* and *B* combined will give:

$$\log BR_{A+B} = 0.5\pi + 1.5I + 1.5 \tag{1.27}$$

The regression coefficient for the indicator variable is fitted by regression analysis to give the difference in the intercepts. According to the definition of *I*, this difference is then added to all points belonging to subgroup *B* so that the two lines are united. This principle is, of course, also applicable to *n*-dimensional spaces (equations with *n* variables), and it is also possible to combine, through indicator variables, more than two subgroups. A case in point is Equation (1.28) which describes the antibacterial potency of lyncomycin derivatives against *Salmonella lutea*:[54]

$$\begin{gathered}\log BR = -0.24(\pm 0.02)\pi^2 + 1.39(\pm 0.12)\pi + 0.23(\pm 0.07)I_1 \\ - 0.20(\pm 0.07)I_2 - 0.43 \\ n = 25,\ r = 0.960, \text{ and } s = 0.162\end{gathered} \tag{1.28}$$

The indicator variables I_1 and I_2 distinguish, respectively, between *trans*-substituted ($I_1 = 1$) and *cis*-substituted ($I_1 = 0$) and between *N*-ethyl ($I_2 = 1$) and *N*-methyl ($I_2 = 0$) derivatives. Equation (1.28) shows that *trans*-substitution leads to more active compounds, while *N*-ethyl derivatives are less potent than their *N*-methyl analogs.

Indicator variables have been used extensively in QSAR work to account for a variety of structural and other features such as hydrogen bonding (intra- and intermolecular), *ortho* effects, different test organisms, different parent skeletons, etc. Even though they have turned out to be very useful, some care is necessary as the physical meaning of such variables is not always clear in the context of Hansch analysis.

A special case arises if, in a series of compounds with multiple substitution, many substituents are varied at some sites but only a few at other sites. In such cases, substitution at the sites with many substituents can be described by Hansch-type expressions while substituent effects at the other sites can be characterized by a Free–Wilson description (a set of indicator variables that correspond to the b_{ijk} in Equation (1.3)). This amounts to a mixed Hansch/Free–Wilson analysis which is possible, as these two methods are formally equivalent.[2,7] The examples presented above are already simple cases of such a mixed approach.

Sometimes, indicator variables also have to be introduced in the form of cross-products with the other descriptor variables. In the two-dimensional case presented in Figure 1.5 this would then be necessary if the two lines would not only differ in intercept but also in slope. An example is shown in Equation (4.8) in Chapter 4.

1.4.3 Building and Evaluating Hansch Equations

At the beginning of an investigation, one has, for a set of compounds (usually called *training series* or also *learning set*), values for an observed biological potency and the structures of the compounds. In a first step, a set of molecule parameters, x_i, describing chemical structure is selected from tabulations or computed. The basic assumption of Hansch analysis is that log *BR* can be described by a weighted linear combination of the x_i:

$$\log BR = a_0 + a_1 x_1^{n1} + \ldots + a_i x_i^{ni} + \ldots + a_n x_n^{nn} \tag{1.29}$$

where the exponents *ni* can be 1 (linear term) or 2 (quadratic term). At the start of an analysis it is not known whether a relationship according to Equation (1.29) exists for the problem under investigation, which of the x_i are related to log *BR* (which of the regression coefficients a_i are significantly different from zero), and what the precise form of Equation (1.29) is (e.g., occurrence of squared terms). In addition, there may be more than one solution (e.g., if some of the x_i are correlated; usually such variables are termed *colinear*). As a consequence, many possible combinations of the x_i have to be screened, and the resulting equations have to be validated and evaluated to select the "best" equation.

The regression coefficients a_i are computed by multiple regression analysis[55] and checked for their statistical significance. Only such terms are allowed that are significantly different from zero at a statistical probability of 95%. Usually, the following statistical criteria are presented together with a regression equation:

- The correlation coefficient, r, which is a relative measure of the quality of fit ($r = 0$, no correlation; $r = 1$, perfect correlation). Its squared value

(r^2) measures the percentage of variance of the dependent variable (log *BR*) explained by the equation.
- The standard deviation, *s*, is another measure for the quality of fit. Its value should be as small as possible but never smaller than the error of the biological experiment (overprediction).
- Fisher's *F* value, which is a measure of the statistical significance of the regression model.
- Confidence intervals for the regression coefficients at a statistical level of significance of 95%. These intervals overlap the true values of the regression coefficients at a statistical probability of 95%.
- Number of degrees of freedom usually presented as number of observations, *n*.
- Residuals = log *BR* (observed) – log *BR* (predicted).

Frequently, the robustness and the potential predictive power of a QSAR are further checked by a procedure called *cross-validation*. In cross-validation, each compound is left out once from the analysis (leave one out, or LOO, technique),* and the model is then derived from the remaining objects. With the resulting models, the activity values of the left-out compounds are then predicted. By comparing these predicted values with the observed values, a squared cross-validated correlation coefficient, q^2, can be computed which is usually lower than the squared correlation coefficient, r^2. Values of q^2 can range from 1 to less than zero. A value of one indicates a perfect prediction, and a value of 0 means that the QSAR derived has no modeling power. Negative values arise from a situation where the derived QSAR is a poorer description of data than no model at all. For medium-sized or small datasets typical for Hansch analysis, "cross-validation may incorrectly indicate a lack of validity of the QSAR model."[2] For large datasets, the situation is different (see below). Another approach to estimate the potential predictive power of a QSAR model is to divide the datasets into two parts by means of series design methods (see below) which, of course, requires a fairly large number of observations. One part is then taken as a training series to derive the QSAR model, and the other part is used as a so-called test set for which biological potencies are calculated from the QSAR derived from the training series. A high predictive power is characterized by a good correlation between predicted and observed activity values.

Regarding the goodness of description of log *BR*, a regression model can be accepted if:

- The *F* value shows that the overall significance is 95%.
- The confidence intervals are smaller than the regression coefficient (the regression coefficients are significantly different from zero at a statistical probability of >95%).
- $r \geq 0.8$.
- *s* is not much larger than the standard deviation of the biological experiment.
- $q^2 \geq 0.6$.

* There are also cross-validation procedures where groups of compounds are left out.

The selection of descriptor variables at the beginning of an analysis is a complicated and time-consuming procedure that always contains subjective and ambiguous elements. On the one hand, all properties important for the biological activity considered must be covered, but on the other hand too large a set of descriptors will make monitoring of the results and interpretation difficult and may lead to so-called chance correlations where a statistically significant result, as judged by the usual statistical criteria (see above), is obtained by chance.[56] For medium-sized problems, the best approach probably is to start from a standard set of hydrophobic, electronic, and steric substituent constants and to consider all reasonable combinations of parameters. Variables that have only a very small spread (near constant value) are not to be included. Variable combinations from which to start can be selected, for example, from knowledge about already existing QSARs for similar or the same compounds or for the same type of biological activity or from hypotheses about the mechanism of action, respectively. Simple plots can be of great help at this stage. The equations are then improved in an iterative stepwise procedure, adding more variables, if necessary, until an acceptable result is obtained. At this stage, plots of residuals vs. such variables can be very helpful. A very reasonable first move is to break down the training series into subsets to understand positional dependencies of effects. Subsets can then be reunited by means of indicator variables. Sometimes, variables have to be modified in order to meet specific aspects of drug–target interactions. One typical example is so-called *ring flipping*. Phenyl rings with substituents in the *meta*-position can flip to place a hydrophobic substituent in a hydrophobic environment and a hydrophilic one in the aqueous surrounding. As a consequence, hydrophobic *meta*-substituents are parameterized with normal π values, while hydrophilic substituents receive a π value of 0. A case in point is the following equation describing the Michaelis constant for the hydrolysis of hippurate esters, X-$C_6H_4OC(=O)CH_2NHC(=O)C_6H_5$, by papain:[10]

$$\begin{aligned} \log 1/K_m = {} & 0.57(\pm 0.20)\sigma + 1.03(\pm 0.25)\pi_3{}^{\complement} \\ & + 0.61(\pm 0.29)MR_4 + 3.80(\pm 0.17) \\ & n = 25,\ r = 0.907, \text{ and } s = 0.208 \end{aligned} \tag{1.30}$$

In this equation, $\pi_3{}^{\complement}$ refers to the more hydrophobic of the two possible *meta*-substituents; the more hydrophilic *meta*-substituent is supposed to project into the aqueous phase. The coefficient of the $\pi_3{}^{\complement}$ term is close to unity, indicating binding in a hydrophobic pocket (see Franke[7]). The positive coefficient of MR_4 indicates an increase in binding with substituent size which requires enough space or flexibility of the corresponding part in the binding site. Another interesting aspect about the MR_4 term is that, because the compounds with the substituents 4-CH_3, 4-C_2H_5, and 4-C_3H_7 have essentially the same log $1/K_m$ values, it was assumed that they do not make significant contact with the enzyme; therefore, a MR_4 value of 0 was assigned to these substituents. The positive σ term, finally, reveals that electron-withdrawing substituents support the formation of the enzyme–substrate complex. All these conclusions have subsequently been verified by computer graphics based on x-ray crystallographic structures. This and many additional examples where conclusions

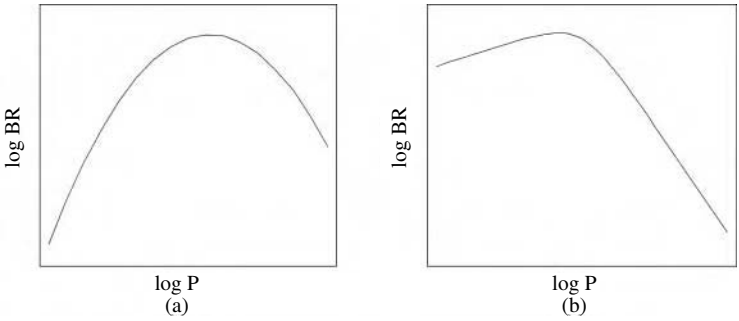


FIGURE 1.6 (left) Shape of a parabola according to Equation (1.31), and (right) shape of a bilinear curve according to Equation (1.33) or Equation (1.35).

from Hansch analysis have later been verified by x-ray crystallography[57] clearly show that the interpretation of QSARs can lead to valid information on the mode of binding.

As already mentioned, molecule parameters in a Hansch equation may occur in linear terms or in a parabolic fashion involving squared terms. Squared hydrophobic terms may occur for both hydrophobic binding as well as pharmacokinetic processes, indicating an optimal value of hydrophobicity (the squared term usually has a negative sign which means that the parabola is curved downward).* In the case of hydrophobic binding, a parabolic relationship with hydrophobicity parameters (see Figure 1.6a) indicates that the hydrophobic binding region at the biological target is limited (see Franke[7]), and for pharmacokinetic processes such relationships reflect an optimum hydrophobicity for the transport to the site of action:

$$\log BR = a\log P - b\log P^2 + c \tag{1.31}$$

The value of log P at the optimum can be computed from:

$$\log P_O = a/2b \tag{1.32}$$

The parabolic function has the disadvantage that the data are forced into a symmetrical relationship while experience has shown that such relationships are not perfectly symmetrical in many cases. A better alternative is the so-called bilinear model according to Kubinyi:[2]

$$\log BR = a\log P - b\log (\mathrm{b}P + 1) + c \tag{1.33}$$

with the optimum at

$$\log P_O = \log (a/\mathrm{b}\ (b - a)) \tag{1.34}$$

If the hydrophobicity parameter is already in the logarithmic scale (log P, p), Equation (1.33) transforms into:

* In some very special cases, a parabola curved upward has also been found.

$$\log BR = a\log P - b\log (b10^{\log P} + 1) + c \qquad (1.35)$$

Equations (1.33) and (1.35) describe a curve with linear ascending and descending sides (see Figure 1.6b) which can have different slopes (a for the ascending and $a - b$ for the descending part). A further advantage of this relationship is that the slopes of the linear parts can be compared with the slopes of linear relationships between log *BR* and hydrophobicity parameters (in such relationships, the variation of hydrophobicity in the compounds of the training series does not cover the region of the optimum). Disadvantages are that more data points are necessary, as one additional adjustable parameter (b) has to be estimated, and computation of the equation requires nonlinear regression analysis. In addition, the slopes are sensitive to the spread in log *P* or p. Examples of the two types of relationships are presented by Equations (1.36) and (1.37) for the antielectric shock activity in mice of miscellaneous compounds (where m is the dipole moment):[10]

$$\log 1/C = 1.15\log P - 0.22\log P^2 - 0.37m + 2.99 \qquad (1.36)$$
$n = 18$, $r = 0.922$, $s = 0.24$, and $\log P_O = 2.59$; confidence intervals not given

$$\log 1/C = 0.86(\pm 0.20)\log P - 1.68(\pm 0.42)\log (b10^{\log P} + 1) - 0.42(\pm 0.14)m + 3.19(\pm 0.20) \qquad (1.37)$$
$n = 18$, $r = 0.938$, $s = 0.221$, and $\log P_O = 2.43$; value of b not given

As the bilinear curve according to Equation (1.37) is fairly symmetrical, the fit with the bilinear model is not much better than with the simple parabola according to Equation (1.36). There is good agreement between the estimates of the optimal value of log *P* which is usually the case if parabolas and bilinear models are compared. The optimal lipophilicity ($\log P_O$) is an important design criterion that can lead to enhancing a desired potency or to decreasing unwanted side effects of drugs. It should be noted that in the case of hydrophobic binding to a target, positional dependencies can occur so that, for example, for substituents in one substitution site a linear relationship exists between log BR and p, while in some other site this relationship is parabolic.

Regarding squared terms, the situation with steric parameters is completely analogous to the behavior of hydrophobic parameters in QSAR equations. Squared terms describing a parabola (see Equation (1.31)) are needed when steric hindrance of binding occurs if substituents exceed a certain size. Frequently, such effects are also better described by relationships corresponding to Equations (1.33) and (1.35). An example is Equation (1.38) describing the rate constant for the inhibition of chymotrypsin by thiophosphonates O= P(SR1)(OR2)(CH_3)(I=1 if a charge occurs in S R1):[10]

$$\log k = 1.47(\pm 0.10)MR(OR2) - 3.43\log (b10^{MR(OR2)} + 1) + 0.34(\pm 0.09)MR(SR1) + 1.25(\pm 0.19)s^*(R1) - 1.06(\pm 0.31)I - 5.26(\pm 0.38) \qquad (1.38)$$
$n = 53$, $r = 0.985$, $s = 0.243$, and $MR(OR2)_O = 3.71$; value of b not given

Equation (1.38) demonstrates a position-dependent steric effect. For substitution at the sulfur, there is a linear increase of inhibition with bulk, while for substitution at the oxygen, a steric optimum exists. The s*(*R*3) term indicates that electron-attracting substituents in *R*1 are favorable, and the indicator variable, finally, shows that charge in the S*R*1 region slows the reaction down.

For electronic terms, quadratic relationships are the exception; however, in some special cases squared s or pKa terms are required.

Sometimes a squared term is supported by only one or two compounds. In such cases, two results should be presented: the parabolic relationship and the linear one resulting after eliminating these compounds. It can then be decided whether it is worthwhile to synthesize additional analogs in order to define the possible optimum.

A severe problem arises from colinearities between variables ($r \geq 0.7$). If such colinearities appear between variables telling very much the same story, such as, for example, between *MR* and *V* (see Equation (1.24)), this situation can simply be handled by omitting that variable that shows the lower correlation with biological potency. If, however, colinearities occur between variables with different physicochemical meaning, alternative equations will result, leading to different interpretations of drug–biosystem interactions with no possibility to decide which is the correct one. Very typical are colinearities between hydrophobic and steric parameters. An example is provided by Equations (1.39) and (1.40) describing the growth inhibition of *Chlorella vulgaris* by piperidinoacetanilides substituted in the phenyl ring:[58]

$$\log 1/C = 1.26\mathrm{s}_{o,m,p} + 0.48\mathrm{p}_m + 0.66\mathrm{p}_p - 1.02MR_o + 2.50 \qquad (1.39)$$
$n = 27$, $r = 0.933$, and $s = 0.252$; confidence intervals not given

$$\log 1/C = 1.34\mathrm{s}_{o,m,p} + 0.51\ \mathrm{p}_m + 0.85MR_p - 1.13MR_o + 2.70 \qquad (1.40)$$
$n = 27$, $r = 0.927$, and $s = 0.253$; confidence intervals not given

Obviously, p_p and MR_p can be freely interchanged without changing anything else. This is due to a high colinearity between p_p and MR_p ($r = 0.843$). As the statistical quality of both equations is the same, it is not possible to decide whether a hydrophobic or a steric effect operates in the *para*-position. In addition to simple colinearities, multicolinearities may exist where one variable is related to a linear combination of two or more other variables (see Franke[7]). The only real possibility to solve such problems is to consider all alternative equations (provided that they are of comparable statistical quality) and then to break down the disturbing colinearities by adding some well-selected additional analogs to the training series in order to understand what features are important for biological potency. Sometimes, comparison with already existing QSARs for the same type of biological activity can also be helpful to make a decision. A strategy of including only one from pairs of related variables in the derivation of the equations is not helpful as the colinearity continues to exist; this is, of course, not true for variables leading to the same or a similar interpretation as for the already-mentioned example of *MR* and *V*.

Another problem in QSAR analyses are compounds that cannot be explained by the derived QSAR model. Such compounds are called *outliers* and are usually

omitted from the corresponding analysis. There are several reasons for a compound to become an outlier as, for example, incorrect biological measurement, incorrect parameter value(s), metabolic inactivation/activation, different mechanism of action, or because it has a unique property not described by the QSAR model. One should always try to rationalize why a compound is an outlier. If this is not possible, the elimination of outliers is a difficult and not unambiguous decision. In any case, it is not an acceptable practice to remove compounds from an analysis until a good fit is obtained. On the other hand, outliers can provide valuable information; for this reason, Hansch has called them "a blessing in disguise."

The selection of the best equation (more than one equation in the case of collinear, not naturally related variables) can be difficult and is very much a matter of personal experience and judgment using statistical, practical, and chemical criteria. From a statistical point of view, the best equation is usually the one with the best fit, provided that the statistical criteria mentioned above are fulfilled. The correlation coefficient is not a good measure here, as it tends to select as many variables as possible. Better criteria are the lowest standard deviation, s, and the highest overall F value. Unfortunately, these two criteria may not lead to the same solution, as the F value sometimes has the highest value for too few variables, while s tends to include too many variables. In this context, an additional criterion may be helpful: Given several equations with high descriptive power, the simplest model should be accepted. In any case, one should have a sufficient number of observations per variable. The value recommended in the literature is a ratio of about 5:1 for medium-sized problems. For small series, however, a ratio of 3:1 may also be acceptable to get a first orientation, and for large datasets, higher ratios are recommended.

A very important point is that the resulting model must be interpretable and consistent with general experience from physical organic chemistry and QSAR work. This aspect is at least as important as statistical criteria. For example, equations with unrealistic regression coefficients must be rejected even if the statistics seem to be acceptable. A process called *lateral validation* by Hansch[10,30,43–45] is a very helpful step to assess the validity of a QSAR equation. In this process, the equation is systematically compared with known QSARs obtained for the same (or a similar) biological target and with known linear free-energy relationships for chemical reactions, if such reactions are suspected to be involved in the biological mechanism of action. As Hansch has put it, "Statistics alone … cannot prove a model … the best test of a model is — does it make sense with our current knowledge of QSARs in chemistry and biology?"[10] Another criterion is that the final equations must be stable in subsets of the training series.

In commercial programs for regression analysis, automated algorithms for deriving an equation from a set of variables are usually included where variables are added and removed in a stepwise procedure guided by statistical criteria. This is not a method to be recommended (particularly when colinearities occur between the variables), as the result often is ambiguous.

We are always faced with the temptation to use available software packages to compute all kinds of parameters that these packages can provide (including, for example, topological indices and a variety of quantum chemical parameters) without much thought about interpretation and selection of the best equation. This is a practice

not to be recommended for Hansch analysis. The number of parameters could exceed the number of observations, and many colinearities between parameters will exist. Sometimes and for certain purposes such a situation cannot be avoided (large and diverse datasets) but in such cases methods other than simple multiple regression analysis are required (see next section). If Hansch analysis is still to be used, the derivation and evaluation of equations are almost entirely based on statistical criteria and approaches as almost no chemical judgment can be introduced; this is a big disadvantage. In order to make such problems manageable, variable selection procedures, for example, cluster significance analysis[59] and genetic[60] and evolutionary[61] algorithms, have been used to find the best (in a statistical sense) equations automatically. The interpretability of the results from such analyses is, at best, very limited.

Hansch analysis is strictly limited to congeneric series of compounds. In such series, the compounds are supposed to be sufficiently similar to be described and compared by the usual hydrophobic, electronic, and steric descriptors within the context of the parameters selected and to have the same biological mechanism of action. Experience has shown that the concept of congenericity can be stretched very far. An example is a QSAR analysis of antimalarials where a large and diverse set of over 600 compounds containing 60 different aromatic/heterocyclic ring systems substituted with a variety of substituents could successfully be described by a relatively simple Hansch equation.[62] The problem is that congenericity very much depends on the biological activity considered and the type of descriptor variables involved. For unspecific biological effects, potency frequently depends only on hydrophobicity, and compounds may behave as congenerics which are not similar at all from a chemical point of view. On the other hand, a family of structurally similar compounds is not necessarily congeneric. A case in point is provided by various phenols acting as growth inhibitors in leukemia cells.[44] Here, two different QSARs are obtained for compounds with electron-withdrawing substituents (Equation (1.41)) and electron-releasing substituents (Equation (1.42)):

$$\log 1/C = 0.62(\pm 0.16)\log P + 2.35(\pm 0.31) \qquad (1.41)$$
$$n = 15,\ r = 0.919, \text{ and } s = 0.232$$

$$\log 1/C = -1.58(\pm 0.26)s^{+} + 0.21(\pm 0.06)\log P + 3.10(\pm 0.24) \qquad (1.42)$$
$$n = 23,\ r = 0.948, \text{ and } s = 0.191$$

The result was rationalized by assuming that the phenols act via two different mechanisms. Electron-releasing substituents support the abstraction of H-radicals which is supposed to be the key step for compounds described by Equation (1.42) (relationships with the enhanced substituent constant s^{+} are typical of radical reactions). This process is blocked by electron-withdrawing substituents so that the s^{+} term disappears (Equation (1.41)), resulting in a nonspecific toxic effect characterized by log *P*.

Hansch equations cannot be taken as causal relationships, but they extract and systematize information of data resulting in hypotheses that can be put to experimental test. They are, thus, an aid to moving in a promising direction, and they can also tell when to stop further structural variations if the optimum is already obtained.

The most important thing is to derive conclusions via interpretation. Unfortunately, there are many pitfalls in deriving, evaluating, and interpreting Hansch equations (and other QSARs) so that extensive experience in the field is a decisive factor. Strictly speaking, the prediction of how new compounds will behave in the biological system is limited to the parameter space spanned by the training series, but even this is an enormous benefit in multidimensional space. Some extrapolation, however, is usually possible. In addition, it must be stressed that QSAR analyses cannot create new information but can only extract information that is present in the available data. This requires, for example, sufficient variation in both biological activity and physicochemical parameters within the training series to be analyzed and biological activity values evenly distributed in physicochemical parameter space. The best way to achieve that is to apply series design methods with the objective of obtaining a maximum of information with a minimum of compounds in the training series (see, for example, Franke,[7] Pleiss and Unger,[63] and Austel[64]).

A special case of QSAR relationships is quantitative activity–activity relationships including structure–selectivity relationships. Such relationships are of growing importance. Typical cases are the separation of desired and undesired effects or comparisons of experimental results from different levels of biological integration (e.g., isolated enzyme/cell/organ/animal) including proper expressions with physicochemical parameters to account for transport processes (see, for example, Kubinyi,[2] Franke,[7] and Ford et al.[16]).

1.5 SOME MULTIVARIATE METHODS

1.5.1 Principal Components and PLS

It was already mentioned that for diverse and difficult to parameterize compounds, it may be necessary to collect a large number of chemical descriptor variables. This is particularly true if, in addition, no hypotheses are available to aid in descriptor selection. Many colinearities are to be expected, and the number of descriptors can exceed the number of biological observations. Clearly, multiple regression analysis cannot be applied in such cases. In order to understand such data in their entirety and to adequately deal with their mathematical properties, methods of multivariate statistics such as principal component analysis are required. Their main objectives are to display multidimensional data in a space of lower dimensionality with a minimum loss of information and to extract basic features behind the data.

If $\mathbf{X}$ is the descriptor matrix with n chemical compounds in the rows and m descriptor variables in the columns, principal component analysis splits $\mathbf{X}$ up into two new matrices, $\mathbf{A}$ and $\mathbf{P}$, so that $\mathbf{X}$ is reproduced within residual error (experimental + model error) according to:

$$(\mathbf{X})_{n,m} = (\mathbf{P})_{n,k}\,(\mathbf{A})_{k,m} + \text{error} \tag{1.43}$$

with the smallest possible k.

$\mathbf{P}$ is called the *principal component matrix* (or *score matrix*) and contains all information about the compounds, while the so-called loading matrix $\mathbf{A}$ represents

all information about the variables. The columns of **P** are called *principal components* (PCs), and the elements of **P** are the so-called scores. Thus, the principal components replace the variables in the original data matrix. The elements of the loading matrix **A** are called *loadings*. Each variable has a loading for each component. The loading a_{ij} ($i = 1, \ldots, k$; $j = 1, \ldots, m$) is a measure of the contribution of the ith PC to the jth variable: a high value of a_{ij} indicates a high importance of the ith PC for the jth variable (the jth variable is said to be highly loaded in the ith PC). Variables with high loadings in the same principal component are similar (correlated).

The principal components are extracted from the correlation matrix of the standardized variables (in this form, the variables have a mean of zero and unity variance) by a mathematical standard procedure in a stepwise manner in such a way that the first component extracts the largest and the last component the smallest part of the data variance. The number of relevant components, k, can be determined by several criteria. It should be as small as possible to reproduce the matrix **X** within experimental error. If colinearities between the variables occur, this number will always be considerably smaller than the number of columns in **X**. As a result, a reduction of dimensionality is obtained (less components than variables), and as the PCs are derived as orthogonal vectors, the colinearity problem is also eliminated in a mathematical sense. The principal components can now be used as variables in QSAR analyses. They are then called *latent variables* or *principal properties*, if a sufficiently large parameter space has been considered for a representative group of compounds (e.g., amino acids).

The use of PCs as independent variables in multiple regression analysis is called *principal component regression*. Today, the so-called PLS (partial least squares) method has become much more important than this technique.[65–67] PLS is also based on principal component analysis and has turned out to be a very efficient and robust method for large datasets. In the most general case, the objective is to describe a matrix of data from different biological tests (matrix **Y** with compounds in the rows and the tests in the columns) in terms of the descriptor matrix **X**. To this end, PCs are derived from both matrices in such a way that they yield an optimal description of **X** and **Y** while, at the same time, the PC pairs $\mathbf{P_k(Y)}$ (kth PC extracted from **Y**) and $\mathbf{P_k(X)}$ (kth PC extracted from **X**) are maximally correlated according to

$$\mathbf{P_k(Y)} = b_k\mathbf{P_k}(\boldsymbol{X}) + h_k \tag{1.44}$$

where b_k is the regression coefficient, and h_k is a residual.

Special algorithms are available to achieve that goal. Cross-validation (see above) is used to estimate the number of relevant components and to check for the validity of the resulting model. In this process, PCs are added step by step until the statistical cross-validation parameters are optimal. The loadings of the **X** PCs will give an impression of which of the original variables are related with which PCs. If biological potencies for new compounds are to be predicted, the descriptor values of these compounds are fitted to the PC model of **X**, leading to new values of the **X** PCs and a residual. If this residual is of the same magnitude as, but not greater than, that of the training series, prediction is possible. New values of the $\mathbf{P_k(Y)}$ are then computed from Equation (1.44), and from these and the PC model of the matrix

Y, biological activity data are obtained. In contrast to Hansch analysis, predictions are strictly limited to the parameter space spanned by the training series; no extrapolation is possible. PLS can also be applied if, instead of a **Y** matrix, only results from one biological test are to be analyzed; in fact, the majority of PLS publications relate to this situation.

The result is very sensitive against noise in the data; strong relationships with single descriptor variables of high relevance for biological potency may completely be obscured by irrelevant variables. For this reason procedures for the elimination of irrelevant descriptors have been developed, such as, for example, GOLPE.[68–71] As is also true for Hansch analysis, PLS works best with a well-designed training series. Special series design methods based on factorial or D-optimal design in principal properties have been proposed for this purpose.[72–74] Design in principal properties works well in connection with PLS but cannot be recommended if multiple regression analysis is to be applied. If one goes back to the individual original variables, the series designed for aromatic substituents using principal properties have low information content, and colinearities as well as multicolinearities exist.[75]

Partial least squares models can be transformed to regression coefficients for the original variables in the **X** matrix resulting in relationships looking like a Hansch equation. Such relationships are, of course, not true regression equations and, in the typical case of colinear descriptor variables, these coefficients are not independent and therefore not individually interpretable. Collinear variables will occur together so that a decision cannot be made regarding what are the true effects on biological potency. Thus, PLS results are of only limited interpretability. As stated by Hansch, "The price one pays for this approach is that … the results cannot be related to mechanistic physical organic or biochemistry as these subjects are now understood."[30] In addition, the results depend on technical details such as, for instance, scaling of variables, variable selection, type of cross-validation, choice of statistical criteria for model selection,[67] so that different PLS programs may lead to different results. This renders a lateral validation (systematic comparison of QSARs; see section on Hansch analysis) of PLS models impossible. For all these reasons, multiple regression analysis is the method of choice for datasets that are not too large and do not have too many variables, provided that it is applied with the necessary care and experience. For large sets of collinear variables as occur, for example, in comparative molecular field analysis (CoMFA; see below), PLS is the only choice and has turned out to be a powerful and effective method.

Continuum regression is a method that contains PLS, principal component regression, and multiple linear regression analysis as limiting cases.[76] By selecting values of an adjustable internal parameter (which can be optimized during calculation), it is possible to optimally adjust this method to the properties of the data to be analyzed.

Another aspect of principal components should be mentioned briefly in regard to the **Y** matrix. With the help of principal components derived from a matrix of compounds measured in a set of tests, the basic effects behind the biological tests may be separated. Such principal components can then replace biological potency in Hansch analysis leading to QSARs for these effects even though these effects have not directly been measured. An example is provided by the work of Seydel

and colleagues[77] on the antibacterial effect of sulfones and sulfonamides. Two principal components were derived from measurements in seven cell-free enzyme extracts and in two whole-cell systems. For these PCs, the following relationships were obtained:

$$PC1 = -7.02(\pm 1.25)\Delta ppm(NH_2) + 1.81(\pm 0.42)f_i - 0.93(\pm 0.19) \quad (1.45)$$
$$n = 17,\ r = 0.969, \text{ and } s = 0.264$$

$$PC2 = 1.40(\pm 0.52)\log k\text{¢} - 3.49(\pm 1.32)\log[0.098(\pm 0.173)k\text{¢} + 1] + 0.51(\pm 0.73) \quad (1.46)$$
$$n = 17,\ \log k\text{¢}_O = 0.834,\ r = 0.934, \text{ and } s = 0.396$$

The first component obviously reflects intrinsic activity at the active site of the enzyme and can be related to electronic parameters expressed as the relative chemical shift ($\Delta ppm(NH_2)$) of the protons of a NH_2-group present in the molecules and the fraction ionized, f_i. The second component shows a bilinear dependence on hydrophobicity expressed by the high-performance liquid chromatography (HPLC) parameter k¢ typical for transport processes. Thus, principal component analysis has led to simultaneous QSARs for intrinsic activity at the target and for the transport phenomena occurring at the cell membrane. A similar separation can be reached by principal component analysis of time series. From a data matrix with measurements of the analgesic potency of fentanyl derivatives in rats at ten different times, two significant factors* were obtained[78] representing pharmacokinetic processes and receptor affinity. As expected, the first factor representing pharmacokinetic processes shows a parabolic relationship with log P typical for adsorption/distribution processes, and the second factor, receptor affinity, could be described by a highly significant Free–Wilson model.

1.5.2 Three-Dimensional QSAR

Three-dimensional QSAR is actually outside the scope of this chapter but will be discussed very briefly because of the increasing importance[79] of such methods. The objective is to derive QSARs for drug–receptor interactions taking into account the three-dimensional structure of the drugs; pharmacokinetic aspects cannot be considered. Comparative molecular field analysis (CoMFA)[80–83] is the most commonly used approach in this area. In CoMFA, the molecules of the training series are placed into a grid following a predefined rule (alignment). This is a critical step especially for flexible molecules. Properties (probes) are then assigned to the grid points (e.g., CH_3, H^+), and the interaction energy with each grid point is then computed for every molecule. This results in a **X** matrix with thousands of columns (one column for each type of interaction energy in each grid point) which is then analyzed by PLS (see above). The resulting PLS model can be used to estimate interaction energies

* Factors are similar to PCs. The only difference between principal component and factor analysis is that in factor analysis only the variance in the so-called *common factor space* is considered. For highly correlated variables, PCs and factors are nearly identical.

for further molecules and to color-code the grid points with respect to the type and intensity of their interaction with the molecules. Even though the resulting contour map is not a true receptor map, it can provide valid information of the true structure of the binding site.[84] Comparisons between the results from CoMFA and classical Hansch analysis have been made by Kim.[85] A number of cases are presented for which Hansch analysis and CoMFA results have led to the same conclusions. The following example can demonstrate this point. For the catalytic step in the enzymatic hydrolysis of pyridyl hippurates, the following QSAR models are obtained (no confidence intervals given):

- Hansch analysis

$$\log k_{cat} = 0.45\mathrm{s} + 0.16 \qquad (1.47)$$
$$n = 13,\ r = 0.929, \text{ and } s = 0.093$$

- *CoMFA*

$$\log k_{cat} = 0.02Z1(\mathrm{H}^+) + 0.02Z2(\mathrm{H}^+) + 0.09Z3(\mathrm{H}^+) + 1.17 \qquad (1.48)$$
$$n = 13,\ r = 0.960,\ s = 0.071, \text{ and } s_{\text{cross-validated}} = 0.109$$

$Z1(\mathrm{H}^+)$, $Z2(\mathrm{H}^+)$, and $Z3(\mathrm{H}^+)$ are the first three PLS components with a H^+ probe at the grid points (electrostatic interaction energies). Both models come to the conclusion that electronic properties of substituents play the most important role, but the Hansch equation is much simpler, more straightforward to interpret, and certainly computationally much easier to obtain. The same is also true for the other examples presented in Kim.[85] This does not mean, of course, that CoMFA is not necessary, as CoMFA can handle structural variations that cannot be treated by Hansch analysis, and considering the three-dimensional structure in CoMFA adds an extremely important new quality. What it does mean, however, is that it is always worthwhile to start with a simple Hansch analysis in a first step whenever this is possible. The results can then aid in subsequent CoMFA analysis, if necessary, to derive hypotheses on the type of fields to be considered and how to align the molecules. In this context it should again be mentioned that conclusions on the mode of binding for a variety of ligands interacting with several enzymes derived from Hansch equations have later been verified by x-ray crystallography.[57] This not only is an argument for the validity of the Hansch approach but also supports the suggestion that such equations (or results from other classical QSAR methods) can be used to aid in the development of CoMFA models.

The area of three-dimensional QSAR is under steady and rapid development leading to improvements of the CoMFA technology but also to the development of alternative methods such as, for example, CoMSIA (comparative molecular similarity analysis) and CoMMA (comparative molecular moment analysis). In CoMSIA,[86] three-dimensional structures of the molecules are aligned as in CoMFA, but instead of the interaction energies at predefined grid-points, similarity indices related to steric, electrostatic, and hydrophobic potentials are calculated between all pairs of molecules. The resulting similarity matrix can then be analyzed in a GOLPE/PLS

procedure. The critical alignment step is possibly avoided in the CoMMA approach, where descriptors are used that describe shape and charge distribution.[87] Neuronal nets and genetic algorithms have also been used in three-dimensional QSAR.

1.5.3 Classification Methods

Quite frequently biological properties present themselves in the form of a classification with respect to either the strength (biological activity) or the type (e.g., agonist/antagonist) of an effect. Classification methods (frequently also referred to as *pattern recognition methods*) aim at deriving mathematical expressions (classifiers) in terms of molecule descriptors to describe the distribution of compounds over the respective classes. The most typical case is a classification with respect to biological potency, if biological measurements are not precise enough to present biological potency on a continuous scale. Once a classifier is known, it is possible to assign new compounds to the classes (classification) which amounts to a prediction of their biological properties. The most frequently applied classification methods in QSAR work are non-elementary discriminant analysis[7] and the simple classification analysis (SIMCA) method,[88,89] which will be discussed briefly. For the sake of simplicity, a case with two classes (e.g., biologically active vs. biologically inactive compounds) will be considered, although, in principle, multiclass problems can be treated in the same way.

In the first step of discriminant analysis, descriptor variables that are supposed to be related to the distribution of compounds over the classes are collected. A so-called *discriminant function*, w, is then derived which is of the general form:

$$w = a_0 + a_1 x_1^{n1} + \ldots + a_i x_i^{ni} + \ldots + a_n x_n^{nn} \tag{1.49}$$

The same variables and functional relationships as in Hansch analysis are used, and all that has been outlined for Hansch analysis is also true for discriminant analysis. The coefficients a_i are so determined that the separation of classes is optimal. This is done by solving a special eigenvalue problem. As redundant variables (variables that do not contribute to the separation of classes) are a disturbing factor, they are eliminated prior to calculation of the discriminant function by multivariate variance analysis in a stepwise procedure. In a two-class case, the discriminant function w can be visualized as the axis of a one-dimensional coordinate system with the two classes occupying different regions (see Figure 1.7). The further these regions are apart, the better is the separation of classes achieved by the respective discriminant function.

Examples for discriminant functions are Equations (4.10) to (4.13) in Chapter 4. As follows from the discussion presented there, discriminant functions can be interpreted in much the same way as Hansch equations. The conclusions are, of course, less precise, as the information content of classified data is smaller than that of a continuous quantity. When interpreting a discriminant function, the positions of the classes on the w-axis must be known and are usually expressed in terms of the class-means of w. In Figure 1.7, $w_{(\text{mean, class 1})}$ is greater than $w_{(\text{mean, class 2})}$. In such cases, variables with positive coefficients will increase the probability that a com-

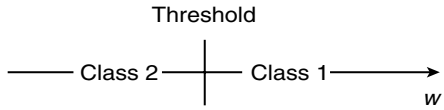


FIGURE 1.7 The discriminant function spans the axis of a one-dimensional coordinate system.

pound belongs to class 1 if the variables have a positive value. If, however, $w_{(\text{mean, class 1})}$ is smaller than $w_{(\text{mean, class 2})}$, the reverse is true.

Once a discriminant function is known, a compound can be classified by computing the value of w for this compound by inserting the values of the respective descriptor variables into the discriminant function. If the classes are placed on the w-axis as in Figure 1.7, then a compound will be assigned to class 1 if the computed value is greater than a certain threshold (usually, this decision is made by means of a special F test).

The quality of a discriminant function can be judged by reclassifying all compounds of the training series. For an acceptable discriminant function, the error of reclassification should be <20% (>80% of the compounds are assigned to the correct class). If enough measurements are available, the compounds can also be divided into two sets. One is used to derive the discriminant function (learning set), and the other one serves as test set, the compounds of which are then classified. The result characterizes the predictive power of the discriminant function; the error of classification is usually somewhat higher than the error of reclassification. A final possibility is cross-validation, where compounds are left out from the analysis and are then classified by the discriminant function derived from the other compounds. Cross-validated errors of classification are also usually higher than reclassification errors. As in Hansch analysis, however, cross-validation is not a safe criterion to reject a discriminant analysis result.

Colinear variables provide the same problems as in Hansch analysis, and similar strategies to deal with such cases are to be used. In no case should a discriminant function be derived from a set of variables containing many colinearities, as spurious results are then to be expected. Discriminant functions can provide insight into the mechanism of action via interpretation, and they can predict the class membership of new compounds within spanned substituent space. It is not possible, however, to tell whether the new compounds will be more (or less) active than those already belonging to the class considered. If possible, the selection of compounds for the training series by series design techniques (same methods as for Hansch analysis) is recommended.

Discriminant analysis can only be applied in the so-called symmetric case where the classes are clearly separated in the parameter space considered; however, sometimes this is not true. The active class can, for example, be imbedded into a scatter of points representing the inactive class (asymmetric case: only one class has a clear structure). This may happen because a compound can be inactive for many different reasons. Such problems can be handled by simple classification analysis (SIMCA). SIMCA is based on principal components and is closely related to PLS. In the first step, parameters as in Hansch or discriminant analysis are collected. Irrelevant

parameters can be eliminated by various procedures including GOLPE (see Section 1.5.1). Then, a principal component model is built for each class separately. With the help of the principal component models, the classes are represented as a kind of hyperboxes in parameter space. For classification, compounds are fitted to these hyperboxes and assigned to that box (class) for which the fit is best. In contrast to discriminant analysis, the principal component models allow estimation of the position of individual compounds within the classes. Because of the principal component analysis step, SIMCA has no problem handling large numbers of collinear variables, which is not possible with discriminant analysis.

A comparison of discriminant analysis and SIMCA is similar to the comparison of Hansch analysis and PLS. SIMCA is mathematically more powerful, while results of discriminant analyses lend themselves to a more straightforward interpretation and, to a certain extent, lateral validation. For medium-sized datasets with chemical structures that do not present problems in parameter selection, discriminant analysis is recommended. For larger datasets and in situations where a larger number of colinear variables must be considered, SIMCA is to be preferred. For the asymmetric case, discriminant analysis cannot be applied.

A variety of other classification methods has also been used in QSAR work such as, for example, adaptive least squares and fuzzy adaptive least squares[90] or non-parametric techniques such as the linear learning machine[7] or the kNN-method.[7,91] A complete list would be outside the scope of this chapter. Even though classification methods can handle less precise biological measurements, they are still restricted to congeneric series, and their "abuse to correlate and predict global toxic, mutagenic, teratogenic, carcinogenic, and other biological properties must be criticized."[2]

1.6 SOME OTHER QSAR-RELATED METHODS

To demonstrate the complexity of the field, some additional QSAR-related methods will be listed in this section without attempting completeness. Artificial neural networks (ANNs) simulate the functioning of human neurons and have found fairly wide application for several drug design problems. [92] After training with a training set they can predict properties of new compounds. In comparison with the classical statistical QSAR methods, they offer advantages and disadvantages (for a critical discussion, see Manallak and Livingstone [93] and Livingstone and Ford [94]). Advantages are that large numbers of variables, colinearities between them, and nonlinearities do not present a technical problem. In addition, it is not necessary to specify the functional form of a relationship. A real disadvantage is that the results are very difficult to interpret. Some authors have reported that a better statistical fit can be obtained with ANNs than with multiple linear regression analysis (see references in Lui and Trinajsti [95]), while the reverse has also been reported. [95,96] A good description of data by ANN models is frequently accompanied by overfitting with low predictive power as a consequence. It seems that, for problems typical of the classical statistical QSAR methods, no reason exists to replace these methods by ANNs. ANNs can be useful for special purposes or if large and diverse datasets are to be analyzed. Examples are the mapping of molecular surface properties, [97] the analysis of CoMFA fields, [98] and the prediction of "drug-likeness" of molecules from chemical structure. [99–101]

Genetic algorithms have already been mentioned. Based on the mechanism of Darwinian evolution, a genetic algorithm will breed better models or solutions from an originally random starting sample by random mutation, crossover, and selection procedures. An introduction and a good review of the application of genetic algorithms in drug design are presented by Devillers.[102] Genetic algorithms cannot replace statistical QSAR methods but can be used as an instrument to support them. They have been used, for example, for variable selection, for series design, and in combination with PLS. An important field of application is molecular modeling. Frequently, genetic algorithms are linked with neural networks.

Knowledge-based expert systems provide another possibility to rationalize structure–activity relationships. A powerful method is the CASE program developed by Klopman[103] that was later modified into the improved MULTICASE approach.[104] Starting from a learning set of structurally diverse compounds, these approaches automatically identify substructures that have a high probability of being responsible for or related with an observed biological activity. Inputs are chemical structures in KLN code and biological activity in the form of a classification. The substructures are found as biophores (substructures essential for biological activity) and as modulators (substructures capable of modifying the effect of biophores). An expert prediction of the activity of new compounds can then be obtained. This approach has been applied to various types of activity including, for example, toxicity;[105] however, predictions must be made with care as the results depend very much on the properties of the learning set. They cannot replace experimental values but can be used to rank and prioritize chemicals for evaluation.

Another method that is supposed to be capable of analyzing large and diverse datasets also based on substructures is hologram QSAR (HQSAR).[106] In this method, fragments of adjustable length are automatically generated, and a matrix is built with the fragments in the columns, the compounds in the rows, and the occurrence number of each fragment for each compound as elements. This matrix is then submitted to PLS analysis, resulting in activity contributions for the fragments. These contributions are assumed to behave in an additive manner (similar assumption as in Free–Wilson analysis) so that the potency of new compounds possessing these fragments can be estimated. Several successful applications have been reported (see, for example, Pungpo et al.[107]); however, the examples considered so far relate to datasets of fairly limited size and structurally similar compounds. What this method can achieve for real large and diverse series still remains to be determined.

A number of QSAR approaches start from a parent structure with which the molecules of the training series are compared. This parent structure can be an artificial hyperstructure that is so defined that it includes all structural features of the molecules to be analyzed. In the minimum steric difference (MSD) method, followed by the minimum topological difference (MTD) method,[108,109] the assumption is made that receptor affinity decreases linearly with steric misfit. Steric misfit is defined as the receptor cavity volume not occupied by the drug molecules plus the volume of the molecules falling into the walls of the (rigid) receptor cavity. MSD or MTD values are taken as measures of this misfit and are derived from superposing all molecules over an artificial hypermolecule, which, in turn, has been derived by superposition of all molecules of the series. A fairly complex iterative technique that is not without

ambiguities is then used to derive the MSD or MTD values. The resulting values can then be used as steric descriptors in Hansch analysis.

Philosophically similar but technically much more advanced is molecular shape analysis,[110,111] which takes into account conformational flexibility. Molecules in their minimum energy conformation are compared with a reference compound, and the common steric overlap volume is then used as a QSAR descriptor. Resulting QSARs have shown predictive power. In the methods LOGANA, LOCON, and EVAL,[112–115] topological pharmacophores are derived. Starting from an artificial hypermolecule, a library of substructures (potential centers of interaction) is derived. Each compound is then superposed over the hyperstructure and described in terms of the presence or absence of these substructures by means of Free–Wilson type descriptors. The descriptor variables are then combined into more complex expressions in a stepwise procedure using logical operations (e.g., and, or, not). Each combination of variables represents a pattern of substructures that becomes more complex with each step and is, thus, present in fewer compounds. That means that compounds are eliminated in each step, and the process is so organized that the (highly) active compounds are retained. The resulting patterns are thus characteristic of (high) activity and are called *topological pharmacophores*. The methods can be applied to very diverse datasets. Physicochemical parameters can be included after transformation into binary variables. The selection of meaningful features is crucial.

Another method also based on special substructural descriptors is the PASS method.[116,117] These descriptors are derived from two-dimensional chemical structures in a recursive sequence. Trained with 30,000 compounds representing 500 different biological activities, the PASS algorithm aims at predicting pharmacological profiles for new structures.

The last method to be mentioned is the VolSurf approach introduced by Cruciani.[118,119] VolSurf is a three-dimensional technique that avoids solving alignment problems, the most difficult and time-consuming steps in CoMFA and related approaches. VolSurf compresses the information obtained from the interaction of molecules with GRID points into simple quantitative descriptors using a holistic transformation. VolSurf descriptors characterize size, shape, polarity, and hydrophobicity and are relatively independent of conformational sampling. They can be used as variables in statistical QSAR analysis.

1.7 CONCLUDING REMARKS

In this chapter, an attempt was made to outline some important aspects of QSAR methods with an emphasis on the classical statistical approaches. As the drug discovery process is of a very complex nature, effective drug design requires an entire spectrum of techniques in which QSAR methods still play an important role. It must always be realized that drug design models, and QSAR results in particular, do not represent causal relationships so that a very careful evaluation and interpretation are absolutely essential. The real power of drug design methods is to extract and systematize information from data to obtain hypotheses that can be put to experimental test. No dramatic overnight discoveries of wonder drugs will result, but an increase in the chance of success due to indications of promising directions is a realistic

expectation. A very close interaction with experimental work is a key factor. As stated by Kubinyi,[2] "QSAR cannot and will never substitute the creativity and intuition of an experienced medicinal chemist or biologist." It can be regarded, however, as an amplifier of human intelligence. Drug design methods have limitations and pitfalls. Thus, an exact knowledge of applicability and access to the entire toolbox of methods is a prerequisite to making drug design successful. One of the achievements of drug design has been to contribute to the development of science in medicinal chemistry. In this respect, interpretability and the systematic comparison of QSARs (lateral validation) are of the utmost importance.

REFERENCES

1. Martin, Y.C., Kim, K.-H., and Bures, G.M., Computer assisted drug design in 21st century, in *Medicinal Chemistry for the 21st Century*, Wermuth, C.G. et al., Eds., Blackwell Scientific, London, 1992, chap. 20.
2. Kubinyi, H., *QSAR: Hansch Analysis and Related Approaches*, VCH, Weinheim, 1993.
3. Böhm, H.-J., Klebe, G., and Kubinyi, H., *Wirkstoffdesign*, Spektrum Akademischer Verlag, Heidelberg, 1996.
4. Martin, Y.C, *Drug Design Methods: A Critical Introduction*, Marcel Dekker, New York, 1978.
5. Seydel, J.K. and Schaper, K.-J., *Chemische Struktur und biologische Aktivität von Wirkstoffen*, VCH, Weinheim, 1979.
6. Franke, R., *Optimierungsmethoden in der Wirkstofforschung*, Akademie-Verlag, Berlin, 1980.
7. Franke, R. *Theoretical Drug Design Methods*, Elsevier, Amsterdam, 1984.
8. Ramsden, C.A., Ed., Quantitative drug design, in *Comprehensive Medicinal Chemistry*, Vol. 4, Hansch, C. et al., Eds., Pergamon Press, Oxford, 1990.
9. Draber, W. and Fujita, T., Eds., *Rational Approaches to Structure, Activity, and Eco Toxicology of Agrochemicals*, CRC Press, Boca Raton, FL, 1992.
10. Hansch, C. and Leo, L., *Exploring QSAR*, American Chemical Society, Washington, D.C., 1995.
11. van de Waterbeemd, H., Ed., *Chemometric Methods in Molecular Design*, VCH, Weinheim, 1995.
12. van de Waterbeemd, H., Ed., *Advanced Computer-Assisted Techniques in Drug Discovery*, VCH, Weinheim, 1995.
13. Hansch, C. and Fujita, T., Eds., *Classical and Three-Dimensional QSAR in Agrochemistry*, American Chemical Society, Washington, D.C., 1995.
14. Livingstone, D.J., *Data Analysis for Chemists*, Oxford University Press, Oxford, 1995.
15. Herrmann, E.C. and Franke, R., Eds., *Computer Aided Drug Design in Industrial Research*, Springer-Verlag, Berlin, 1995.
16. Ford, M.G. et al., Eds., *Bioactive Compound Design: Possibilities for Industrial Use*, Bios Scientific Publishers, Oxford, 1996.
17. Pliska, V., Testa, B., and van de Waterbeemd, H., Eds., *Lipophilicity in Drug Action and Toxicology*, VCH, Weinheim, 1996.
18. Kubinyi, H., Ed., *3D QSAR in Drug Design*, ESCOM, Leiden, 1993.
19. Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., *3D QSAR in Drug Design*, Vol. 2, Kluwer/ESCOM, Dordrecht, 1998.

20. Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., *3D QSAR in Drug Design*, Vol. 3, Kluwer/ESCOM, Dordrecht, 1998.
21. Testa, B. et al., Eds., *Pharmacokinetic Optimization in Drug Research*, Wiley-VCH, Weinheim, 2001.
22. Sanz, F., Giraldo, J., and Manaut, F., *QSAR and Molecular Modeling: Concepts, Computational Tools, Applications*, Prous Science, Barcelona, 1995.
23. van de Waterbeeemd, H., Testa, B., and Folkers, G., Eds., *Computer-Assisted Lead Finding and Optimization*, Wiley-VCH, Weinheim, 1997.
24. Gundertofte, K. and Jørgensen, F.S., Eds., *Molecular Modeling and Prediction of Bioactivity*, Kluwer/Plenum Press, New York, 2000.
25. Höltje, H.-D. and Sippl, W., Eds., *Rational Approaches to Drug Design*, Prous Science, Barcelona, 2001.
26. Hansch, C. and Fujita, T., r-s-p Analysis: a method for the correlation of biological activity and chemical structure, *J. Am. Chem. Soc*., 86, 1616, 1964.
27. Free, Jr., S.M. and Wilson, J.W., A mathematical contribution to structure activity studies, *J. Med. Chem*., 7, 395, 1964.
28. Todeschini, R. and Consonni, V., *Handbook of Molecular Descriptors*, Wiley-VCH, Weinheim, 2000.
29. Hansch, C., Leo, A., and Taft, W., A survey of Hammett substituent constants and resonance and field parameters, *Chem. Rev.*, 91, 165, 1991.
30. Hansch, C., Hoekman, D., and Gao, H., Comparative QSAR: toward a deeper understanding of chemicobiological interactions, *Chem. Rev.*, 96, 1045, 1996.
31. Taylor, P.J., Hydrophobic properties of drugs, in *Comprehensive Medicinal Chemistry*, Vol. 4, Hansch, C. et al., Eds., Pergamon Press, Oxford, 1990, chap. 18.6.
32. Franke, R., Kühne, R., and Dove, S., Dependence of hydrophobicity on solvent and structure, in *Quantitative Approaches to Drug Design*, Pharmacochem. Libr. 6, Dearden, J.C., Ed., Elsevier, Amsterdam, 1983, p. 15.
33. Ganellin, C.R. et al., Use of partition coefficients as a model for brain penetration applied to the design of H_2-receptor histamine antagonists, in *QSAR: Rational Approaches to the Design of Bioactive Compounds*, Pharmacochem. Libr. 16, Silipo, C. and Vittoria, A., Eds., Elsevier, Amsterdam, 1991, p. 103.
34. Seiler, P., Interconversion of lipophilicities from hydrocarbon/water systems into octanol/water system, *Eur. J. Med. Chem.*, 9, 473, 1974.
35. Seydel, J.K. and Wiese, M., *Drug–Membrane Interactions*, Wiley-VCH, Weinheim, 2002.
36. Leahy, D.E. et al., Model solvent systems for QSAR. 3. An LSER analysis of the critical quartet: new light on hydrogen bond strength and directionality, *J. Chem. Soc. Perkin Trans*., 2, 705, 1992.
37. Abraham, H.M. and Chadha, H.S., Applications of a solvation equation to drug transport properties, in *Lipophilicity in Drug Action and Toxicology*, Pliska, V., Testa, B., and van de Waterbeemd, H., Eds., VCH, Weinheim, 1996, p. 311.
38. Raevsky, O.A., Hydrogen bond strength estimation by means of the HYBOT program package, in *Computer-Assisted Lead Finding and Optimization*, van de Waterbeeemd, H., Testa, B., and Folkers, G., Eds., Wiley-VCH, Weinheim, 1997, p. 367.
39. van de Waterbeemd, H., Quantitative structure-absorption relationships, in *Pharmacokinetic Optimization in Drug Research*, Testa, B. et al., Eds., Wiley-VCH, Weinheim, 2001, p. 499.
40. Stenberg, P. et al., Experimental and computational screening models for the prediction of intestinal drug absorption, *J. Med. Chem.*, 44, 1927, 2001.

41. Österberg, T. and Norinder, U., Prediction of polar surface area and drug transport processes using simple parameters and PLS statistics, *J. Chem. Inf. Comput. Sci.*, 40, 1408, 2000.
42. Feher, M., Sourial, F., and Schmidt, J.M., A simple model for the prediction of blood–brain partitioning, *Int. J. Pharmac.*, 201, 239, 2000.
43. Leo, A. and Hansch, C., Role of hydrophobic effects in mechanistic QSAR, *Perspectives Drug Discov. Design*, 17, 1, 1999.
44. Hansch, C. et al., Chem-bioinformatics and QSAR: a review of QSAR lacking positive hydrophobic terms, *Chem. Rev.*, 101, 619, 2001.
45. Hansch, C. et al., Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology, *Chem. Rev.*, 102, 783, 2002.
46. Leo, A., The future of log P calculation, in *Lipophilicity in Drug Action and Toxicology*, Pliska, V., Testa, B., and van de Waterbeemd, H., Eds., VCH, Weinheim, 1996, p. 157.
47. Duban, M.E. et al., Virtual screening of molecular properties: a comparison of log P calculators, in *Pharmacokinetic Optimization in Drug Research*, Testa, B. et al., Eds., Wiley-VCH, Weinheim, 2001, p. 485.
48. Leo, A., Calculating log P_{oct} from structure, *Chem. Rev.*, 93, 1281, 1993.
49. Rekker, R.F. and Mannhold, R., *Calculation of Drug Lipophilicity*, VCH, Weinheim, 1992.
50. Franke, R. and Gruska, A., Principal component and factor analysis, in *Chemometric Methods in Molecular Design*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, chap. 4.1.
51. Verloop, A., *The STERIMOL Approach to Drug Design,* Marcel Dekker, New York, 1987.
52. Kier, L.B. and Hall, L.H., *Molecular Connectivity in Structure–Activity Analysis*, Wiley, Chichester, 1986.
53. Kier, L.B. and Hall, L.H., *Molecular Structure Description: The Electrotopological State*, Academic Press, London, 1999.
54. Martin, Y.C. and Lynn, K.R., Quantitative structure–activity relationships in leucomycin and lincomycin derivatives, *J. Med. Chem.*, 14, 1162, 1971.
55. Draper, N.R. and Smith, H., *Applied Regression Analysis*, 2nd ed., Wiley, New York, 1981.
56. Topliss, J.G. and Edwards, R.P., Chance factors in studies of quantitative structure–activity relationships, *J. Med. Chem.*, 22, 1238, 1979.
57. Selassie, C.D. and Klein, T.E., Building bridges: QSAR and molecular graphics, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 257.
58. Franke, R., unpublished results.
59. McFarland, J. and Gans, D.J., On identifying likely determinants of biological activity in high dimensional QSAR problems, *Quant. Struct.–Act. Relat.,* 13, 11, 1994.
60. Rogers, D. and Hopfinger, A.J., Application of genetic function approximation to quantitative structure–activity relationships and structure–property relationships, *J. Chem. Inf. Comput. Sci.*, 34, 854, 1994.
61. Kubinyi, H., Variable selection in QSAR studies: an evolutionary algorithm, *Quant. Struct.–Act. Relat.*, 13, 285, 1994.
62. Kim, K.H. et al., Quantitative structure-activity relationships in 1-aryl-2-(alkylamino)ethanol antimalarials, *J. Med. Chem.*, 22, 366, 1979.
63. Pleiss, M.A. and Unger, S.H., The design of test series and the significance of QSAR relationships, in *Comprehensive Medicinal Chemistry*, Vol. 4, Hansch, C. et al., Eds., Pergamon Press, Oxford, 1990, p. 561.

64. Austel, V., Experimmental design, in *Chemometric Methods in Molecular Design*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, chap. 3.1.
65. Wold, S., PLS for multivariate linear modeling, in *Chemometric Methods in Molecular Design*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, chap. 4.4.
66. Wold, S., Johansson, E., and Cochi, M., PLS: partial least-squares projections to latent structures, in *3D QSAR in Drug Design,* Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 523.
67. Kubinyi, H. and Abraham, U., Practical Problems in PLS analysis, in *3D QSAR in Drug Design,* Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 711.
68. Baroni, M. et al., An advanced chemometric tool for handling three-dimensional QSAR problems, *Quant. Struct. Act. Relat*., 12, 9, 1993.
69. Cruciani, G., Clementi, S., and Baroni, M., Variable selection in PLS analysis, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 551.
70. Cruciani, G. and Clementi, S., GOLPE: philosophy and application in 3D QSAR, in *Advanced Computer-Assisted Techniques in Drug Discovery*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, chap. 2.3.
71. Cruciani, G., Clementi, S., and Pastor, M., GOLPE-guided region selection, in *3D QSAR in Drug Design*, Vol. 2, Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., Kluwer/ESCOM, Dordrecht, 1998, p. 71.
72. Skagerberg, G. et al., Principal properties for aromatic substituents: a multivariate approach for design in QSAR, *Quant. Sruct.–Act. Relat*., 8, 32, 1989.
73. Baroni, M. et al., D-optimal design in QSAR, *Quant. Sruct.–Act. Relat.*, 12, 225, 1993.
74. Clementi, S., Series design, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 567.
75. Franke, R., unpublished results.
76. Malpass, J.A. et al., Continuum regression: a new algorithm for the prediction of biological potency, in *Advanced Computer-Assisted Techniques in Drug Discovery*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, p. 163.
77. Coats, E.A. et al., Multiple regression and principal component analysis of antibacterial activities of sulfones and sulfonamides in whole-cell and cell-free systems of various DDS sensitive and resistant bacterial strains, *Quant. Struct.–Act. Relat*. 4, 99, 1985.
78. Franke, R. and Gruska, A., Decomposition of time dependent response data by factor analysis, *Quant. Struct.–Act. Relat.*, 13, 184, 1994.
79. Martin, Y.C., 3D QSAR: current state, scope and limitations, in *3D QSAR in Drug Design*, Vol. 3, Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., Kluwer/ESCOM, Dordrecht, 1998, p. 3.
80. Cramer, R.D. et al., The developing practice of comparative molecular field analysis, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 443.
81. Folkers, G., Merz, A., and Rognan, D., CoMFA: scope and limitations, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 583.
82. Thibaut, U., Applications of CoMFA and related 3D QSAR approaches, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 661.
83. Norinder, U, Recent progress in CoMFA technology and related techniques, in *3D QSAR in Drug Design*, Vol. 3, Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., Kluwer/ESCOM, Dordrecht, 1998, p. 25.
84. Kim, K.H., Building a bridge between G-protein-coupled receptor modeling, protein crystallography and 3D QSAR studies for ligand design, p. 234 in Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., *3D QSAR in Drug Design*, Vol. 3, Kluwer/ESCOM, Dordrecht, 1998.

85. Kim, K.H., Comparison of classical and three-dimensional QSAR, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 619.
86. Klebe, G., Comparative molecular similarity analysis: CoMSIA, in *3D QSAR in Drug Design*, Vol. 3, Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., Kluwer/ESCOM, Dordrecht, 1998, p. 87.
87. Silverman, B.D. et al., Comparative molecular moment analysis (CoMMA), in *3D QSAR in Drug Design*, Vol. 3, Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., Kluwer/ESCOM, Dordrecht, 1998, p. 183.
88. Dunn, W.J. and Wold, S., Pattern recognition techniques in drug design, in *Comprehensive Medicinal Chemistry*, Vol. 4, Hansch, C. et al., Eds., Pergamon Press, Oxford, 1990, p. 691.
89. Dunn, W.J. and Wold, S., SIMCA pattern recognition and classification, in *Chemometric Methods in Molecular Design*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, p. 179.
90. Schaper, K.-J., Quantitative structure–activity–class relationships by (fuzzy) adaptive least squares, in *Advanced Computer-Assisted Techniques in Drug Discovery*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, p. 245.
91. Rose, V.S., Wood, J., and MacFie, H.J.H., Analysis of embedded data: k-nearest neighbor and single class discrimination, in *Advanced Computer-Assisted Techniques in Drug Discovery*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, p. 228.
92. Zupan, J. and Gasteiger, J., *Neural Networks for Chemistry and Drug Design*, Wiley-VCH, Weinheim, 1999.
93. Manallak, D.T. and Livingstone, D.J., Neural networks and expert systems, in *Advanced Computer-Assisted Techniques in Drug Discovery*, van de Waterbeemd, H., Ed., VCH, Weinheim, 1995, p. 293.
94. Livingstone, D.J. and Ford, M.G., Artificial networks as an alternative to statistics, in *Bioactive Compound Design: Possibilities for Industrial Use*, Ford, M.G. et al., Eds., Bios Scientific Publishers, Oxford, 1996, p. 99.
95. Lui, B. and Trinajsti, N., Multivariate regression outperforms several robust architectures of neural networks in QSAR modeling, *J. Chem. Inf. Comput. Sci.*, 39, 121, 1999.
96. Lui, B., Amic, D., and Trinajsti, N., Nonlinear multivariate regression outperforms several concisely designed neural networks on three QSPR data sets, *J. Chem. Inf. Comput. Sci.*, 40, 403, 2000.
97. Anzali, S. et al., The use of self-organizing neural networks in drug design, in *3D QSAR in Drug Design*, Vol. 2, Kubinyi, H., Folkers, G., and Martin, Y.C., Eds., Kluwer/ESCOM, Dordrecht, 1998, p. 273.
98. Tetko, I.V. et al., Application of volume learning artificial neural network to calculate 3D QSAR models with enhanced predictive properties, in *Rational Approaches to Drug Design*, Höltje, H.-D. and Sippl, W., Eds., Prous Science, Barcelona, 2001, p. 229.
99. Sadowski, J. and Kubinyi, H., A scoring scheme for discriminating between drugs and nondrugs, *J. Med. Chem.*, 41, 3325, 1998.
100. Frimurer, T.M. et al., Improving the odds in discriminating "drug-like" from "non-drug-like" compounds, *J. Chem. Inf. Comput. Sci.*, 40, 1315, 2000.
101. Sadowski, J., Optimization of the drug-likeness of chemical libraries, *Perspect. Drug Discovery Design*, 20, 17, 2000.
102. Devillers, J., *Genetic Algorithms in Molecular Modeling*, Academic Press, London, 1996.

103. Klopman, G., Artificial intelligence approach to structure–activity studies: computer automated structure evaluation of biological activity of organic molecules, *J. Am. Chem. Soc.*, 106, 7315, 1984.
104. Klopman, G., MULTICASE 1: a hierarchical computer automated structure evaluation program, *Quant. Struct.–Act. Relat.*, 11, 176, 1992.
105. Klopman, G. and Rosenkranz, H.S., Toxicity estimation by chemical substructure analysis: the Tox II program, *Toxicol. Lett.*, 79, 145, 1995.
106. Lowis, D.R., Hologram QSAR, *Tripos Tech. Notes*, 1(5), 1997.
107. Pungpo, P., Wolschann, P., and Hannongbua, S., Quantitative structure–activity relationships of HIV-1 reverse transcriptase inhibitors, using hologram QSAR, in *Rational Approaches to Drug Design*, Höltje, H.-D. and Sippl, W., Eds., Prous Science, Barcelona, 2001, p. 206.
108. Simon, Z. et al., *Minimum Steric Difference*, Research Study Press, Letchworth, England, 1984.
109. Simon, Z., MTD and hyperstructure approaches, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 307.
110. Hopfinger, A.J., A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based on molecular shape analysis, *J. Am. Chem. Soc.*, 102, 7196, 1980.
111. Burke, B.J. and Hopfinger, A.J., Advances in molecular shape analysis, in *3D QSAR in Drug Design*, Kubinyi, H., Ed., ESCOM, Leiden, 1993, p. 276.
112. Streich, W.J. and Franke, R., Topological pharmacophores: new methods and their application to a set of antimalarials. Part 1. The methods LOGANA and LOCON, *Quant. Struct.–Act. Relat.,* 4, 13, 1985.
113. Franke, R. and Streich, W.J., Topological pharmacophores: new methods and their application to a set of antimalarials. Part 2. Results from LOGANA, *Quant. Struct.–Act. Relat.,* 4, 51, 1985.
114. Franke, R. and Streich, W.J., Topological pharmacophores: new methods and their application to a set of antimalarials. Part 3. Results from LOCON, *Quant. Struct.–Act. Relat.,* 4, 63, 1985.
115. Hübel, S. and Franke, R., EVAL: a new tool to evaluate topological pharmacophores, *Software Dev. Chem.*, 5, 85, 1991.
116. Filimonov, D.A. et al., Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with other descriptors, *J. Chem. Inf. Comput. Sci*., 39, 666, 1999.
117. Poroikov, V.V. et al., Robustness of biological activity spectra predicting by computer program PASS for noncongeneric sets of chemical compounds, *J. Chem. Inf. Comput. Sci*., 40, 1349, 2000.
118. Clementi, S. et al., A new set of principal properties for heteroaromatics obtained by GRID, *Quant. Struct.–Act. Relat*., 15, 108, 1995.
119. Cruciani, G. et al., From molecular interaction fields (MIF) to a widely applicable set of descriptors, in *Rational Approaches to Drug Design*, Höltje, H.-D. and Sippl, W., Eds., Prous Science, Barcelona, 2001, p. 159.