MOLECULAR DESCRIPTORS IN QSAR/QSPR

MATI KARELSON
Department of Chemistry
University of Tartu

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Library of Congress Cataloging-in-Publication Data:

Karelson, Mati.
Molecular descriptors in QSAR/QSPR / by Mati Karelson.
p. cm.
"A Wiley-Interscience publication."
Includes bibliographical references and index.
ISBN 0-471-35168-7 (alk. paper)
1. QSAR (Biochemistry)—Mathematical models. I. Title.
QP517.S85K37 2000
572'.4—dc21 99-38911

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Tiia

CONTENTS

Preface ix

1 Introduction 1

2 Empirical Molecular Descriptors 13

2.1 Historical Background, 13
2.2 Structural Descriptors, 19
2.3 Solvational Descriptors, 66
References, 124

3 Theoretical Molecular Descriptors 141

3.1 Classification of Theoretical Molecular Descriptors, 141
3.2 Constitutional and Geometrical Descriptors, 142
3.3 Topological Descriptors, 160
3.4 Charge-Distribution-Related Descriptors, 220
3.5 Molecular Field Approach, 265
3.6 Quantum Chemical Descriptors, 275
3.7 Solvational Descriptors, 329
References, 354

4 Methods for Development of QSAR/QSPR 385

4.1 Introduction, 385
4.2 Linear and Multiple Linear Regression Methods, 386

4.3 Derivation of Best Multilinear Regression Models in Large Descriptor Spaces, 396
4.4 Nonlinear Methods, 404
References, 411

Appendix **415**

Computer Software for Calculation of Molecular Descriptors and Derivation of QSAR/QSPR

Index **423**

PREFACE

The revolutionary progress in computer technology has created an entirely new environment for efficient use of the theoretical constructions of natural science in many areas of applied research. The theoretical approach has proven to be most beneficial in chemistry and bordering sciences, where the experimental study and synthetic development of new compounds and materials is often time consuming, expensive, or even hazardous. The contemporary quantum theory of molecular matter and the respective ab initio computational methods can predict the properties of isolated small molecules within the experimental error. However, the majority of industrially and environmentally important chemical processes, and all biochemical transformations in living organisms, take place in heterogeneous condensed media. The extreme complexity of such media is often prohibitive for the ab initio theory to be used, and thus the relationship between the chemical activity and the molecular structure in these systems is often poorly described and understood.

The direct development of empirical equations that are usually referred to as the quantitative structure–activity/property relationships (QSAR/QSPR) represents an attractive alternative approach to predict the molecular properties in complex environments. Notably, the QSAR methodology has been extremely productive in pharmaceutical chemistry and in computer-assisted drug design. Probably thousands of new medications have been first developed on computer screens before their implementation in synthetic laboratories. In analytical chemistry, the QSPR equations are commonly used to predict the spectroscopic, chromatographic, and other analytical properties of compounds. In recent years, the QSPR approach is rapidly expanding to various areas of industrial and environmental chemistry. In most contemporary applications, *empirical* molecular descriptors that rely on some experimental data have been used in the

development of QSAR/QSPR equations. Such descriptors, starting from the original Hammett substituent σ constants to the most popular partition coefficients between water and octanol (log P) are, strictly speaking, restricted to the compounds for which the necessary experimental data are available. Another shortcoming of experimental descriptors evolves from the fact that many of them reflect a complicated combination of different physical interactions. Alternatively, the molecular descriptors can be derived using only the information encoded in the chemical structure of the compound. Importantly, such *theoretical* descriptors can be developed for the compounds that have never been synthesized or experimentally explored.

The motivation for this book is to provide a comprehensive overview of theoretical molecular descriptors used in chemistry and related sciences. During the last half of the twentieth century, much attention has been paid to the interrelation between the topology and chemical properties of molecules in QSAR/QSPR equations. The respective molecular descriptors have been particularly successfully used in prediction of the pharmaceutical activity of compounds. On the other hand, the semiempirical quantum chemical calculations for large molecules have become accessible even using small personal computers. Correspondingly, many descriptors of well-defined physical nature can be acquired from the results of such calculations. The simplest molecular descriptors can be defined just as the counts of different atoms and chemical bonds in a compound, whereas others proceed from the geometry of the molecule. In the present book, a systemic and critical overview of all main classes of theoretical molecular descriptors is given.

The critical analysis of descriptors is accompanied with a review of computational methods that can be employed in the development of QSAR/QSPR equations. The text is supplemented with software that allows to calculate many molecular descriptors for almost any chemical structure and includes basic methods for the development of QSAR/QSPR equations.

The present book is intended to be a helpful tool for both the academic and industrial chemists, from students to advanced researchers, for anyone who is interested in relating the properties of biological activity of compounds with their basic chemical structure.

Finally, I wish to thank my co-workers Dr. Uko Maran, Dr. Rein Hiob, Dr. Jaan Leis, Ms. Helle Kuura, and Mr. Tarmo Tamm for help in writing this book. In particular, I would like to acknowledge the efforts by Mr. Sulev Sild for help in creating the software.

Mati Karelson
Tartu, Estonia

MOLECULAR DESCRIPTORS IN QSAR/QSPR

1

INTRODUCTION

The development of quantum theory in the twentieth century has resulted in methods and techniques that, in principle, should enable us to predict the physical properties and chemical affinity of molecular matter with experimental precision. However, most *real-life* systems, including the processes in chemical reactors, drug–receptor interactions in biological systems, and the degradation of environmental pollutants, are characterized by such a complexity of intra- and intermolecular interactions that their theoretical description from first principles is impossible, even using the most powerful foreseeable computers. Different "model approximations" have therefore been used in ab initio quantum theory to make it applicable for larger chemical systems. In most cases, such simplifications are introduced on the basis of chemical intuition or proceed from various mathematical considerations.

Furthermore, because of a lack of an analytical solution to the so-called many-body problem, ab initio quantum theory of molecules is approximate by its very fundamental mathematical nature. The respective methods for molecular electronic structure calculations are thus often reduced to exercises in applied mathematics, with indirect connection to basic physical phenomena and introduction of various pseudo- or quasiphysical effects. Therefore, the applicability of approximate and semiempirical theories has to be verified for a given property and often even for a given restricted set or class of chemical compounds. In the case of loosely controlled theoretical approximations, the causal relationship between the molecular property and its structure can be obscured.

Over the past several decades, the quantitative structure–activity/property relationships (QSAR/QSPR) have become an alternative powerful theoretical tool for the description and prediction of properties of complex molecular systems in different environments. The QSAR/QSPR approach proceeds from the

assumption of the one-to-one correspondence between any physical property, chemical affinity, or biological activity of a chemical compound and its molecular structure. The latter can be represented by the chemical composition, connectivity of atoms, potential energy surface, and electronic wave function of a compound (cf. Fig. 1.1). Various physico-chemical molecular descriptors reflecting the structure can be determined empirically or by using theoretical and computational methods of different complexity. It must be underlined that a necessary requirement for the application of the QSAR/QSPR approach is the knowledge of the exact chemical constitution and/or the three-dimensional molecular structure of the chemical compounds studied.

The QSAR/QSPR relationships are usually derived using the (multiple) linear least-squares regression of the experimentally measured property values P against a preselected set of molecular descriptors (D_1, D_2, D_3, . . .):

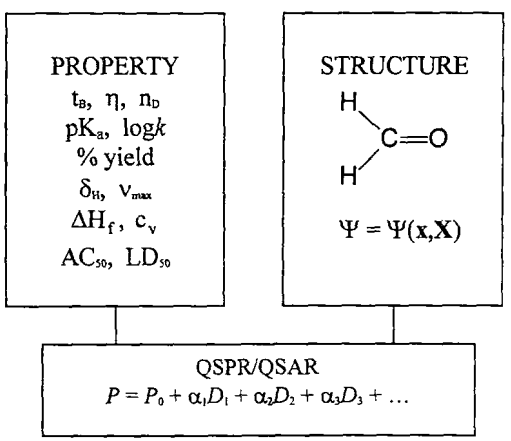


Figure 1.1 QSPR equations relate the property of a molecular system to its structure presented by the chemical constitution, atomic connectivity, and molecular electronic-nuclear wave function $\Psi(\mathbf{x}, \mathbf{X})$, where $\mathbf{x}$ and $\mathbf{X}$ denote the electronic and nuclear coordinates of the system, respectively. In many cases, the QSPR relationship is derived using the (multiple) linear least-squares fitting of the experimentally measured property values P with a selected set of molecular descriptors (D_1, D_2, D_3, . . .). The molecular property P can be a physical constant of the bulk of a compound (such as the normal boiling point t_B, viscosity η, or refractive index n_D at normal temperature), some measure of chemical reactivity (such as the acidic dissociation constant pK_a, rate constant k, or the yield of a chemical reaction), spectroscopic characteristic of a molecule [such as the proton chemical shift δ_H in the nuclear magnetic resonance (NMR) spectrum or the transition frequency ν_{max} in the electronic or vibrational spectrum of a compound], a thermodynamic function (such as the enthalpy of formation ΔH_f or heat capacity c_v), or the biological activity [such as the active concentration (AC_{50}) or the half-lethal dose (LD_{50})] of the compound.

$$P = P_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \cdots \tag{1.1}$$

In this expansion, the property P is a dependent variable whereas the descriptors D_i represent the independent variables. The coefficients α_i measure the weight of each descriptor involved in the relationship. Equation (1.1) is valid only if certain requirements are fulfilled for a given property and a given set of descriptors. First, the property value for a given compound has to be divisible into additive terms, each of which corresponds to a single molecular descriptor. Second, in the case of the variation of chemical structure, such terms should depend linearly on the respective descriptor values. In Figure 1.2, two possible situations are demonstrated. Apparently, the use of Eq. (1.1) is justified if the variation of the chemical structure corresponds to the interval of a descriptor $[D_0, D_1]$. Within the descriptor interval $[D_1, D_2]$, the linear QSAR/QSPR will not be valid.

In such case it is, of course, still possible to find a functional relationship between molecular property and descriptors. Instead of linear regression with molecular descriptors, some nonlinear form of the quantitative structure–property relationship has to be used. The simplest approach involves some nonlinear transformation of the descriptor, for example, the square, the square root, or the logarithm of the natural descriptor can be applied as an independent variable in Eq. (1.1). More systematically, the first few terms of the polynomial expansion of the appropriately rescaled molecular descriptor can be used as the independent variables in multiple linear regression. Various techniques are available for the direct nonlinear least-squares fitting of the property with the suitably chosen mathematical function on the descriptor. The application of artificial neural networks to develop dependence between the property and molecular descriptors of a chemical system also allows accounting for the intrinsically nonlinear relationship between them. Thus, in general, the absence of the (multiple) linear relationship between the observable property and molecular descriptor(s) of a system does not necessarily imply the absence of the causal or functional dependence between them.

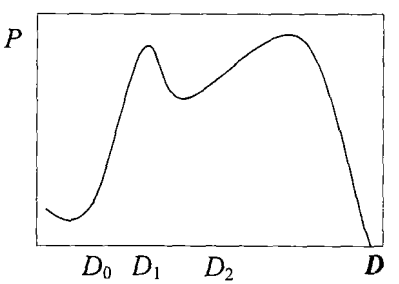


Figure 1.2 Dependence of property P of a molecular system on the molecular descriptor D, presented as a continuous function of the chemical structure.

The first requirement for the validity of Eq. (1.1) that demands the additivity of the different linear terms in the QSAR/QSPR equation may also be violated. In general, this additivity is strictly valid only when the intra- or intermolecular interactions, corresponding to different descriptors, are not affected by each other. In real-life situations, this may often not be the case. A well-known example of the interdependence of different intermolecular interactions is the steric restriction to the resonance effect in the conjugated π-electron molecular systems. One possible way to account for the nonadditive effects is the use of the cross terms between the respective molecular descriptors in the QSAR/QSPR regression:

$$P = P_0 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_{12} D_1 D_2 + \cdots \tag{1.2}$$

The addition of the cross terms decreases the number of statistical degrees of freedom for a given regression equation. However, it is acceptable when leading to substantial improvement of the respective QSAR/QSPR description.

The success of the QSAR/QSPR approach is critically dependent on the accurate definition and appropriate use of molecular descriptors. In this book, we attempt to provide a systematic and unified overview of the whole variety of *empirical* and *theoretical* molecular descriptors. This differentiation, while somewhat arbitrary, evolves from the possible limits of the applicability of QSAR/QSPR relationships using the descriptors from every individual group. The QSAR/QSPR equations obtained using solely the theoretically derived descriptors can be used, in principle, for the prediction of the respective properties of any molecular structure. Therefore, such equations can be expanded to the compounds for which the experimental information necessary for the definition of empirical descriptors is missing or which have not yet even been synthesized. The use of empirical descriptors restricts the predictive power of QSAR/QSPR equations to the compounds for which this experimental information is available.

The empirical descriptors can be divided into two general classes (Table 1.1). The first reflects the intramolecular electronic interactions (*structural descriptors*) whereas the second accounts for the intermolecular interactions in

TABLE 1.1 General Classification of Empirical Molecular Descriptors

Class	Subclass
Structural descriptors	Induction constants
	Resonance constants
	Steric constants
Solvational descriptors	Polarity scales
	Polarizability scales
	Acidity scales
	Basicity scales
	Mixed scales

condensed media such as liquids and solutions (*solvational descriptors*). The most widespread structural descriptors are defined to quantify the induction, the mesomeric or resonance, and the steric effects in chemical compounds. The solvational descriptors reflect the interactions of the solute with the bulk of the surrounding solvent (*macroscopic or nonspecific solvent effects*) and the specific bonding, mostly hydrogen bonding between the solute and individual solvent molecules (*microscopic or specific solvent effects*). The macroscopic solvent effects are quantified using various polarity and polarizability scales. The microscopic solvent effect descriptors include general acidity and general basicity scales. Some empirical solvent effect scales (*mixed scales*) may involve both the macroscopic and microscopic effects. A typical representative of such descriptors is the water–octanol partition coefficient, log P.

The theoretical molecular descriptors can be conventionally divided into a number of different classes, proceeding from either their complexity or the method of calculation. The simplest theoretical descriptors are the *constitutional* descriptors that can be constructed from the information about the chemical composition of the chemical compound. The absolute and relative counts of different types of atoms and chemical bonds, the molecular weight, and the number of different rings in the compound are some typical constitutional descriptors. The *topological* descriptors (also called topological indices) describe the atomic connectivity in the molecule. It has been debated that the topological indices may encode more subtle molecular interactions than just the branching of chemical bonds or specific mass distribution in the molecule. The *geometrical* descriptors are derived from the three-dimensional structure of molecules defined by the coordinates of atomic nuclei and the size of the molecule represented, for instance, by the atomic van der Waals radii. For most of the chemical compounds, the molecules possess certain conformational flexibility, and the respective molecular potential surfaces have multiple local minima. Depending on the structure of the molecule, the number of these minima can be very large and, therefore, it may be rather difficult to find the global energy minimum at given experimental conditions. Obviously, the geometrical descriptors may vary significantly depending on the conformation of the molecule. Therefore, it is important to verify the correctness of the conformation used in the calculation of these descriptors. To some extent, the *charge-distribution*-related theoretical descriptors may be also conformation dependent. These descriptors are based on the three-dimensional structure and the charge distribution in the molecule. The latter may be presented by atomic partial charges derived from some empirical schemes or from more sophisticated functions based on the quantum chemically calculated wave function of the molecule. A very interesting and rapidly growing direction in the area of charge-distribution-related descriptors is the use of molecular electrostatic fields. This direction, known primarily as the comparative molecular field analysis (CoMFA), has been successfully applied to the investigation of biological activity of compounds and to the computer-aided molecular design (CAMD) of new materials and pharmaceuticals.

A number of different *molecular-orbital*-based quantum chemical descriptors have been employed in the development of QSAR/QSPR equations. The most widely used are the frontier molecular orbital energies, that is, the calculated energy of the highest occupied molecular orbital (ε_{HOMO}), the energy of the lowest unoccupied molecular orbital (ε_{LUMO}), and the difference between these energies. Also, various reactivity indices derived from Fukui's theory of superdelocalizability or other theoretical constructions have gained popularity among researchers.

Modern quantum chemical program packages include the calculation of the statistical-physical partition function and its derivative thermodynamic functions of molecular systems. These molecular characteristics can be considered as molecular descriptors, particularly appropriate for systems at elevated temperatures where the thermal motion of molecules may substantially influence the process or property studied. It has to be stressed that the possible temperature dependence of molecular properties is not accounted for by traditional QSAR/QSPR descriptors. The respective effects may be, however, of great importance in studying many real-life systems, and thus the application of the theoretical *thermodynamic* or other *temperature-dependent* descriptors would be necessary to obtain a realistic picture of molecular systems. Various possibilities exist to include the temperature dependence of the properties and molecular descriptors in the QSAR/QSPR analysis.

During the last decade, great strides have been made in the development of the quantum theory of solvation. Numerous methods and algorithms introduced calculate the solvent effects on the molecular structure and properties. The theoretically computed individual contributions to the free energy of solvation and other characteristics are attractive as new theoretical *solvational* descriptors of compounds.

Not all theoretical descriptors can be strictly classified according to the above-given scheme (cf. Table 1.2). For example, the topographical indices are derived using the information about both the topology and the geometry of the molecules. The electrotopological indices are based on the topology and charge distribution whereas the charged partial surface area descriptors encode simultaneously the charge distribution and the geometry of compounds. Such descriptors can be classified as the *mixed* or *combined* molecular descriptors.

Molecular descriptors can be defined for the whole molecular system under study or for any part (fragment) of it. For instance, most empirical structural descriptors relate traditionally to molecular fragments called substituents. Accordingly, the molecules in a congeneric series of chemical compounds are formally divided into two or more fragments that correspond to a constant structural unit Y (e.g., the reaction center) and to the variable structural units X_i (substituents). The QSAR/QPSR relationships are thus presented as follows:

$$P = P_0^{(Y)} + \sum_i \sum_k \alpha_i^{(Y)} D_{ik}^{(X)} \qquad (1.3)$$

TABLE 1.2 General Classification of Theoretical Molecular Descriptors

Class	Subclass
Constitutional descriptors	Counts of atoms or bonds
	Atomic-weight-based descriptors
Topological descriptors	Topological (connectivity) indices
	Information-theoretical descriptors
	Topochemical descriptors
Geometrical descriptors	Distance-related descriptors
	Surface-area-related descriptors
	Volume-related descriptors
	Molecular steric field descriptors
Charge-distribution-related descriptors	Atomic partial charges
	Molecular electrical moments
	Molecular polarizabilities
	Molecular electrostatic field descriptors
Molecular-orbital-related descriptors	Frontier molecular orbital energies
	Bond orders
	Fukui's reactivity indices
Temperature-dependent descriptors	Thermodynamic functions
	Boltzmann factor-weighted descriptors
Solvational descriptors	Electrostatic energy of solvation
	Dispersion energy of solvation
	Free energy of cavity formation
	Hydrogen bonding descriptors
	Entropy of solvation
	Theoretical linear solvation energy descriptors
Mixed descriptors	Topographical descriptors
	Electrotopological descriptors
	Charged partial surface area descriptors

where $P_0^{(Y)}$ is the intercept corresponding to the constant molecular fragment Y, $D_{ik}^{(X)}$ are the molecular descriptors of type k for the variable fragments X_i, and $\alpha_{ik}^{(Y)}$ are the expansion coefficients characteristic for a given series of compounds X_iY. Many applications, especially the computer-aided molecular design of new compounds and materials, require the description and prediction of properties of compounds with large structural variability. The general QSAR/QSPR equations applicable for compounds involving different chemical functionalities will be useful in these cases. Instead of descriptors referring to structural fragments, the descriptors corresponding to the whole molecule are more appropriate to develop the necessary relationships. Notably, most of the theoretical descriptors listed in Table 1.2 can be calculated either for the whole molecule or for a predefined molecular fragment.

Many QSAR/QSPR applications involve complex multicomponent systems such as solutions and mixtures of chemical compounds, biological objects in vivo and in vitro, chemical reactors and aquifers, and others. Typically, in these cases only one molecular structure is singled out as presumably responsible for the variance of the given property of interest. The quantitative relationship is developed between the property of the whole system and the molecular descriptor(s) of this responsible component. However, the properties of a multicomponent system may depend from the individual contributions from each component and/or from the intermolecular interactions between the different components. In each case, this requires a special execution of the QSAR/QSPR treatment. In the case of negligible intercomponent interaction effects, the QSPR/QSPR expansion for a multicomponent system can be developed using the molar fraction weighted descriptors for individual components. A simple way to introduce the interaction terms between the different components is the use of the cross terms of descriptors belonging to the interacting components.

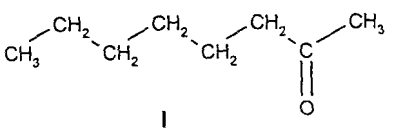
$$P = P_0 + \sum_i \sum_k \beta_{ik} D_i D'_k \qquad (1.4)$$

where D_i and D'_k are molecular descriptors for components i and k, respectively, and β_{ik} is the specific interaction constant. Depending on the nature of the intermolecular interactions between the different components of a multicomponent system, descriptors D_i and D'_k may be the same or different and reflect the physical mechanism of the interaction. Simple examples of such interactions are the electrostatic repulsion between the similarly charged molecular surfaces and the electrostatic attraction between the oppositely charged surfaces, respectively. In the first case, the same descriptor (fractional negative or positive surface area) is used as the interaction term [Eq. (1.4)]. In the second case, the product of two different descriptors (fractional negative surface area of one component multiplied by the fractional positive surface area of another component) has to be employed. The mathematical form of the dependence of some property of a multicomponent system on the molecular descriptors of different components may be more complex but known. In this case, a least-squares treatment of data according to the respective, possibly nonlinear, equation leads to the physically justified quantitative structure–property relationships.

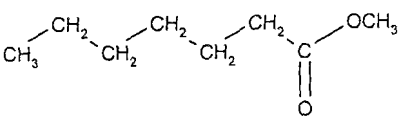
A large variety of statistical structure–property correlation techniques can be used for the analysis of experimental data in combination with the calculated molecular descriptors. First, the multiple linear regression methods can be applied in the scalar space of the original descriptors, in the principal-component orthogonalized space of the descriptors, or in the target-transformed descriptor space. Different stepwise and stagewise strategies are available for effective search of the best (most informative) multiparameter correlations in large spaces of the natural descriptors. The molecular properties, molecular descriptors, or their combinations can be analyzed using factor-analysis-based pattern recognition methods, including principal component analysis (PCA), partial

least squares (PLS), or nonlinear partial least squares (NIPALS). In the case of intrinsically nonlinear dependence between the experimental property of compounds and molecular descriptors, nonlinear regression methods can be applied for the development of QSAR/QSPR equations. The intrinsic nonlinear dependence may also be encoded in the respective artificial neural networks. Notably, the choice of method for the data treatment is largely independent of the descriptors applied. In other words, most molecular descriptors can be applied universally in different QSAR/QSPR treatments. This flexibility makes the molecular descriptors, especially the theoretically derived descriptors, attractive and efficient for the construction of working models to predict the physical, chemical, and biomedical properties of compounds.

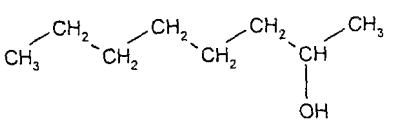
Frequently, the concept of molecular similarity is used for the prediction of properties for new and previously unknown molecular systems. It is important to emphasize, however, that there is no absolute similarity measure between any two molecular structures. Let us consider a set of the following six compounds: 2-octanone (**I**), methyl heptanoate (**II**), 2-octanol (**III**), acetone (**IV**), methyl acetate (**V**), and isopropanol (**VI**).



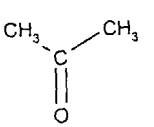
I



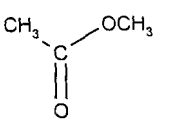
II



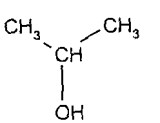
III



IV



V



VI

By looking at the simple constitutional descriptors, such as the counts of carbon or hydrogen atoms in the molecule, compounds **I** to **III** belong to one class of structurally similar compounds whereas compounds **IV** to **VI** relate to distinctly different class of compounds (cf. Table 1.3). The same classification is applicable if we compare the molecular volumes or the surface areas of molecules. The theoretically calculated dipole moments of compounds do not show, however, any distinct grouping of molecules. The experimentally determined properties of these compounds exhibit similar diversity. For instance, the boiling point of compounds **I** to **III** is very similar. According to this property, acetone (**IV**) and methyl acetate (**V**) constitute another group of very similar compounds whereas isopropanol (**VI**) stays separately. By comparison of the wavelength of the spectral maxima corresponding to the lowest energy transition in the ultraviolet spectra of compounds, another classification of the six compounds could be obtained. Compounds **I**, **II**, **IV**, and **V** have a carbonyl group in their structure, and the spectral maxima of the accompanying $n - \pi^*$ transitions are located substantially above 200 nm. The maxima for compounds **III** and **VI** are shifted to substantially shorter wavelengths (<200 nm). Thus, according to this property, they form another class of compounds. Finally, the comparison of the chemical reactivity with water results in yet another classification by similarity of the six compounds. Ketones **I** and **IV** undergo the enolization in aqueous solutions whereas esters **II** and **V** are hydrolyzed by water. Alcohols **III** and **VI** do not react with water.

Consequently, the similarity of compounds is related to the specific property investigated. On the other hand, different molecular descriptors also discriminate between the molecular structures differently. Therefore, the multitude of

TABLE 1.3 Numerical Values of Some Molecular Descriptors and Experimental Properties for a Set of Six Compounds: 2-octanone (I), methyl heptanoate (II), 2-octanol (III), acetone (IV), methyl acetate (V), and isopropanol (IV)

Property/Compound	I	II	III	IV	V	VI
Count of C atoms	8	8	8	3	3	3
Count of H atoms	16	16	18	6	6	8
Molecular volume	148.60	157.20	154.71	63.84	72.84	70.34
Molecular surface area	178.15	187.07	179.67	81.72	91.68	85.64
Dipole moment (AM1)	2.742	1.663	1.692	2.922	1.775	1.615
Boiling point (°C)	173	173.5	174	56	57.5	82
λ_{max}(ultraviolet), nm	>200	>200	<200	>200	>200	<200
Reaction with water	E[a]	H[b]	I[c]	E[a]	H[b]	I[c]

[a]Enolization.
[b]Hydrolysis.
[c]Inert.

molecular descriptors available for the development of QSAR/QSPR equations is not really a result of the fantasy of researchers but based on the specific relationships between the properties and molecular characteristics of compounds. It is thus very important to make a correct choice of descriptors for the treatment of a given property. In the following, we proceed with a systematic analysis of the genesis and applicability of the multitude of empirical and theoretical descriptors.

2

EMPIRICAL MOLECULAR DESCRIPTORS

2.1 HISTORICAL BACKGROUND

Investigation of the dependence between the molecular structure and properties of chemical compounds has been an important subject of research throughout modern chemistry. Some authors have noted that the first attempts to relate biological activity with the chemical properties of compounds date back to the middle of nineteenth century. Already in 1863, Cros observed that toxicity of alcohols to mammals increased as the water solubility of the alcohols decreased [1]. At the turn of the last century, Meyer and Overton related the toxicity of organic compounds to their lipophilicity [2–4].

An important step toward the development of quantitative structure–property relationships was made by Brönsted and Pedersen who introduced the equations to correlate the rate constants of different general acid- or base-catalyzed reactions with the respective acidic or basic dissociation constants of the catalytic acids or bases:

$$k_{HA} = g_A(K_{HA})^{\alpha} \tag{2.1}$$

for acids, and

$$k_B = g_B(K_B)^{\beta} \tag{2.2}$$

for bases. The coefficients g_A and g_B and exponents α and β are characteristic for a given catalytic process. In principle, these equations define a linear relationship between the free energy of activation of one process (catalytic re-

action) and the free energy of another chemical reaction (acid–base equilibrium) as:

$$\Delta G^{\neq}_{\text{cat}(HA)} = -RT \ln g_{HA} + \alpha \Delta G_{HA} \tag{2.3}$$

and

$$\Delta G^{\neq}_{\text{cat}(B)} = -RT \ln g_{B} + \beta \Delta G_{B} \tag{2.4}$$

where $\Delta G^{\neq}_{\text{cat}(HA)}$ and $\Delta G^{\neq}_{\text{cat}(B)}$ denote the free energies of activation of the acid- and base-catalyzed reactions, respectively, and ΔG_{HA} and ΔG_B are the free energies of the acid–base equilibria involving the corresponding catalytic acids or bases. The approach that relates the free energies of two processes involving the same molecular structure by equations similar to the Brönsted equations [(2.3) and (2.4)] is known in physical organic chemistry as the method of *linear free energy relationships* (LFER). It has been suggested that this name be replaced by linear Gibbs energy relations (LGER) [5], but the latter is still rarely used.

An important landmark in the development of LFERs was set by Hammett whose classical work (1935) associated the chemical reactivity of meta- and para-substituted benzenes with the acidic dissociation constants of the similarly substituted benzoic acids [6,7]. The well-known Hammett equation has the following form:

$$\log \left(\frac{k_X}{k_H}\right) = \rho\sigma_X \tag{2.5}$$

where k_H and k_X are the rate constants for some chemical reaction involving the unsubstituted benzene and the benzene derivative with the substituent X, respectively. The so-called σ constants

$$\sigma_X = \log \left(\frac{K_X}{K_H}\right) \tag{2.6}$$

were defined as the logarithmic ratio of the acidic dissociation constants of the substituted (by substituent X) and the unsubstituted benzoic acids. The constant ρ is characteristic for a given reaction. The Hammett equation has been extended to other aromatic systems. In addition, σ constants have been developed for the ortho-substituted compounds, but the situation is more complicated in this case because of the spatial closeness of the reaction center and the substituent.

After the first success with the Hammett equation, it was understood that at least two different intermolecular interaction mechanisms are reflected by the σ constants. First, the *induction* effect is originated from the polarization of the chemical bonds by electronegative atoms or atomic groups. Second, depending on the chemical nature and mechanism of the process, direct polar conjugation between the substituent and the reaction center may affect the reactivity in substituted aromatic systems (*mesomeric* or *resonance* effect). A multiparameter extension of the Hammett equation was thus developed to quantify the role of enhanced resonance effects on the reactivity of meta- and para-substituted benzene derivatives as [8–10]:

$$\log \left(\frac{k_X}{k_H}\right) = \rho\sigma + r(\sigma^- - \sigma) \tag{2.7}$$

or

$$\log \left(\frac{k_X}{k_H}\right) = \rho\sigma + s(\sigma^- - \sigma) \tag{2.8}$$

The second term in Eq. (2.7) reflects the enhanced resonance effect by the electron-donating substituents, quantified by the substituent constant $(\sigma^+ - \sigma)$. The second term in Eq. (2.8) describes the increase of the direct resonance between the reaction center and the electron-accepting substituent in the conjugated para-position using the scale $(\sigma^- - \sigma)$. A more recent generalization of the Hammett equation belongs to Taft and co-workers who introduced the dual substituent parameter (DSP) equation [11,12]:

$$P = P_0 + \rho_I\sigma_I + \rho_R\sigma_R \tag{2.9}$$

which has been applied to the description of chemical, physical, and spectroscopic properties of substituted benzenes. In Eq. (2.9), P denotes the magnitude of a given property for the compound with substituent X, and P_0 is to the property value for the unsubstituted benzene. The terms σ_I and σ_R denote the inductive, or polar, and the resonance substituent constants, respectively. The coefficients ρ_I and ρ_R characterize the relative susceptibility of the property studied to the induction and resonance effect, respectively.

Already in 1953, Taft had extended the LFER approach to the aliphatic organic compounds [13–16]. He noticed that the transition states (TS) of the acid- and base-catalyzed reactions of the ester hydrolysis are geometrically similar, the difference being only in the charge and two additional protons in the TS of the acid-catalyzed process. Therefore, it was plausible to assume that the possible steric effects would cancel if the free energies of activation of these two reactions for the same ester were compared. At the same time, the influence of the possible induction effect by the substituents in the substrate ester molecules has to be significantly different in the case of the acid- and base-catalyzed reaction, respectively. This observation allowed Taft to define a quantitative scale of the induction effect by the substituents in the aliphatic series of compounds as:

$$\sigma^* = \frac{1}{2.48}\left[\log\left(\frac{k_X}{k_0}\right)_B - \log\left(\frac{k_X}{k_0}\right)_A\right] \tag{2.10}$$

where k_X denotes the rate constant for the compound with the substituent X and k_0 is the rate constant for the standard compound, respectively. Subscripts B and A correspond to the base-catalyzed and to the acid-catalyzed reaction, respectively.

The quantitative LFER treatment of steric effects was also pioneered by Taft [14,17]. It was observed that the Hammett ρ constant for the acid hydrolysis of esters was close to zero. Consequently, the induction effect by substituents on this reaction is practically negligible. Thus, the original E_S substituent constants were defined as the logarithmic ratio of the rate constants for the acid hydrolysis of the ester with the substituent X and of the standard compound:

$$E_S = \log\left(\frac{k_X}{k_0}\right)_A \tag{2.11}$$

and employed in the respective general Taft equation as:

$$\log\left(\frac{k_X}{k_H}\right) = \rho^*\sigma^* + \delta E_S \tag{2.12}$$

where δ scales the susceptibility of the given process to the steric effect.

It has to be emphasized that all of these substituent constants were defined using some experimental data; thus, in principle, they are empirical structural molecular descriptors.

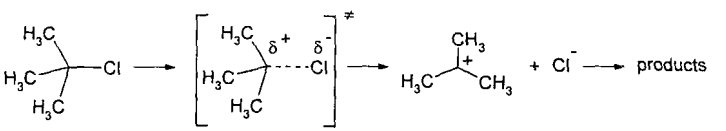
In addition to the structural effects on the chemical reactivity and properties of compounds, the solvent effects reflecting the influence of the surrounding environment on the properties of molecules in condensed media have been quantified using the empirical LFER approach. As in the case of LFER constants for structural effects, solvent effect parameters have been introduced proceeding from different physical models of solvation and intermolecular interactions in condensed media. These parameters can be divided into two general groups. The first kind of parameter has been developed using the concept of electrostatic interactions between the solute and the bulk of the solvent. Within this approach, the surrounding medium (i.e., the solvent) is presented as a homogeneous dielectric continuum, polarizable by the electrostatic field created by a discrete solute molecule. Various theoretical constructions predict a linear relationship between the free energy of the observed process and some function of the dielectric constant of the medium [18]. Because of the different characteristic relaxation times of different dielectric relaxation processes in solution (shift of the electron distribution, conformational changes of the molecule), it has been proposed to distinguish between *polarity* and *polarizability* effects [19]. The first refers to the full dielectric polarization of the solvent in

the field of the solute molecule. In the respective solvent LFER scales, the macroscopic dielectric permittivity at zero frequency has been employed. The polarizability effects arise from only the short-time electron–nuclear polarization of the solvent, and thus they relate to the dielectric permittivity of solvent at infinite frequency of the external electric field.

Another group of LFER solvent effect scales has been developed to describe the short-range specific solute–solvent interactions. These involve hydrogen bonding and other semichemical interactions between the discrete solute and solvent molecules. The solute–solvent interactions, where the solvent is acting as a hydrogen bond donor and the solute molecule as an acceptor, are referred to as the solvent *acidity*, or *electrophilicity*, effects. The mirror interactions involving the solvent molecules as hydrogen bond acceptors and the solute molecule as the donor of the hydrogen bond are related to the solvent *basicity*, or *nucleophilicity*. It has been recommended to use the first terms in these pairs (acidity and basicity) in the case of equilibrium processes whereas the second terms (electrophilicity and nucleophilicity) are preferable in the case of kinetic or dynamic processes.

The influence of the solvent on the rates and equilibria of chemical reactions was recognized already in the nineteenth century. The solvent effects on the rate of chemical reactions were first reported by Berthelot and Péan de Saint-Gilles [20,21] who noticed that the addition of some solvents to the reaction medium decelerate the esterification of acetic acid by ethanol. The influence of the solvent on the chemical equilibria was discovered in studies of the keto-enol tautomerism of 1,3-dicarbonyl compounds [22–24]. In his classical work on the reaction named after him, Menshutkin related the solvent effects on the rate of reaction to the chemical nature of the solvent [25,26].

One of the first LFER-type solvent effect treatments belongs to Grünwald and Winstein who proceeded from the observation that the S_N1 solvolysis of *tert*-butyl chloride is substantially accelerated by the polar and protic solvents [27].



Accordingly, they defined a solvent ionizing power constant, Y, as:

$$Y = \log k_S - \log k_0 \tag{2.13}$$

where k_S is the rate constant of the solvolysis of *tert*-butyl chloride in a given solvent S and k_0 denotes the rate constant of this reaction in the standard solvent (80% aqueous ethanol) at 25°C. The LFER describing the solvent effect on some other reaction (A) can then be presented as:

$$\log k_S^A = \log k_0^A + mY \tag{2.14}$$

where k_S^A and k_0^A are the rate constants of this reaction in a given solvent S and in the standard solvent, and m denotes the sensitivity of the reaction to the solvent effect. It was shown later that the Y parameter involves, apart of the solvent polarity effects, the solvent polarizability and the electrophilicity terms, the latter being related to the electrophilic solvent assistance of the chloride ion elimination during the reaction [19]. Winstein himself extended Eq. (2.14) to account for the solvent nucleophilic assistance effects on the chemical reactions by introducing an additional term $\ell \cdot N$, where N is the nucleophilicity of the solvent and ℓ is the respective sensitivity coefficient. The nucleophilic (basic) solvation effects were also considered within the LFER formalism by Gutmann [28,29]. He provided an empirical scale of the Lewis basicity by defining a donor number (DN) of the electron pair donating solvents as the negative value of the molar enthalpy for adduct formation between antimony pentachloride and a given solvent. Another scale of solvent basicity was defined by Koppel and Paju [30] using the band shifts of the O—H stretching vibration of the phenol $\Delta\tilde{\nu}$ in tetrachloromethane, induced by the hydrogen bond formation with the added hydrogen-bonding accepting solvent S:

$$B \equiv \Delta\tilde{\nu} = \tilde{\nu}_{\mathrm{PhOH}}^{\mathrm{CCl_4}} - \tilde{\nu}_{\mathrm{PhOH\cdots S}}^{\mathrm{CCl_4}} \tag{2.15}$$

In general, the LFER equations of solvent effects should include terms accounting both for the nonspecific and specific solute–solvent interactions in solutions. A variety of multiparameter equations has been proposed including different solvent effect scales. One of the first such equations was suggested by Koppel and Palm [31,32] as:

$$P = P_0 + y \cdot Y + p \cdot P + e \cdot E + b \cdot B \tag{2.16}$$

where P and P_0 are the measured property values in a given solvent and in the standard solvent, respectively; $Y = (\varepsilon - 1)/(2\varepsilon + 1)$ is the solvent polarity parameter, $P = (n_D^2 - 1)/(2n_D^2 + 1)$ reflects the solvent polarizability, E is the solvent electrophilicity, and B is the basicity of the solvent. The coefficients y, p, e, and b are determined by the least-squares regression of the experimental property values P against the respective four solvent scales.

An alternative multiparameter approach to the solvent effects was proposed by Taft and Kamlet [33–37] who developed the so-called linear solvation energy relationship (LSER) as:

$$P = P_0 + s(\pi^* + d\delta) + a\alpha + b\beta + h\delta_H^2 + e\xi \tag{2.17}$$

In this equation, P and P_0 express again the measured property values in a given solvent and in the standard solvent, respectively, and π^* is an index of the solvent dipolarity/polarizability. The latter characterizes the solvent ability

to stabilize the dipole or the charge of the solute by the dielectric reaction field of the solvent. It was shown that for a series of solvents with a dominant dipolar group, the last parameter is proportional to the total dipole moment of molecules. In Eq. (2.17), δ is a discontinuous polarizability correction term with the value $\delta = 0$ for the nonchlorine substituted aliphatic solvents, $\delta = 0.5$ for polychlorinated aliphatics, and $\delta = 1$ for aromatic solvents. In the same equation, α denotes the measure of the hydrogen bond donating ability of the solvent and corresponds to parameter E in Eq. (2.16). Descriptor β is the measure of the hydrogen bond accepting ability of the solvent similar to parameter B in Eq. (2.16), and δ_H^2 represents the square of the Hildebrandt solubility parameter. Parameter ξ has been introduced to correlate certain types of so-called family-dependent solute basicity properties.

It is, however, doubtful that any of the empirical solvent effect scales would correspond to a single mechanism of physical interactions between the solute and solvent molecules. This has caused an unnecessary confusion and misunderstanding between researchers as the same terms have been used for different experimentally derived solvent effect scales that involve the fundamental physical intermolecular interactions with different weights.

In the following, we proceed with the systematic presentation of different empirical molecular descriptors, together with the analysis of the interrelation between them and the limits of applicability.

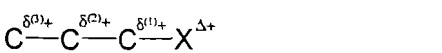
2.2 STRUCTURAL DESCRIPTORS

2.2.1 Induction Constants

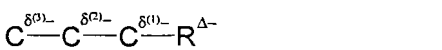
The induction or inductive effect is one of the fundamental terms in physical organic chemistry that describes the intramolecular interaction between the functional groups or molecular fragments in organic molecules [38]. Two different physical mechanisms have been considered as a cause for this effect [39–44]. No conclusive discrimination between them has been achieved, and thus we briefly describe both. The first mechanism of induction effect was suggested by Kirkwood and Westheimer [45–48], who formulated this effect as resulting from the electrostatic interaction between the charges and/or dipoles in the molecule. The polarization of a chemical bond between the atoms of different electronegativity was conceived as the cause for the formation of partial charges of opposite sign on these atoms and a respective local dipole moment of this bond. Thus, in the model of Kirkwood and Westheimer, the electrostatic interaction between the bond dipoles or the partial charges on atoms is considered the origin of the induction effect.

According to the second physical model, the induction effect is caused by the consecutive polarization of the bonds along the chain of chemically bonded atoms involving at least one bond between the atoms of different electronegativity [49]. The magnitude of this polarization depends on the distance

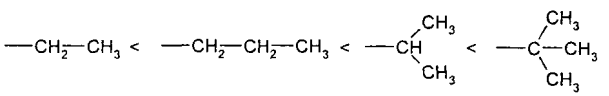
from such bonds and the difference in the electronegativity of the respective atoms. Two groups of substituents can be distinguished according to the direction of the polarization along the chain of carbon atoms. First, the more electronegative substituents X as compared to the carbon atom induce the positive atomic partial charges on the carbon atoms.



By convention, these substituents are denoted as the −I groups. The inductive effect fades with distance from the electronegative group and, consequently, $\delta^{(1)+} > \delta^{(2)+} > \delta^{(3)+}$. The majority of functional groups in organic compounds (halogens, carbonyl-, hydroxyl-, amino-, amide-, nitro-, cyano-groups, etc.) possess the −I inductive effect. The alkyl groups, R, are assumed to have an opposite effect by creation of negative atomic partial charges along the carbon chain.



The methyl group has been often taken as a standard for which the bond polarization along the chain and the respective inductive effect is postulated to be zero. According to this definition, the longer and more branched hydrocarbon radicals possess the increasing +I effect in the following order:



This definition of the induction effect for alkyl substituents has been disputed by Charton who ascribed the respective change in the free energy of a standard process to the difference in steric repulsion by such substituents [50,51]. Nevertheless, the induction constants for alkyl groups are still used together with the induction constants for electronegative groups in the same LFER correlations.

The unique standardization of the quantitative scale for the induction effect has been also problematic. First, it has to be mentioned that neither the electrostatic nor the bond polarization model of the induction effect can be reduced to fundamental physical interactions. The total energy, E, of a molecule consisting of M nuclei and n electrons is quantum mechanically determined as the solution of the respective Schrödinger equation:

$$\hat{H}\Psi = \left[-\sum_{a=1}^{M} \frac{\nabla_a^2}{2M_a} - \sum_{i=1}^{n} \frac{\nabla_i^2}{2} - \sum_{a=1}^{M} \sum_{i=1}^{n} \frac{Z_a}{r_{ia}} + \frac{1}{2} \sum_{a=1}^{M} \sum_{b=1}^{M} \frac{Z_a Z_b}{r_{ab}} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{r_{ij}} \right] \Psi = E\Psi \qquad (2.18)$$

This energy can be partitioned into separate contributions, corresponding to individual terms of Hamiltonian $\hat{H}$ in Eq. (2.18). Those include, in the following order, the kinetic energy of nuclei, the kinetic energy of electrons, the electron–nuclear electrostatic attraction, the nuclear–nuclear electrostatic repulsion, and the electron–electron electrostatic repulsion energy, respectively. In Eq. (2.18) the r_{ia}, r_{ab}, and r_{ij} correspond to the distances between the electrons and nuclei, to the internuclear distances and to the interelectronic distances, respectively. The term Ψ denotes the total wave function of the molecule. The induction effect is expected to influence each of the energy terms given in Eq. (2.18), by the perturbation caused with variable substituents in the molecule. The effect on each individual term, however, cannot be scaled out. Therefore, the quantitative measure of the induction effect cannot be determined theoretically, by inspection of the respective terms in the mathematical expression of the quantum mechanical total energy of the molecule. Accordingly, the absolute magnitude of the induction effect is unknown, and thus it has been conventionally defined with respect to some standard level of energy.

The use of different standards for the quantitative definition of induction effect has resulted in numerous empirical scales for this effect. It is evident that the response of different chemical or physical characteristics of a molecule (e.g., the energies of spectral transitions, the rate of chemical reactions, or the position of chemical equilibria) to the inductive effect would be different. In most cases, it is also difficult to prove that other intermolecular (mesomeric, steric) effects do not influence the standard process. Therefore, one reason for the differences in different induction effect scales is the systematic error due to the presence of other effects of unknown extent. In addition, different standard processes take place in different environment, e.g., in different media or at different temperatures. These external factors can also substantially influence the intermolecular induction effects. This influence may be specific for each molecular structure, and thus the substituent effects may be affected differently. An additional systematic deviation of one induction effect scale from another may arise from that reason.

In conclusion, obviously no general empirical induction effect scale is available for every process. Two remarks are essential on this point. First, it is advisable to use the scale of induction effect that is more closely related to the phenomenon studied by the LFER analysis. Second, "bare" induction effect that is not disturbed by surrounding medium should be observed in the case of standard processes occurring in the gas phase at low pressure (e.g., the gas-phase proton affinities).

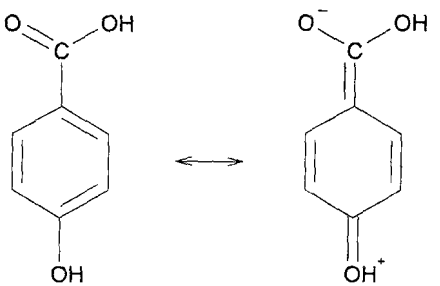
The original induction effect scales, Hammett's σ and Taft's σ^*, were defined on the basis of standard processes in aqueous solutions [cf. Eqs. (2.6) and (2.10)]. Because of the difficulties with the solubility of many substituted benzoic acids in water, the original σ scale was soon extended using the pK_a of benzoic acids in mixed solvents. In addition, other processes for which the log k or pK is linearly related to the pK_a of benzoic acids in aqueous solutions had been used for the definition of induction constants. For example, it was

noticed that the pK_a of acids XCH_2COOH are linearly correlated with the original σ^* constants and thus applicable for the further extension of the induction effect scale in aliphatic and alicyclic systems. In fact, a new scale

$$\sigma_I^{(X)} = pK_a^{(XCH_2COOH)} - pK_a^{(CH_3COOH)} \tag{2.19}$$

was defined on the basis of the acidic dissociation constants of substituted acetic acids [52]. The similar σ_I constants have been developed also for heterocyclic systems [53].

As discussed, the original Hammett σ constants determined from the dissociation constants of substituted benzoic acids included, at least for the electron-donating mesomeric groups in para-position, a contribution related to the direct mesomeric interaction through the aromatic cycle; for example:

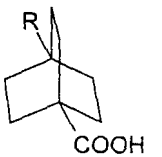


The comparison of the σ and σ_I constants revealed that even in the case the meta-substituents the two scales are poorly correlated. One possible reason for this discrepancy is the possible mesomeric conjugation between the electronegative substituents and the aromatic phenyl ring. Thus, a new scale for the induction effect in aromatic systems (σ^0) was defined on the basis of the following relation:

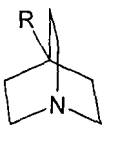
$$\sigma^0_{(XC_6H_4)} = \sigma^*_{(XC_6H_4)} - 0.600 \tag{2.20}$$

where $\sigma^*_{(XC_6H_4)}$ is determined from the pK_a of phenyl-substituted acetic acids [54,55]. In result, these constants have the same scaling as the σ^* constants for the aliphatic substituents. The scale of σ^0 constants has been extended using the rate constants of the hydrolysis of substituted phenylacetic acids [56] and by applying a general statistical treatment of similar data on the reactivity of substituted phenyl systems [57]. In each of these systems, the reaction center was separated from the aromatic ring by the methylene group, which disabled the possible mesomeric conjugation with the substituent in the ring.

To eliminate the possible mesomeric effects, several other standard reactions have been employed for the definition of the induction constants in cyclic systems. Those include the pK_a of dissociation 4-substituted bicyclo[2,2,2]octane carboxylic acids (**I**) [58] and the pK_a of protonation of 4-substituted quinuclidines (**II**) [59–61].



(I)



(II)

In these systems, the substituent R is geometrically rigidly fixed with respect to the reaction center and the possible mesomeric interaction between these two groups has been disabled because of aliphatic $—CH_2—CH_2—$ bridges between them. Table 2.1 lists various substituent induction constants together with reference to the standardization process used. Table 2.2 presents the numerical values of some common induction constants for a selection of substituents [62,63,63a].

Two principles have been employed in the calculation of the induction constants of complex substituents involving several electronegative groups, aliphatic chains, and multiple substitutions in aromatic rings. First, the induction effect is assumed to be additive, that is, the induction constant σ_I for a multiply substituted radical can be added to the σ_I constants of the respective substituent groups [64–66]:

$$\sigma_I^{(—CX_1X_2X_3)} = \sum_{i=1}^{3} \sigma_I^{(—CH_2X_i)} \tag{2.21}$$

It has been argued, however, that in the case of several strong electron-withdrawing groups (e.g., Cl, F) at the same carbon atom, the additivity of the substituent constants can break down, leading to a "saturation" of the induction effect. Originally, the induction effect by several substituents in an aromatic system (e.g., in the phenyl ring) has been also assumed to be additive [67]. However, it was later shown that this additivity does not hold strictly [68,69]. An alternative "experimental" scale of Hammett σ constants for multiply substituted phenyl groups has been suggested by Hansch and others [70]. The Hammett σ constants are also expected to be position dependent. Nevertheless, a strong collinearity has been reported between the σ constants for the same substituent at the para- and the meta-position of phenyl ring, σ_m and σ_p, respectively [62], expressed by the following equation:

$$\sigma_p = (0.08 \pm 0.02) - (1.19 \pm 0.04)\sigma_m$$
$$R^2 = 0.885 \qquad s = 0.137 \qquad n = 530$$

The second principle applied to the induction interaction asserts that the effect