CHAPTER 2

Chemical Information and Descriptors

Contents

2.1 Introduction 48

2.2 Concept of Descriptors 48

2.3 Type of Descriptors 53

2.3.1 Substituent constants 53

2.3.2 Whole molecular descriptors 54

2.4 Descriptors Commonly Used in QSAR Studies 54

2.4.1 Physicochemical descriptors 54

2.4.1.1 Hydrophobic parameters 54

2.4.1.2 Electronic parameters 56

2.4.1.3 Steric parameters 58

2.4.2 Topological descriptors 61

2.4.3 Structural descriptors 61

2.4.4 Indicator variables 61

2.4.5 Thermodynamic descriptors 67

2.4.6 Electronic parameters 68

2.4.7 Quantum chemical descriptors 69

2.4.7.1 Mulliken atomic charges 69

2.4.7.2 Quantum topological molecular similarity indices 69

2.4.8 Spatial parameters 69

2.4.8.1 RadofGyration 70

2.4.8.2 Jurs descriptors 70

2.4.8.3 Shadow indices 71

2.4.8.4 Molecular surface area 71

2.4.8.5 Density 72

2.4.8.6 Principal moment of inertia 73

2.4.8.7 Molecular volume 73

2.4.9 Information indices 73

2.4.9.1 Information of atomic composition index 74

2.4.9.2 Information indices based on the A-matrix 74

2.4.9.3 Information indices based on the D-matrix 74

2.4.9.4 Information indices based on the E-matrix and the ED-matrix 75

2.4.9.5 Multigraph information content indices (IC, BIC, CIC, SIC) 75

2.4.10 Molecular shape analysis descriptors 75

2.4.11 Molecular field analysis parameters 76

2.4.12 Receptor surface analysis parameters 76

2.5 Overview and Conclusion 78

References 78

Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment.
ISBN: 978-0-12-801505-6, DOI: http://dx.doi.org/10.1016/B978-0-12-801505-6.00002-8

2.1 INTRODUCTION

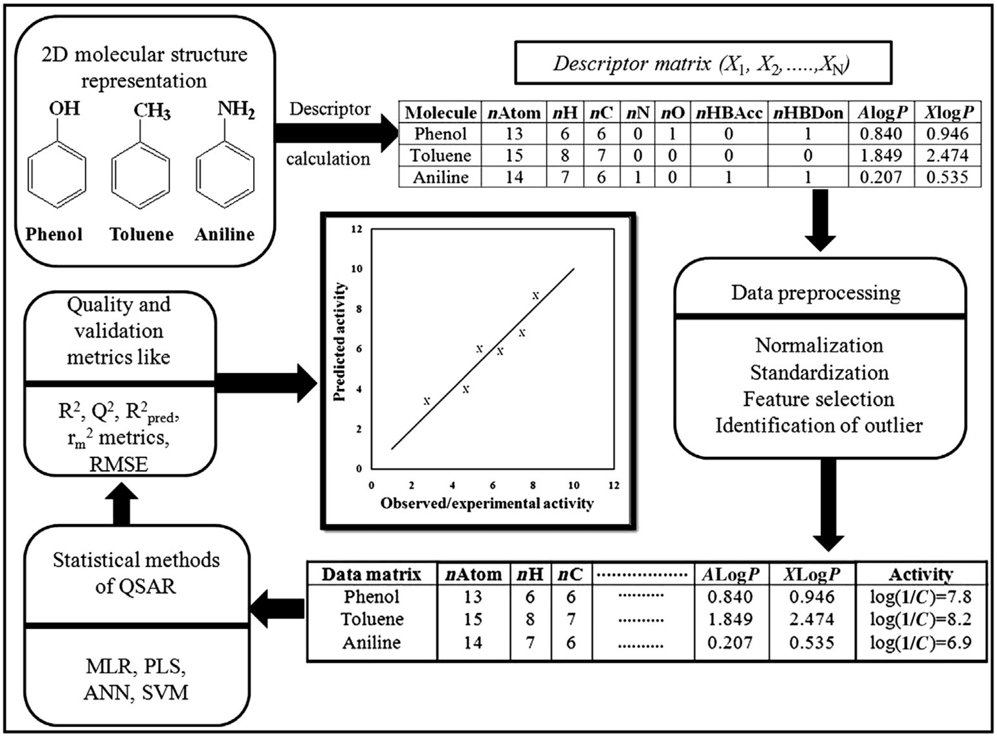
The quantitative structure−activity relationship (QSAR) technique, being directly related to the molecular structures of chemicals, can explain the effects exerted by the chemicals in relation to their structures and properties. Any significant search for the required chemical information of molecules for a particular end point can provide a strong tool for the predictive assessment of the response of existing untested as well as new chemicals [1]. QSAR is a simple mathematical model that can correlate chemistry with the properties (physicochemical/biological/toxicological) of molecules using various computationally or experimentally derived quantitative parameters known as *descriptors*. These descriptors are correlated with the response variable using a variety of chemometric tools in order to obtain a meaningful QSAR model. The developed models provide a significant insight regarding the essential structural requisites of the molecules, thus enabling us to identify the features contributing to the biological activity/property/toxicity of the studied molecules [2].

2.2 CONCEPT OF DESCRIPTORS

Molecular descriptors are terms that characterize specific information about a studied molecule. They are the "numerical values associated with the chemical constitution for correlation of chemical structure with various physical properties, chemical reactivity, or biological activity" [3,4]. In other words, the modeled response (activity/property/toxicity of query molecules) is represented as a function of quantitative values of structural features or properties that are termed as descriptors for a QSAR model. Cheminformatics methods depend on the generation of chemical reference spaces into which new chemical entities are predictable by the developed QSAR model. The definition of chemical spaces significantly depends on the use of computational descriptors of studied molecular structure, physical or chemical properties, or specific features.

$$\text{Response(activity/property/toxicity)} = f(\text{Information in form of chemical structure or property}) = f(\text{Descriptors})$$

The type of descriptors used and the extent to which they can encode the structural features of the molecules that are correlated to the response are critical determinants of the quality of any QSAR model. The descriptors may be physicochemical (hydrophobic, steric, or electronic), structural (based on frequency of occurrence of


2D molecular structure representation
OH
CH$_3$
NH$_2$
Phenol
Toluene
Aniline
Descriptor
calculation
Descriptor matrix (X_1, X_2,.....,X_N)
Data preprocessing
Normalization
Standardization
Feature selection
Identification of outlier
Quality and validation metrics like
R^2, Q^2, R$^2_{pred}$, r$_m^2$ metrics, RMSE
Predicted activity
Observed/experimental activity
0
2
4
6
8
10
12
Statistical methods of QSAR
MLR, PLS, ANN, SVM

Molecule	*n*Atom	*n*H	*n*C	*n*N	*n*O	*n*HBAcc	*n*HBDon	Alog*P*	Xlog*P*
Phenol	13	6	6	0	1	0	1	0.840	0.946
Toluene	15	8	7	0	0	0	0	1.849	2.474
Aniline	14	7	6	1	0	1	1	0.207	0.535

Data matrix	*n*Atom	*n*H	*n*C	....	ALog*P*	XLog*P*	Activity
Phenol	13	6	6	....	0.840	0.946	log(1/*C*)=7.8
Toluene	15	8	7	....	1.849	2.474	log(1/*C*)=8.2
Aniline	14	7	6	....	0.207	0.535	log(1/*C*)=6.9

Figure 2.1 How chemical structure is used to calculate descriptors and QSAR model development.

a substructure), topological, electronic (based on molecular orbital calculations), geometric (based on a molecular surface area calculation), or simple indicator parameters (dummy variables). A schematic overview is presented in Figure 2.1 in order to show the steps how a chemical structure is used to calculate descriptors and used in QSAR model development.

A dimension in the QSAR analysis acts as the constraint that controls the nature of the analysis. The term *dimension* in predictive model development is roughly associated with the complexity of the modeling technique that directly signifies the degree of descriptors. The dimension of an object can be mathematically attributed to the minimum number of coordinates needed for specifying a particular point in it [1]. The addition of dimension to a specific geometric object assists in identifying it in a different way by adding more information. Thus, it is clear that dimension is an intrinsic property of an object and does not depend on the space of the object [1]. The addition of new dimensions to the QSAR technique helps in deriving structural information at a higher level of analysis. With the use of ascending dimensions of descriptors in the modern QSAR analysis, a QSAR modeler may be able to reveal new features of the molecules. The dimensionality of descriptors depends on the type of algorithm employed and defines the nature of QSAR analysis. In the development of a predictive model, the dimension is assigned on the basis of the nature of the independent variables (descriptors) and the corresponding QSAR modeling is named likewise; that is, a QSAR model comprising of one-dimensional (1D) parameters is called *1D-QSAR*. In other words, one can conclude that the dimension of the performed QSAR analysis follows the dimension of the descriptor.

In order to pursue a quantitative analysis on structure of chemical compounds, generation of data encoding chemical information is an essential first step in the development of the QSAR model. It is therefore envisaged that QSAR analysis attempts to develop predictive models in the form of mathematical relations by using chemical information about molecules. Descriptors represent the chemical information that encodes the behavior of a molecular entity. They are the numerical or quantitative representations of chemical compounds derived using suitable algorithms and are used as independent variables for predictive model development. In summary, any apt structural information quantitatively describing the biological activity/property/toxicity of a molecule can be defined as a descriptor. Hence, molecular descriptors range from simple atomic counts or molecular weight measures to complex spatial or geometrical features [5].

One can describe a single molecule in many ways. It is possible to compute thousands of numerical descriptors for a given chemical. Many of these descriptors are very closely related to each other and even capture the same information at times. Thus, the selection of relevant descriptors is a well-known problem, and it requires a lot of experience for the QSAR modeler to select the appropriate ones for the model development [6]. In addition, one has to take into account the nature of the chemical structure being considered. A set of descriptors may efficiently encode the chemical information perfectly for the small molecules, but the same set of descriptors may not be able to encode the required features for polymers, protein structures, and inorganic molecules. Thus, not only the calculation but also the selection of suitable descriptors requires a lot of knowledge and experience in QSAR model development.

Counts of types of atoms or bonds can be considered as constitutional descriptors that only consider atom and bond labels of the compound. Topological descriptors take into account connectivity and labeled graph theory [7]. An advantage of these descriptors is that they do not require exhaustive three-dimensional (3D) coordinate generation and conformational analysis. On the contrary, geometric and 3D descriptors require a 3D structure as input, and therefore the analysis needs to be extensive in order to get the output. Geometrical descriptors are those that describe the molecular shape that is a key factor in ligand–receptor interactions and is an important approach to virtual screening. A variety of descriptors have been invented to characterize molecular shape. Apart from the numerical descriptors, other helpful descriptors are called *fingerprint descriptors.* Conventionally, these descriptors are symbolized in the form of bit strings. In case of inorganic materials, traditional small molecule descriptors are not at all useful. In recent times, a number of periodic table–derived descriptors for inorganic materials, which are very similar to constitutional descriptors, are proving to be helpful [8,9].

This discussion has focused on various facets of descriptors that make them useful in developing QSAR models. It is interesting to point out that the efficacy of a descriptor can rely heavily on the problem being considered. More precisely, certain end points may need to take into account exact molecular features. The best possible features that make a descriptor ideal for the construction of a QSAR model are summarized here:

1. A descriptor must be correlated with the structural features for a specific end point and show negligible correlation with other descriptors.
2. A descriptor should be applicable to a broad class of compounds.
3. A descriptor that can be calculated rapidly and does not depend on experimental properties can be considered more suitable than one that is computationally exhaustive and relies heavily on experimental results.
4. A descriptor should generate dissimilar values for structurally different molecules, even if the structural differences are small. This means that the descriptor should show minimal degeneracy. In addition to degeneracy, a descriptor should be continuous. It signifies that small structural changes should lead to small changes in the value of the descriptor.
5. It is always important that the descriptor has some form of physical interpretability to encode the query features of the studied molecules.
6. Another significant aspect is the ability to map descriptor values back to the structure for visualization purposes [10]. These visualizations are sensible only when descriptor values can be associated to structural features.

Different features of ideal descriptors are summarized in a graphical way in Figure 2.2.

Figure 2.3 illustrates an outline of the types of descriptors and the form of molecular structure required to calculate them. Here, the representation is very general and focuses only on small molecule descriptors. In the following section, various molecular descriptors are discussed for different chemical entities, not just only for small organic molecules.

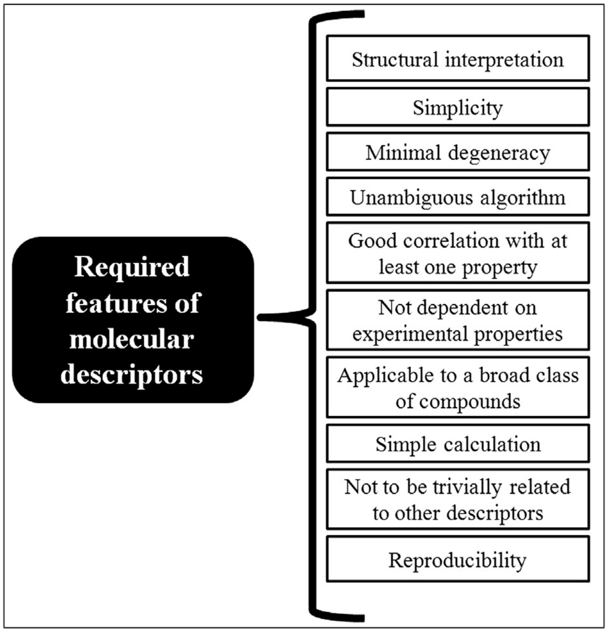

Required features of molecular descriptors
Structural interpretation
Simplicity
Minimal degeneracy
Unambiguous algorithm
Good correlation with at least one property
Not dependent on experimental properties
Applicable to a broad class of compounds
Simple calculation
Not to be trivially related to other descriptors
Reproducibility

Figure 2.2 Ideal features of descriptors for the development of the QSAR model.

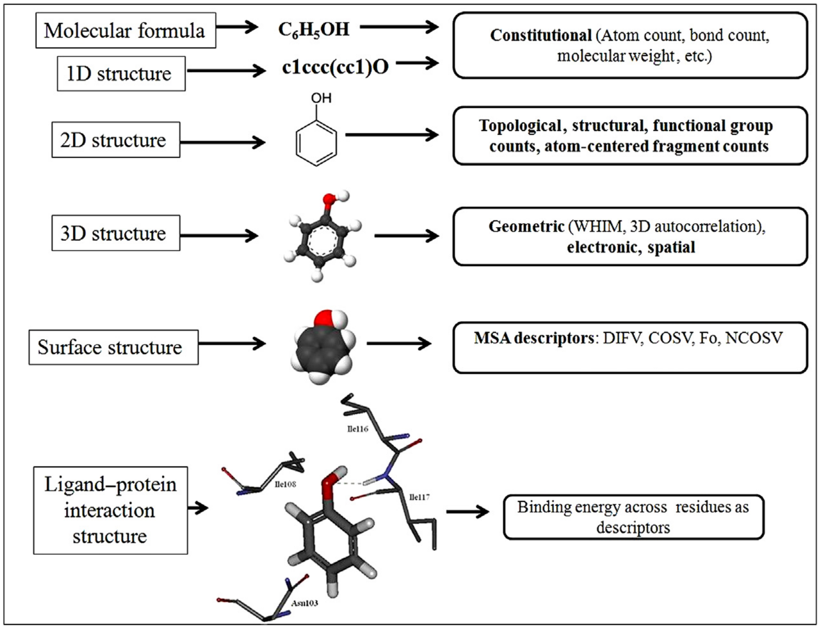

Molecular formula
C6H5OH
1D structure
c1ccc(cc1)O
Constitutional (Atom count, bond count, molecular weight, etc.)
2D structure
OH
Topological, structural, functional group counts, atom-centered fragment counts
3D structure
Geometric (WHIM, 3D autocorrelation), electronic, spatial
Surface structure
MSA descriptors: DIFV, COSV, Fo, NCOSV
Ligand–protein interaction structure
Ile116
Ile108
Ile117
Asn103
Binding energy across residues as descriptors

Figure 2.3 An outline of the types of descriptors can be calculated from the different forms of molecular structure.

2.3 TYPE OF DESCRIPTORS

Descriptors can be classified in multiple ways. In general, there are several types of descriptors like structure explicit descriptors (topological), structure implicit (hydrophobicity and electronic), and cryptic descriptors (quantum chemical). It is interesting to point out that the majority of QSAR researchers prefer to classify the types of descriptors in respect to their dimensions. Considering this aspect, Table 2.1 gives a useful illustration of largely used molecular descriptors based on dimensions [5].

In a broader perspective, descriptors (specifically, physicochemical descriptors) can be classified into two major groups: (1) substituent constants and (2) whole molecular descriptors.

2.3.1 Substituent constants

QSAR grew out of physical organic chemistry based on studies to show how differential reaction rates of chemical reactions depend on the differences in molecular structure. Characterization of these differences in structure, which are due to functional group substitutions into a fixed core structure, led to the development of substituent constants [11]. The substituent constants may encode the electronic, hydrophobic, and steric aspects of a series of compounds by which QSAR models can be generated. Substituent constants are basically physicochemical descriptors that are designed on

Table 2.1 Commonly used molecular descriptors based on different dimensions

Dimension of descriptors	Parameters
0D descriptors	Constitutional indices, molecular property, atom, and bond count.
1D descriptors	Fragment counts, fingerprints.
2D descriptors	Topological, structural, physicochemical parameters including thermodynamic descriptors.
3D descriptors	Electronic, spatial parameters, MSA parameters, MFA parameters, RSA parameters.
4D descriptors	Volsurf, GRID, Raptor, etc. derived descriptors.
5D descriptors	These descriptors consider induced-fit parameters and aim to establish a ligand-based virtual or pseudoreceptor model. These can be explained as 4D-QSAR + explicit representation of different induced-fit models. Example: flexible-protein docking.
6D descriptors	These are derived using the representation of various solvation circumstances along with the information obtained from 5D descriptors. They can be explained as 5D-QSAR + simultaneous consideration of different solvation models. Example: Quasar.
7D descriptors	They comprise real receptor or target-based receptor model data.

the basis of factors, which govern the physicochemical properties of chemical entities. Due to changes in physicochemical properties, absorption, distribution, and transport of chemical entities may be changed.

2.3.2 Whole molecular descriptors

Advancement of structural representation and exploitation of chemical structures have led to the generation of novel methods for representing entire molecular structures. Many of the whole molecule descriptors are expansions of the substituent constant approach, but many of them are also derived from entirely new approaches or from experiments. There are now many commercial molecular modeling programs that can produce descriptors from the whole molecule. Some examples of most commonly used whole molecule descriptors in the QSAR study include the octanol–water partition coefficient, acidic dissociation constant (pK_a), and van der Waals volume (Vw).

2.4 DESCRIPTORS COMMONLY USED IN QSAR STUDIES

Most commonly used descriptors in the QSAR studies are elaborately described in the following sections.

2.4.1 Physicochemical descriptors

These descriptors are derived from the results of some physicochemical experimental findings, and they have connections with the physicochemical properties of the molecules.

2.4.1.1 Hydrophobic parameters

2.4.1.1.1 Partition coefficient (log *P*)

The relative affinity of a drug molecule for an aqueous or lipid medium is important for the drug's activity because absorption, transport, and excretion depend on partitioning phenomena [12,13]. The most widely used molecular descriptor to encode this property is the logarithm of the partition coefficient, *P*, between *n*-octanol and water:

$$P = [C]_{\mathrm{octanol}}/[C]_{\mathrm{aqueous}} \tag{2.1}$$

In Eq. (2.1–2.3), $[C]_{\mathrm{octanol}}$ is the concentration of a solute in the lipid phase (*n*-octanol) and $[C]_{\mathrm{aqueous}}$ is the concentration of the solute in the aqueous phase. Compounds for which $P > 1$ are lipophilic or hydrophobic, and compounds for which $P < 1$ are hydrophilic. Lipophilicity represents the affinity of a molecule or a moiety for a lipophilic environment. It is commonly measured by its distribution behavior in a biphasic system, either liquid–liquid or solid–liquid. In addition, log *P* values can also be computed based on atomic/fragmental contributions and various correction factors, and several such algorithms like $C\log P$, $A\log P$, $A\log P98$, $M\log P$, and $X\log P$ are available.

One of the important methods for calculating $\log P$ from the molecular structure is substituent additivity based on fragmental and atomic contributions, with consideration of surface area, molecular properties, and solvatochromic parameters [14]. In this method, after summing fragment constants for the molecule in query, any necessary correction factors for intramolecular interactions between the fragments, such as electronic, steric, or hydrogen-bonding effects, are added. This fragment addition method developed by Hansch et al. [14] is widely used in recent times. Here, the $\log P$ of a compound is computed by summing the contributions for the fragments and then applying a number of correction factors as needed:

$$\log P = \sum_i a_i f_i + \sum_j b_j f_j \tag{2.2}$$

where f_i are fragment constants and f_j are correction factors. The contribution for each fragment f_i multiplied by the occurrence of that fragment (a_i) is added cumulatively. This sum is then corrected for a number of factors according to the solvent theory. Each correction factor has an associated value, f_j, and this is multiplied by the number of instances of the correction in the structure, b_j. Correction factors include those due to molecular flexibility, branching, polar fragment interaction factors, *ortho* effects, and aromatic interactions [14].

2.4.1.1.2 Hydrophobic substituent constant (π)

Hydrophobicity [12,13] is the association of nonpolar groups or molecules in an aqueous environment, which arises from the tendency of water to exclude nonpolar molecules. The hydrophobicity of the compounds in the series can be represented on a relative scale with the hydrophobic substituent constant π. The value for the substituent X is defined as follows:

$$\pi_X = \log P_X - \log P_H \tag{2.3}$$

In Eq. (2.3), P_X is the partition coefficient of the derivative and P_H is the partition coefficient of the parent compound. The variable π_X expresses the variation in lipophilicity, which results when the substituent X replaces H in RH. For example, the value of the chloro substituent π_{Cl} is the difference between the partition coefficient values of chlorobenzene and benzene. As another example, one may note that the $\log P$ values for benzene and benzamide are 2.13 and 0.64, respectively. Since benzene is the parent compound, the substituent constant for $CONH_2$ is -1.49 (Figure 2.4).

A positive value of π indicates that the substituent is more hydrophobic than hydrogen, and a negative value indicates that the substituent is less hydrophobic. The π value

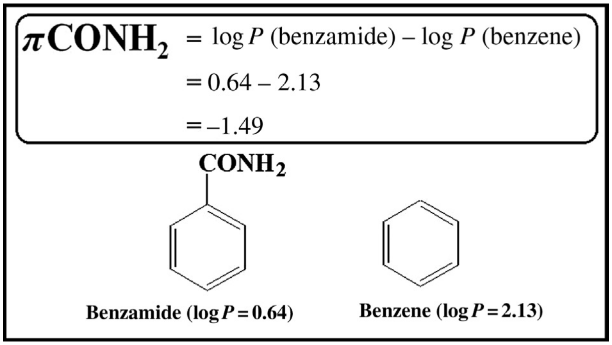


Figure 2.4 Sample calculation of the hydrophobic substituent constant for $CONH_2$ group.

is a characteristic for an individual substituent and can be used to calculate how the partition coefficient of a drug would be affected by adding that particular substituent.

2.4.1.1.3 Hydrophobic fragmental constant (*f*, *f′*)

The hydrophobic fragmental constant of a substituent or molecular fragment represents the lipophilicity contribution of that molecular fragment [12–14].

2.4.1.2 Electronic parameters

Electronic substituent constants were developed as a direct result of an empirical observation made from certain chemical systems that substituents have the same relative effects on the rates of reaction equilibria, regardless of which reaction was being studied [13,15].

2.4.1.2.1 Acid dissociation constant

An important whole molecular parameter defining electronic nature of the tested molecules is the acid dissociation constant [5], which can be explained by the following equation:

$$K_a = \frac{[A^-][H^+]}{[HA]} \tag{2.4}$$

where A^- is the conjugate base of acid HA and H^+ is the proton. The negative logarithmic function (pK_a) is used for modeling purposes and can be defined as $pK_a = -\log_{10} K_a$. It is usually determined using the famous Henderson–Hasselbalch equation:

$$pK_a = pH - \log\frac{[A^-]}{[HA]} \tag{2.5}$$

where pH is the negative logarithmic concentration of H^+ ion; that is, $pH = -\log[H^+]$.

2.4.1.2.2 Hammett constant

Hammett proposed the electronic substituent constant from the rate constants of ionization reaction of *meta*- and *para*-substituted benzoic acid derivatives:

$$\log \frac{k}{k_0} = \rho\sigma \tag{2.6}$$

In Eq. (2.6), the slope ρ is a proportionality reaction constant pertaining to a given equilibrium that relates the effect of substituents on that equilibrium to the effect on the benzoic acid equilibrium. The parameter σ describes the electronic properties of aromatic substituents; that is, electron withdrawing or donating power. The constant k_0 refers to the rate constant for the unsubstituted compound, while k refers to that of a *meta*- or *para*-substituted congener. The substituent constant, σ, reflects the intrinsic polar effect of a given substituent relative to hydrogen. This effect is independent of the reaction.

One of the limitations of the Hammett constant is that it does not hold for *ortho* substituents. This characteristic is known as the *ortho effect*. Taft and Newman [16] proposed a quantitative measure for separating the inductive influence of a substituent from its steric effect. The substituent constant σ^* is based on the rates of acid- and base-catalyzed hydrolysis of esters of the form X—CH$_2$—COOR:

$$\sigma^* = (1/2.48)[\log(k/k_0)_{\mathrm{BASE}} - \log(k/k_0)_{\mathrm{ACID}}] \tag{2.7}$$

where X = H for k_0. Taft argued that σ^* should measure only the inductive influence of a substituent.

There are several other variants of electronic substituent constants. For example, σ^- may be used instead of σ_{p} when cross-conjugation with an electron-withdrawing substituent occurs. Similar to the σ^- constants for electron-withdrawing groups, there is another variant σ^+, which can be used for groups that release electron density via resonance. The *meta*-substituted systems are employed to find out the appropriate reaction constant ρ, and these, in turn, may be used to find the normalized substituent constant, σ^0, for *para* substituents. The σ^0 values indicate that electron donor groups such as —N(CH$_3$)$_2$, —NH$_2$, and —OCH$_3$ have much less influence in *p*-substituted phenylacetic acid systems than in the corresponding *para*-substituted benzoic acids. This helps to prove that cross-conjugation is a significant component of the Hammett constants for such substituents.

Efforts were made to include inductive and resonance components in the general quantitation for the electronic effects, which can be shown as follows:

$$\sigma_{\mathrm{p}} = \sigma_{\mathrm{I}} + \sigma_{\mathrm{R}} \tag{2.8}$$

$$\sigma_{\mathrm{m}} = \sigma_{\mathrm{I}} + \alpha\sigma_{\mathrm{R}} \tag{2.9}$$

In Eqs. (2.8) and (2.9), σ_I is the field/inductive component, which is a scaled version of Taft's σ^* parameter; α is the transmission effect; and σ_R is the resonance component.

2.4.1.3 Steric parameters

In a homologous series of compounds, different biological activities for the compounds are often related to the size of the substituents [13,16,17]. Bulky substituents can interfere with the intermolecular reactions, which lead to the drug's activity. The quantitative encoding of the steric aspect of the drug structure can be accomplished by a series of steric substituent constants.

2.4.1.3.1 Taft steric constant

The first steric parameter to be quantified and used in QSAR studies was Taft's steric (E_S) constant [16,17], which was proposed as a measure of steric effects that a substituent X exerts on the acid-catalyzed hydrolytic rate of esters of substituted acetic acids XCOOR. The basic assumption is that the effect of X on acid hydrolysis is purely steric, as the reaction constant ρ for acid hydrolysis of substituted esters is close to zero. It was a modification to the Hammett constant equation. While the Hammett equation accounts for how field, inductive, and resonance effects influence the reaction rates, the Taft equation describes the steric effects of a substituent. E_S is defined as

$$E_S = \log(k_x)_A - \log(k_{CH3})_A = \log(k_x/k_{CH3})_A \qquad (2.10)$$

where k_x and k_{CH3} are the rate constants for the substituted (substituent is X) and unsubstituted ($X = CH_3$) esters or acids, respectively; and the subscript A denotes hydrolysis in acid solution. The bulkier is the substituent, the more negative are the E_S constant values.

2.4.1.3.2 Charton's steric parameter (ν) and van der Waals radius

Charton found that Taft's steric (E_S) constant is linearly dependent on the van der Waals radius of the substituent, which led to the development of Charton's steric parameter (υ_X) [18]. Taft also pointed out that E_S varies parallel to the atom group radius. Charton's steric parameter can be defined as

$$\upsilon_X = r_X - r_H = r_X - 1.20 \qquad (2.11)$$

where r_X and r_H are the minimum van der Waals radii of the substituent and hydrogen, respectively.

Charton's steric parameter is related to the van der Waals radius of any symmetrical substituent or to the minimum width of unsymmetrical ones. To overcome the problem of asymmetrical substituents, Verloop introduced the STERIMOL set of five parameters, which is explained in a later section of this chapter [13,16].

2.4.1.3.3 Effective Charton's steric parameter (υ_{ef})

Taft developed his steric effect constants based on the assumption that rates of esterification of carboxylic acids with alcohols and of acid-catalyzed hydrolysis of carboxylate esters were structurally reliant on steric effects. This assumption was supported by the work of Charton, who showed that the Taft E_S values for H, a no conformational dependence (NCD) group, and for CH_3, CCl_3, CBr_3, and CF_3, minimal conformational dependence (MCD) groups, were well correlated by Eq. (2.12):

$$E_{S,X} = a_1 r_{\upsilon,min,X} + a_0 \tag{2.12}$$

where *min* denotes minimum radius. Effective values, υ_{ef}, of the steric parameter for some intermediate conformational dependence (ICD) groups have been obtained by means of a two-step procedure:

1. The log k_X values, which are available for NCD and MCD groups, are correlated with Eq. (2.13):

$$\log(k_X)_A = s\upsilon_{min,X} + h \tag{2.13}$$

2. The s and h values obtained from this correlation are used to calculate the values of υ_{eff} from Eq. (2.14):

$$\upsilon_{eff,X} = \frac{\log(k_X)_A - h}{s} \tag{2.14}$$

The values of υ_{eff} for some frequently occurring substituents are given in Table 2.2.

2.4.1.3.4 STERIMOL parameters

In an attempt to go beyond the Taft parameter, which was designed for simple homogenous organic reactions, Verloop [19] designed a multiparametric method for characterizing the steric features of the substituents in more complex biological systems. Verloop [19] developed the STERIMOL parameters, which are a set of five descriptors (L, $B1$, $B2$, $B3$, and $B4$), to describe the shape of a substituent. L is the length of the substituent along the axis of the bond between the first atom of the substituent and the parent molecule. The width parameters $B1-B4$ are all orthogonal to L and form angles of 90° to each other. The large number of parameters required to define each substituent and the large number of compounds necessary to incorporate all the parameters in a QSAR resulted in pruning of the descriptors to L, $B1$, and $B5$, with $B1$ as the smallest and $B5$ the largest width parameter, which does not have any directional relationship to L [19]. Verloop STERIMOL parameters for the carboxylic acid group in *para*-hydroxy benzoic acid are presented in Figure 2.5.

Table 2.2 υ_{ef} values for common substituents

Groups	υ_{ef}	Groups	υ_{ef}
CH_3	0.52	CH_3CHOH	0.50
CH_2CH_3	0.56	C_2H_5CHOH	0.71
$CH_2CH_2CH_3$	0.68	C_6H_5CHOH	0.69
$CH_2CH_2CH_2CH_3$	0.68	CH_2NH_2	0.54
$C_6H_5(CH_2)_3$	0.70	CH_3CHNH_2	0.58
CH_2F	0.62	$C_2H_5CHNH_2$	0.89
CH_2Cl	0.60	H	0
CH_2Br	0.64	F	0.27
CH_2I	0.67	Cl	0.55
CHF_2	0.68	Br	0.65
$CHCl_2$	0.81	I	0.78
$CHBr_2$	0.89	OH	0.32
CHI_2	0.97	NH_2	0.35
CF_3	0.90	SH	0.60
CCl_3	1.38	CH_3OCH_2	0.63
CBr_3	1.56	$CH_3CH_2OCH_2$	0.61
CI_3	1.79	$CH_3OCH_2CH_2$	0.89
CH_2CH_2Cl	0.97	CH_2OH	0.53
CH_2CH_2Br	0.92	CH_2CH_2OH	0.77
CH_2CH_2I	0.93	$C_3H_7OCH_2$	0.65

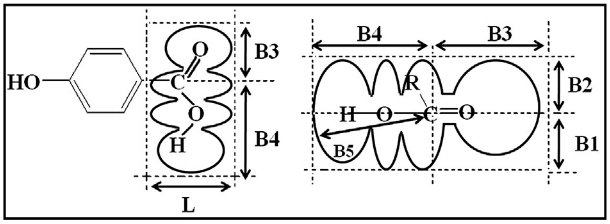


Figure 2.5 Verloop STERIMOL parameters for the carboxylic acid group of *para*-hydroxy benzoic acid.

2.4.1.3.5 Molar refractivity

The molar refractivity (MR) [13,16] is the molar volume corrected by the refractive index. It represents the size and polarizability of a fragment or a molecule. In an atom-based approach, each atom of the molecule is assigned to a particular class, with additive contributions to the total value of MR:

$$\mathrm{MR} = \left[\frac{(n^2 - 1)}{(n^2 + 2)}\right]\left(\frac{\mathrm{MW}}{d}\right) \tag{2.15}$$

In Eq. (2.15), n is the refractive index, MW is the molecular weight, and d is the density of the compound.

2.4.1.3.6 Parachor

An important whole molecular parameter defining the steric nature is parachor, which can be derived by the following equation:

$$PA = \gamma^{1/4} \cdot \frac{MW}{\rho_L - \rho_V} \tag{2.16}$$

where γ is the surface tension of the liquid, MW is the molecular weight, and ρ_L and ρ_V are the densities of the liquid and vapor states, respectively. Parachor depends on molecule volume [5].

We have enlisted here some of the most commonly used substituent constant parameters for a representative list of common aromatic substituents (Table 2.3).

2.4.2 Topological descriptors

Topological descriptors are calculated based on the graphical representation of the molecules and thus neither require estimation of any physicochemical parameters nor need the rigorous calculations involved in the estimation of the quantum chemical descriptors. The structure representation of the molecule depends on its topology, which indicates the position of the individual atoms and the bonded connections between them. Topological indices are computed by applying a specific algorithm using information obtained from the hydrogen-suppressed graph; that is, the number of elements defining it and their connectivity information [20]. The graph theoretic determination of the molecular structure involves the covalently bonded compounds, considering atoms as the vertices and bonds as the edges. Various topological indices constitute a major portion of the development of successful and predictive QSAR models. Computation of such parameters is very fast and efficient, as they require only hydrogen-suppressed 2D-structural information of the molecule under consideration [21,22]. Their simplicity and easy calculability make them useful for studying large databases of compounds.

In Table 2.4, we list the most commonly used topological descriptors [5,20–38] along with their formal mathematical definitions. The topological descriptors are discussed more thoroughly in Chapter 4 due to their immense application in QSAR model development.

2.4.3 Structural descriptors

Structural parameters [5] are classified and described in Table 2.5.

2.4.4 Indicator variables

Indicator variables have been employed in QSAR models due to their simplicity. Substructure descriptors can be easily employed as indicator variables. Two sets of

Table 2.3 Representative list of substituent constant parameters for common aromatic substituents

Substituent	π	MR	σ_m	σ_p	L	$B1$	$B2$	$B3$	$B4$	$B5$
H	0.00	0.103	0.00	0.00	2.06	1.00	1.00	1.00	1.00	1.00
CH_3	0.56	0.565	−0.07	−0.17	3.00	1.52	2.04	1.90	1.90	2.04
CH_2CH_3	1.02	1.030	−0.07	−0.15	4.11	1.52	2.97	1.90	1.90	3.17
CH_2OH	−1.03	0.719	0.00	0.00	3.97	1.52	2.70	1.90	1.90	2.70
CH_2CN	−0.57	1.011	0.16	0.01	3.99	1.52	4.12	1.90	1.90	4.12
CH_2Cl	0.17	1.049	0.11	0.12	3.89	1.52	3.46	1.90	1.90	3.46
CH_2Br	0.79	1.339	0.12	0.14	4.09	1.52	3.75	1.95	1.95	3.75
CH_2I	1.50	1.886	0.10	0.11	4.36	1.52	4.15	2.15	2.15	4.15
$CH_2C_6H_5$	2.01	3.001	−0.08	−0.09	3.63	1.52	6.02	3.11	3.11	6.02
$CH(CH_3)_2$	1.53	1.496	−0.07	−0.15	4.11	2.04	2.76	3.16	3.16	3.17
n-C_3H_7	1.55	1.496	−0.07	−0.13	5.05	1.52	3.49	1.90	1.90	3.49
n-C_4H_9	2.13	1.969	−0.08	−0.16	6.17	1.52	4.42	1.90	1.90	4.54
C_5H_{11}	2.67	2.426	−0.08	−0.16	7.11	1.52	4.94	1.90	1.90	4.94
C_6H_5	1.96	2.536	0.06	−0.01	6.28	1.70	1.70	3.11	3.11	3.11
$COCH_3$	−0.55	1.118	0.38	0.50	4.06	1.90	1.90	2.36	2.93	3.13
$CONH_2$	−1.49	0.981	0.28	0.36	4.06	1.60	1.60	2.42	3.07	3.07
COC_6H_5	1.05	3.033	0.34	0.43	4.57	2.36	5.98	3.11	3.11	5.98
OH	−0.67	0.285	0.12	−0.37	2.74	1.35	1.93	1.35	1.35	1.93
$OCOCH_3$	−0.64	1.247	0.39	0.31	4.87	1.35	3.68	1.90	1.90	3.68
OCH_3	−0.02	0.787	0.12	−0.27	3.98	1.35	2.87	1.90	1.90	3.07
OCH_2CH_3	0.38	1.247	0.10	−0.24	4.92	1.35	3.36	1.35	1.90	3.36
OC_3H_7	0.85	1.706	0.10	−0.25	6.05	1.35	4.30	1.90	1.90	4.42
OC_4H_9	1.55	2.166	0.10	−0.32	6.99	1.35	4.79	1.90	1.90	4.79
OC_6H_5	2.08	2.768	0.25	−0.03	4.51	1.35	5.89	3.11	3.11	5.89
CH_2OCH_3	−0.78	1.207	0.02	0.03	4.91	1.52	2.88	1.90	1.90	3.41
CHO	−0.65	0.688	0.35	0.42	3.53	1.60	1.60	2.00	2.36	2.36
COOH	−0.32	0.693	0.37	0.45	3.91	1.60	1.60	2.36	2.66	2.66
CN	−0.57	0.633	0.56	0.66	4.23	1.60	1.60	1.60	1.60	1.60
CF_3	0.88	0.502	0.43	0.54	3.30	1.98	2.61	2.44	2.44	2.61
F	0.14	0.092	0.34	0.06	2.65	1.35	1.35	1.35	1.35	1.35
Cl	0.71	0.603	0.37	0.23	3.52	1.80	1.80	1.80	1.80	1.80
Br	0.86	0.888	0.39	0.23	3.83	1.95	1.95	1.95	1.95	1.95
I	1.12	1.394	0.35	0.18	4.23	2.15	2.15	2.15	2.15	2.15
NO	−1.20	0.520	0.62	0.91	3.44	1.70	2.44	1.70	1.70	2.44
NO_2	−0.28	0.736	0.71	0.78	3.44	1.70	1.70	2.44	2.44	2.44
NH_2	−1.23	0.542	−0.16	−0.66	2.93	1.50	1.50	1.84	1.84	1.97
$N(CH_3)_2$	0.18	1.555	−0.15	−0.83	3.53	1.50	2.56	2.80	2.80	3.08
$NHCH_3$	−0.47	1.033	−0.30	−0.84	3.53	1.50	3.08	1.90	1.90	3.08
NHC_6H_5	1.37	3.004	−0.12	−0.40	4.53	1.50	5.95	3.11	3.11	5.95
$N{=}NC_6H_5$	1.69	3.131	0.32	0.39	8.43	1.70	1.70	1.92	4.31	4.31
SH	0.39	0.922	0.25	0.15	3.47	1.70	2.33	1.70	1.70	2.33
SCH_3	0.61	1.382	0.15	0.00	4.30	1.70	3.26	1.90	1.90	3.26
SO_2CH_3	−1.63	1.349	0.60	0.72	4.37	2.11	3.15	2.67	2.67	3.15
SO_2NH_2	−1.82	1.228	0.46	0.57	3.82	2.11	3.07	2.67	2.67	3.07

Table 2.4 A representative overview of topological descriptors used in QSAR model development

Name	Mathematical definition	Additional information
Wiener index (W)	$W = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\delta_{ij}$ where N is the number of vertices or atoms and δ_{ij} is the distance matrix of the shortest possible path between vertices i and j.	Sum contribution of all connecting bonds in a molecular graph.
Zagreb group indices	$\text{Zagreb} = \sum_{i}\delta_i^2$ where δ_i is the valency of vertex atom i.	Principally depends on vertex adjacency.
Balaban J index	$J = \frac{M}{\mu + 1}\sum_{\text{all edges}}(\delta_i\delta_j)^{-0.5}$ where M is the number of edges, μ represents cyclomatic number, and δ_i (or δ_j) can be defined as: $\delta_i = \sum_{j=1}\delta_{ij}$	Acyclic graph is pruned toward its center in order to obtain the sequences of numbers.
Randic branching index (χ)	$\chi = \sum_{\text{all edges}}(\delta_i\delta_j)^{-0.5}$ where δ_i and δ_j represent the number of other nonhydrogen atoms bonded to atoms (vertices) i and j, respectively, forming an edge ij.	Basic connectivity parameter based on which, various higher order graph connectivities are established.
Molecular connectivity index	${}^m\chi_t = \sum_{j=1}^{n_m}{}^mS_j$ where n_m represents the number of t type subgraphs of order m. The term ${}^mS_j$ may be defined as follows: ${}^mS_j = \prod_{i=1}^{m+1}(\delta_i)_j^{-0.5}$ and δ_i for the ith atom may be defined as $\delta_i = \sigma_i - h_i$, where σ_i is the number of valence electrons in σ orbital of the ith atom and h_i represents the number of hydrogen atoms attached to vertex i.	The δ_i represents the number of skeletal neighbors, whereas the valence delta value explicitly considers the hybridization states of each atom. In cases of saturated carbon systems, that is, alkanes, $\delta_i^v = \delta_i$.

(*Continued*)

Table 2.4 (Continued)

Name	Mathematical definition	Additional information
Valence molecular connectivity index	${}^{m}\chi^{vv}_{t} = \sum_{j=1}^{n_m} {}^{m}S^{vv}_{j}$ Here, the corresponding term δ^{v} is defined as $\delta^{v}_{i} = \frac{(Z^{v}_{i} - h)}{Z - Z^{v}_{i} - 1}$, where Z and Z^{v} are the atomic number and the total number of valence electron respectively for the ith vertex.	
Bond/edge connectivity indices	$\in = \sum_{l=1}^{p_2} [\delta(e_i)\delta(e_j)]_{l}^{-0.5}$ where $\delta(e)$ corresponds to edge degree and is summed (l) over all the p_2 adjacent edges.	This index coincides with Randic connectivity parameter in case of line graph where number of edges equals number of connections.
Extended bond/edge connectivity indices	${}^{m}\in_{t} = \sum_{s} \prod_{i} [\delta(e_i)]_{s}^{-0.5}$ where m represents the order of the index, t is the type of fragment, and $\delta(e_i)$ is the degree of the edge e_i.	They characterize a generalization of the edge connectivity index. The subscript t denotes the subgraph type is represented as follows. ch: chain or ring; pc: path-cluster; c: cluster; p: path
Kappa shape indices	${}^{1}\kappa = 2\frac{{}^{1}P_{\max}\,{}^{1}P_{\min}}{({}^{1}P_{i}}){}^{2};\ {}^{2}\kappa = 2\frac{{}^{2}P_{\max}\,{}^{2}P_{\min}}{({}^{2}P_{i}}){}^{2};\ {}^{3}\kappa = 4\frac{{}^{3}P_{\max}\,{}^{3}P_{\min}}{({}^{3}P_{i}}){}^{2}$ where the numbers of one, two, and three path lengths are denoted by ${}^{1}P_{i}$, ${}^{2}P_{i}$, and ${}^{3}P_{i}$ respectively. Furthermore, the maximum and minimum path lengths of a specific type may be represented in terms of the number of atoms (A) and thus the corresponding Kappa shape indices can be defined as follows: ${}^{1}P_{\max} = (A(A-1))/2;\ {}^{1}P_{\min} = (A-1)$ ${}^{1}\kappa = \frac{A(A-1)^{2}}{({}^{1}P_{i})^{2}};\ {}^{2}\kappa = \frac{(A-1)(A-2)^{2}}{({}^{2}P_{i})^{2}};\ {}^{3}\kappa = \frac{(A-1)(A-3)^{2}}{({}^{3}P_{i})^{2}}$ for odd value of A and ${}^{3}\kappa = \frac{(A-2)^{2}(A-3)}{({}^{3}P_{i})^{2}}$ for even value of A.	Also known as the Kier's shape indices. Here, the shape of a molecule is defined in terms of number of atoms and their bonding pattern. The index Kappa 1 (${}^{1}\kappa$) shows the degree of complexity, Kappa 2 (${}^{2}\kappa$) defines the degree of linearity or "star-likeness," while Kappa 3 (${}^{3}\kappa$) represents the branching degree at the molecular center.

Kappa modified (alpha) shape indices	The Kappa indices are modified by using an α term which is defined as $\alpha_x = \frac{r_x}{r_{\mathrm{Csp}^3}} - 1,$ where r_x and r_{Csp3} are the covalent radii of desired atom and sp^3 hybridized carbon atom, respectively. The corresponding alpha-modified Kappa shape indices are defined here: ${}^1\kappa_\alpha = \frac{(A+\alpha)(A+\alpha-1)^2}{({}^1P_i+\alpha)^2}; {}^2\kappa_\alpha = \frac{(A+\alpha-1)(A+\alpha-2)^2}{({}^2P_i+\alpha)^2};$ ${}^3\kappa_\alpha = \frac{(A+\alpha-1)(A+\alpha-3)^2}{({}^3P_i+\alpha)^2}$ for odd A values and ${}^3\kappa_\alpha = \frac{(A+\alpha-2)^2(A+\alpha-3)}{({}^3P_i+\alpha)^2}$ for even A values.	The basic Kappa shape indices consider equivalency of all atoms, which is avoided here by comparing the atomic radius of individual atoms with C_{sp3} carbon.
E-state index	$S_i = I_i + \Delta I_i$ where I_i is an intrinsic state parameter and ΔI_i is the perturbation factor. Both the terms are defined as $I_i = \frac{[2/N]^2\delta^v + 1}{\delta}$ and $\Delta I_i = \sum_{j \neq i} \frac{(I_i - I_j)}{r_{ij}^2}$ where N is the principal quantum number and r_{ij} being the topological distance between atoms i and j.	Within the electrotopological state atom index, the intrinsic state part denotes the possible partitioning influence of nonsigma (σ) electrons along the path of the chosen atom, whereas the perturbation factor corresponds to an electronegative gradient.
Flexibility index (Kier and Hall's)	$\Phi = \frac{({}^1\kappa_\alpha {}^2\kappa_\alpha)}{A}$ where A is the number of vertices and ${}^1\kappa_\alpha$ and ${}^2\kappa_\alpha$ are the modified Kappa shape indices of the one and two paths respectively.	This index was derived to provide a direct interpretation of degree of linearity, presence of cycles and branching of the studied structural moiety.

(*Continued*)

Table 2.4 (Continued)

Name	Mathematical definition	Additional information
Information theoretic indices	$I = N\log_2 N - \sum_{i=1}^{n} N_i \log_2 N_i$ where the number of elements present in the *i*th set is represented by N_i and the number of different sets of elements are denoted by n.	In order to measure the information content in bits, base-2 logarithmic value has been used to define this index.
Extended topochemical atom (ETA) indices	Definitions of some basic ETA indices are given here: $\alpha = \frac{Z - Z^v}{Z^v} \cdot \frac{1}{PN - 1}, \quad \beta = \Sigma x\sigma + \Sigma y\pi + \delta, \quad \gamma_i = \frac{\alpha_i}{\beta_i},$ $[\eta]_i = \sum_{j \neq i} \left[\frac{\gamma_i \gamma_j}{r_{ij}^2} \right]^{0.5}, \quad \varepsilon = -\alpha + 0.3 \times Z^v, \quad \psi = \frac{\alpha}{\varepsilon}$ where α is the core count, β is the valence electron mobile (VEM) count, γ is the VEM vertex count, η is an atom level index, ε is an electronegativity count, and ψ is a measure of hydrogen-bonding propensity parameter. Z and Z^v are the respective atomic number and valence electron number; PN corresponds to periodic number; σ and π are the representation of sigma and pi bond respectively with their contributions being x and y; δ gives a measure of the resonating lone pair electron in an aromatic system; r_{ij} is the topological distance between two atoms.	The ETA indices were introduced as a refinement of different topologically arrived unique (TAU) scheme indices. All the ETA indices are available under two headings: (a) the basic (first generation) ETA indices and (b) the more novel ETA indices. The ETA indices are thoroughly discussed in Chapter 4.
Subgraph count index	It is the number of subgraphs of a given type and order. Subgraph count index is classified from zero order to third order (SC_0, SC_1, SC_2, SC_3). It is notable that third-order subgraphs are divided into three types on the basis of path, cluster, and ring (SC_3_P, SC_3_C, SC_3_CH). Subgraph count index are thoroughly discussed in Chapter 4.	

Table 2.5 Structural parameters used in the development of QSAR models

Parameter	Explanation
Chiral centers	It counts the number of chiral centers (R or S) in a molecule.
Molecular weight (MW)	It is the simple molecular weight of a chemical entity.
Rotatable bonds (Rotlbonds)	This descriptor counts the number of bonds in the molecule having rotations that are considered to be meaningful for molecular mechanics. All terminal H atoms are ignored.
Hbond donor	It counts the number of groups or moieties capable of donating hydrogen bonds.
Hbond acceptor	This descriptor calculates the number of hydrogen-bond acceptors present in the molecule.

compounds, whose only difference is that a substructure exists in one set but not the other, can be studied as an entire set when using an indicator variable. This creates a model that simultaneously utilizes all other independent variables and then combines the models via the indicator variable. The major limitation of this variable is that this approach should be used only when the two sets of compounds are identical in every respect, except for the substructure being coded with the indicator variable. It can be considered as the extension of a structure-based descriptor.

2.4.5 Thermodynamic descriptors

The most commonly used thermodynamic descriptors in QSAR models are described in Table 2.6 [5].

Table 2.6 Thermodynamic parameters used in the development of QSAR models

Descriptor	Description
*A*log*P*	Log of the partition coefficient using Ghose and Crippen's method
*A*log*P*98	The *A*log*P*98 descriptor is an implementation of the atom-type-based *A*log*P* method
*A*logp_atypes	The 120 atom types defined in the calculation of *A*log*P*98 are available as descriptors. Each *A*log*P*98 atom-type value represents the number of atoms of that type in the molecule.
Fh2o	Desolvation free energy for water derived from a hydration shell model developed by Hopfinger
Foct	Desolvation free energy for octanol derived from a hydration shell model developed by Hopfinger
Hf	Heat of formation

2.4.6 Electronic parameters

Electronic descriptors [5] are used to describe electronic aspects of both the whole molecule and particular regions, such as atoms, bonds, and molecular fragments. Electronic charges in the molecule are the driving force of electrostatic interactions, and it is well known that local electron densities or charges play a fundamental role in many chemical reactions and physicochemical properties. The electronic descriptors are summarized in Table 2.7.

Table 2.7 Electronic parameters used in the development of QSAR models

Parameter	Explanation
Sum of atomic polarizabilities	It is the summation of atomic polarizabilities (A_i). The polarizabilities are calculated as follows: $P_a = \sum_i A_i$
Dipole moment (Dipole)	This 3D descriptor represents the strength and orientation behavior of a molecule in an electrostatic field. Both the magnitude and the components (X, Y, and Z) of the dipole moment are calculated. It is determined by using partial atomic charges and atomic coordinates.
Highest occupied molecular orbital (HOMO) energy	This is the highest energy level in the molecule that contains electrons. It governs molecular reactivity and properties. When a molecule acts as a Lewis base (an electron-pair donor) in bond formation, the electrons are supplied from this orbital. It measures the nucleophilicity of a molecule.
Lowest unoccupied molecular orbital (LUMO) energy	This is the lowest energy level in the molecule that contains no electrons. It is also important in governing molecular reactivity and properties. When a molecule acts as a Lewis acid (an electron-pair acceptor) in bond formation, incoming electron pairs are received in this orbital. It measures the electrophilicity of a molecule.
Superdelocalizability (S_r)	This is an index of reactivity in aromatic hydrocarbons, represented as follows: $S_r = 2\sum_{j=1}^{m}\left(\frac{c_{jr}^2}{e_j}\right)$ S_r = superdelocalizability at position r, e_j = bonding energy coefficient in jth molecular orbital (eigenvalue), c = molecular orbital coefficient at position r in the HOMO, m = index of the HOMO. The index is based on the idea that early interaction of the molecular orbitals of two reactants may be regarded as a mutual perturbation, so that the relative energies of the two orbitals change together and maintain a similar degree of overlap as the reactants approach one another.

2.4.7 Quantum chemical descriptors

2.4.7.1 Mulliken atomic charges

Charges (e.g., Mulliken atomic charges) computed from structures optimized at different levels of theory may be used as descriptors. Energy minimization may be carried out at different levels of theory: (i) the semiempirical AM1 (or PM3) method, (ii) the Hartree—Fock method at the HF/3-21G(d) level, (iii) Hartree—Fock method at the HF/6-31G(d) level, (iv) B3LYP/6-31 + G(d,p), (v) B3LYP/6-311 + G(2d,p), and (vi) MP2/6-311 + G(2d,p). The output from each level may be used as the input for the next level for energy minimization [39].

2.4.7.2 Quantum topological molecular similarity indices

Quantum topological molecular similarity (QTMS) descriptors focus on bond critical points (BCPs), which occur when the gradient of the electron density, ρ vanishes ($\nabla\rho = 0$) at some point between two bonded nuclei. The electron density at a BCP, denoted by ρ_b, can be related to bond order via an exponential relationship. Seven types of descriptors (ρ, $\nabla^2\rho$, λ, ε, K, G, and equilibrium bond lengths) can be calculated for each of the bonds connecting the adjacent common atoms [40,41]. For the molecules sharing a common skeleton, properties are calculated at each BCP formed by the common atoms. At a BCP, the Hessian of ρ has two negative eigenvalues ($\lambda_1 < \lambda_2 < 0$) and one positive value ($\lambda_3 > 0$). The eigenvalues express local curvature of ρ in a point: negative eigenvalues are curvatures perpendicular to the bond, while the positive eigenvalue measures the curvature along the bond [42,43]. If the positive eigenvalue λ_3 dominates, electron density is accumulated along the bond path toward the nuclei. The descriptor λ_3 gives a measure the σ character of a bond, while the summation of values of $\lambda_1 + \lambda_2$ measure the degree of π character. The Laplacian, denoted by $\nabla^2\rho$, refers to the sum of eigenvalues and is a measure of how much ρ is concentrated ($\nabla^2\rho < 0$) or depleted ($\nabla^2\rho > 0$) in a point. Another descriptor in this series is the ellipticity of a bond, which also measures the degree of π character of a bond together with the susceptibility of the ring bonds to rupture and is defined as, $\varepsilon = \lambda_1/\lambda_2 - 1$. In the QTMS bond descriptor vector, there are two more components: the kinetic energy density $K(r)$ and a more classical kinetic energy $G(r)$. In addition, the equilibrium bond length (R_e) has also been used as one of the descriptors, along with other QTMS descriptors. It has been reported that the BCP descriptors have been successful at translating the predicted electronic effects of orbital theories into observable consequences of variation in bond electron densities [44,45].

2.4.8 Spatial parameters

They comprise a series of descriptors calculated based on the spatial arrangement of the molecules and the surface occupied by the molecules [5].

2.4.8.1 RadofGyration

Radius of gyration (RadofGyration) is a measure of the size of an object, a surface, or an ensemble of points. It is calculated as the root mean square distance of the objects' parts from either its center of gravity or an axis [5]. This can be calculated as per Eq. (2.17):

$$\text{RadofGyration} = \sqrt{\left[\sum \frac{(x_i^2 + y_i^2 + z_i^2)}{N}\right]} \tag{2.17}$$

where N is the number of atoms and x, y, z are the atomic coordinates relative to the center of mass.

2.4.8.2 Jurs descriptors

These descriptors combine shape and electronic information to characterize molecules. These descriptors are calculated by mapping atomic partial charges on solvent-accessible surface areas of individual atoms [46]. The various descriptors included in this category are listed in Table 2.8.

Table 2.8 List of Jurs descriptors used in QSAR model development

Category of descriptors	**Definition/Remarks**
Partial negative surface area (PNSA1)	Sum of the solvent-accessible surface areas (SASAs) of all negatively charged atoms.
Partial positive surface area (PPSA1)	Sum of the SASAs of all positively charged atoms.
Total charge-weighted negative surface area (PNSA2)	Partial negative SASA multiplied by the total negative charge.
Total charge-weighted positive surface area (PPSA2)	Partial positive SASA multiplied by the total positive charge.
Atomic charge-weighted negative surface area (PNSA3)	Sum of the products of atomic SASAs and partial charges over all negatively charged atoms.
Atomic charge-weighted positive surface area (PPSA3)	Sum of the products of atomic SASAs and partial charges over all positively charged atoms.
Difference in charged partial surface area (DPSA1)	Partial positive SASA minus the partial negative SASA.
Difference in total charge-weighted surface area (DPSA2)	Total charge-weighted positive SASA minus the total charge-weighted negative SASA.
Difference in atomic charge-weighted surface area (DPSA3)	Atomic charge weighted positive SASA minus the atomic charge weighted negative SASA.
Fractional charged partial negative surface areas (FNSA1, FNSA2, FNSA3)	They are obtained by multiplication of PNSA1, PNSA2, and PNSA3 descriptors with SASA and then dividing the fraction by 1,000, respectively.

(*Continued*)

Table 2.8 (Continued)

Category of descriptors	Definition/Remarks
Fractional charged partial positive surface areas (FPSA1, FPSA2, FPSA3)	They are obtained by multiplication of PPSA1, PPSA2, and PPSA3 descriptors with SASA and then dividing the fraction by 1,000, respectively.
Surface-weighted charged partial negative surface areas (WNSA1, WNSA2, WNSA3)	They are obtained by multiplication of PNSA1, PNSA2, and PNSA3 descriptors with SASA and then dividing the fraction by 1,000, respectively.
Surface-weighted charged partial positive surface areas (WPSA1, WPSA2, WPSA3)	They are obtained by multiplication of PPSA1, PPSA2, and PPSA3 descriptors with SASA and then dividing the fraction by 1,000, respectively.
Relative negative charge (RNCG)	Partial charge of the most negative atom divided by the total negative charge.
Relative positive charge (RPCG)	Partial charge of the most positive atom divided by the total positive charge.
Relative negative charge surface area (RNCS)	SASA of the most negative atom divided by the relative negative charge.
Relative positive charge surface area (RPCS)	SASA of the most positive atom divided by the relative positive charge.
Total hydrophobic surface area (TASA)	Sum of SASAs of atoms with absolute value of partial charges less than 0.2.
Total polar surface area (TPSA)	Sum of SASAs of atoms with absolute value of partial charges greater than or equal to 0.2
Relative hydrophobic surface area (RASA)	TASA divided by the total molecular SASA.
Relative polar surface area (RPSA)	TPSA divided by the total molecular SASA.

2.4.8.3 *Shadow indices*

These indices help to characterize the shape of the molecules. These are calculated by projecting the molecular surface on three mutually perpendicular planes; that is, *XY*, *YZ*, and *XZ*. These descriptors depend not only on conformation, but also on the orientation of molecules. Molecules are rotated to align principal moments of inertia with *X*-, *Y*-, and *Z*-axes [47]. The various descriptors included in this category are listed in Table 2.9. Projections and embedding rectangles of query molecule in the three principal planes for shadow indices are presented in Figure 2.6.

2.4.8.4 *Molecular surface area*

Molecular surface area is a 3D descriptor that describes the van der Waals area of a molecule. It measures the extent to which a molecule exposes itself to the external environment. It is related to binding, transport, and solubility [5].

Table 2.9 List of shadow descriptors used in QSAR model development

Mode of calculation	Descriptors	Description
Areas of molecular shadows	Shadow-*XY*	Area of the molecular shadow in the *XY* plane (*Sxy*)
	Shadow-*XZ*	Area of the molecular shadow in the *XZ* plane (*Sxz*)
	Shadow-*YZ*	Area of the molecular shadow in the *YZ* plane (*Syz*)
Fractional areas of molecular shadows	Shadow-*XY*fr	Fraction of the area of molecular shadow in the *XY* plane over the area of enclosing rectangle (*Sxy,f*)
	Shadow-*XZ*fr	Fraction of the area of molecular shadow in the *XZ* plane over the area of enclosing rectangle (*Sxz,f*)
	Shadow-*YZ*fr	Fraction of the area of molecular shadow in the *YZ* plane over the area of enclosing rectangle (*Syz,f*)
Extents of molecular shadows	Shadow-*X*length	Length of molecule in the *X* dimension (*Lx*)
	Shadow-*Y*length	Length of molecule in the *Y* dimension (*Ly*)
	Shadow-*Y*length	Length of molecule in the *Z* dimension (*Lz*)
	Shadow-nu	Ratio of largest to smallest dimension

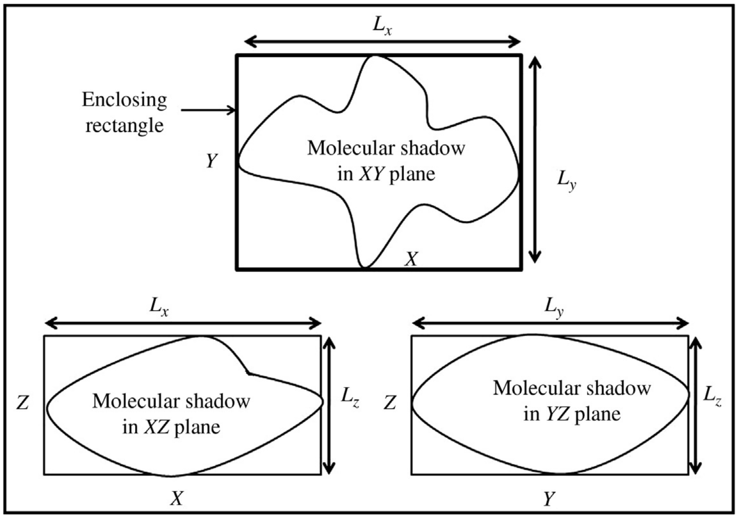

L_x
Enclosing rectangle
Y
Molecular shadow in XY plane
L_y
X
L_x
L_y
Z
Molecular shadow in XZ plane
L_z
Z
Molecular shadow in YZ plane
L_z
X
Y

Figure 2.6 Projections and embedding rectangles of a query molecule in the three principal planes for shadow indices.

2.4.8.5 Density

The 3D descriptor known as density is the ratio of molecular weight to molecular volume. This descriptor represents the type of atoms and how tightly they are packed in a molecule. It is related to transport and melt behavior [5].

2.4.8.6 *Principal moment of inertia*

The moments of inertia are calculated for a series of straight lines through the center of mass. These are associated with the principal axes of the ellipsoid [48]. If all three moments are equal, the molecule is considered to be a symmetrical top.

2.4.8.7 *Molecular volume*

This 3D descriptor known as molecular volume is the volume inside the contact surface [5]. It is related to binding and transport.

2.4.9 Information indices

In this approach, molecules are viewed as structures that can be partitioned into subsets of elements that are in some sense equivalent. The concept of equivalence depends on the particular descriptor. For a partition of a set of N elements into k subsets, each consisting of N_k elements [49–51]:

$$\text{Equivalence class} = 1, 2, \ldots, k$$

The number of elements in each N_1 N_2 ... N_k and N is mathematically represented as

$$N = N_1 + N_2 + N_3 + \cdots + N_k \tag{2.18}$$

For a given partition P, the following relationship is obtained:

$$P = N(N_1, N_2, N_3, \ldots, N_k) \tag{2.19}$$

A probability distribution can be associated with the partition and may be represented as

$$p_i = \frac{N_i}{N_V} \tag{2.20}$$

where p_i is the probability for a randomly chosen element to belong to class i. This degree of uncertainty also can be expressed by the entropy as follows:

$$H_i = -\text{lb}\ p_i \tag{2.21}$$

where H_i is the entropy and lb is the base-2 logarithm. Then the mean entropy of such a probability distribution can be defined as

$$H = -\sum_{i=1}^{k} p_i\ \text{lb}\ p_i \tag{2.22}$$

This parameter can be considered as a measure of the mean quantity of information contained in each structure element (in bits per element) [49].

2.4.9.1 Information of atomic composition index

Here, equivalence classes of the atoms in a molecule are formed by considering their atomic numbers. Two types of information of atomic composition (IAC) descriptors are IAC-mean and IAC-total. The partition of atoms then yields the descriptor IAC-mean as the mean quantity of information H (as defined previously). The descriptor IAC-total is defined as

$$\text{IAC-total} = N \times \text{IAC-mean} \tag{2.23}$$

where N is the number of atoms in the molecule.

2.4.9.2 Information indices based on the A-matrix

The concept is based on partitioning elements of the A-matrix according to two basic modes:

1. The *equality* mode: The matrix elements are considered as equivalent if their values are equal.
2. The *magnitude* mode: This mode assumes that each matrix element is an equivalence class unto itself whose cardinality (number of elements) is equal to the magnitude of the matrix element.

The two information indices in this category are:

1. *Total vertex adjacency/equality (V_ADJ_equ)*: Here, the A-matrix (N-by-N) consists of zeros and 1s, so the partitioning consists of two classes. If C_k represents the number of matrix elements for each class k in the A-matrix, the vertex adjacency/equality will be defined as

$$V\_ADJ\_equ = -N^2 \sum_k \frac{C_k}{N^2} \operatorname{lb} \frac{C_k}{N^2} \tag{2.24}$$

2. *Total vertex adjacency/magnitude (V_ADJ_mag)*: The magnitude descriptor uses the actual A-matrix values a_{ij}, unlike the equality values used for the populations of each class in the A-matrix. Here, the nonzero elements are not included in the expression of V_ADJ_mag:

$$V\_ADJ\_mag = -N^2 \sum_{a_{ij} \neq o} \frac{a_{ij}}{n^2} \operatorname{lb} \frac{a_{ij}}{N^2} \tag{2.25}$$

2.4.9.3 Information indices based on the D-matrix

The information indices based on the D-matrix are similar descriptors like the vertex adjacency indices, but the difference is that the distance matrix is used instead of the adjacency matrix. Two types of indices based on this matrix are:

1. Vertex distance/equality (V_DIST_equ)
2. Vertex distance/magnitude (V_DIST_mag)

2.4.9.4 Information indices based on the E-matrix and the ED-matrix

The information indices based on the E-matrix and ED-matrix are descriptors based on the edge adjacency and the edge distance matrices. The indices based on these matrices are:

1. Edge adjacency/equality (E_ADJ_equ)
2. Edge adjacency/magnitude (E_ADJ_mag)
3. Edge distance/equality (E_DIST_equ)
4. Edge distance/magnitude (E_DIST_mag)

2.4.9.5 Multigraph information content indices (IC, BIC, CIC, SIC)

For multigraph information content, an *unordered* sequence of *ordered* pairs is assigned to each vertex v, termed as a *coordinate*, as follows:

$$\{(m_1, n_1), (m_2, n_2), \ldots, (m_k, n_k)\}$$

where k is the valence of the vertex [with one ordered pair (m_j, n_j) per each neighboring vertex, v_j], and for every $j = 1, \ldots, k$, n_j is the valence of v_j and the bond between v and v_j is of order m_j.

The coordinates are assigned to vertices, and the partition of vertices is constructed in the usual way, in which two vertices are considered equivalent if their coordinates are the same as unordered k-tuples; that is, the repetitions of ordered pairs are not ignored, as they would be if we treated the k-tuples purely as sets.

After the partition, the index is termed as information content (IC). The different classes of IC are as follows [49–51]:

1. Bonding information content (BIC): This index corresponds to the number of bonds counting bond orders, defined as
$$\text{BIC} = \text{IC}/\text{lb} \tag{2.26}$$
2. Structural information content (SIC): This index refers to the number of vertices, defined as
$$\text{SIC} = \text{IC}/\text{lb} \tag{2.27}$$
3. Complementary information content (CIC): This index measures the deviation of IC from its maximum possible value, corresponding to a partition into classes containing one element each. The definitions of IC_{max} and CIC are derived as follows:
$$\text{IC}_{\text{max}} = -N_X(1/N) \times \text{lb}(1/N) = \text{lb}(N) \tag{2.28}$$
$$\text{CIC} = \text{lb}(N) - IC \tag{2.29}$$

2.4.10 Molecular shape analysis descriptors

Different types of molecular shape analysis (MSA) descriptors [5] are summarized in Table 2.10.

Table 2.10 MSA descriptors

Parameter	Explanation
Difference volume (DIFFV)	It is the difference between the volume of the individual molecule and the volume of the shape reference compound.
Common overlap steric volume (COSV)	This is the common volume between each individual molecule and the reference molecule. It is the measurement of similarity of steric shape between analogs to reference compound.
Common overlap volume ratio (Fo)	It is obtained from the ratio of common overlap steric volume to the volume of the individual molecule.
Noncommon overlap steric volume (NCOSV)	It is the volume of the individual molecule and the common overlap steric volume.
Root mean square to shape reference (ShapeRMS)	This is the root mean square deviation between the individual molecule and the shape reference compound.

2.4.11 Molecular field analysis parameters

The molecular field analysis (MFA) [52] formalism calculates probe interaction energies on a rectangular grid around a bundle of active molecules. The surface is generated from a *shape field*. The atomic coordinates of the contributing models are used to compute field values on each point of a 3D-grid. MFA evaluates the energy between a probe (H^+ and CH_3) and a molecular model at a series of points defined by a rectangular grid. The fields of molecules are represented using grids in MFA, and each energy value associated with an MFA grid point can serve as input for the calculation of a QSAR.

2.4.12 Receptor surface analysis parameters

The energies of interaction between the receptor surface model and each molecular model can be used as descriptors for generating QSARs [52]. The surface points that organize as triangle meshes in the construction of the receptor surface analysis (RSA) store these properties as associated scalar values. Receptor surface models provide compact, quantitative descriptors that capture 3D information of interaction energies in terms of steric and electrostatic fields at each surface point.

QSAR has become more attractive for researchers with the development of new and advanced software tools, which have allowed them to determine and understand how molecular structure is responsible for a compound's activity/property/toxicity. Table 2.11 gives a representative list of various software tools used to generate descriptors from molecular structures.

Table 2.11 List of software tools for computation of molecular descriptors

Software	Web link
4D FAP	http://www.ra.cs.uni-tuebingen.de/software/4DFAP/welcome_e.html
ADAPT	http://research.chem.psu.edu/pcjgroup/adapt.html
ADMET Predictor	http://www.simulations-plus.com/Products.aspx?grpID=1&cID=11&pID=13
ADRIANA.Code	http://www.molecular-networks.com/products/adrianacode
Alchemy 2000	http://www.chemistry-software.com/modelling/10235.htm
ALMOND	http://www.moldiscovery.com/soft_almond.php
BlueDesc	http://www.ra.cs.uni-tuebingen.de/software/bluedesc/welcome_e.html
CAChe	http://www.cache.fujitsu.com/cache/index.shtml
Cerius^2	http://accelrys.com/
ChemEnlightenTM	http://www.tripos.com/sciTech/inSilicoDisc/media/LITCTR/CHEMENLI.PDF
CODESSA PRO	http://www.codessa-pro.com/index.htm
Discovery Studio	http://accelrys.com/
DRAGON	http://www.talete.mi.it/products/dragon_description.htm
GRID	http://www.moldiscovery.com/soft_grid.php
JChem	http://www.chemaxon.com/jchem/intro/index.html
JOELib	http://www.ra.cs.uni-tuebingen.de/software/joelib/index.html
ISIDA	http://infochim.u-strasbg.fr/spip.php?rubrique53
MOE	http://www.chemcomp.com/software.htm
MOLCONN-Z	http://www.edusoft-lc.com/molconn/
MOLGEN-QSPR	http://www.molgen.de/?src=documents/molgenqspr.html
OAK	http://www.ra.cs.uni-tuebingen.de/software/OAKernels/welcome_e.html
OASIS QSAR	http://toolbox.oasis-lmc.org/
OpenBabel	http://openbabel.org/
PaDEL-Descriptor	http://padel.nus.edu.sg/software/padeldescriptor/
Pentacle	http://www.moldiscovery.com/soft_pentacle.php
PowerMV	http://nisla05.niss.org/PowerMV/?q = PowerMV/
PreADMET	http://preadmet.bmdrc.org/index.php?option=com_content&view=frontpage&Itemid=1
QSARModel	http://www.molcode.com/
QuaSAR	http://www.chemcomp.com/feature/qsar.htm
RDKit	http://www.rdkit.org/
SciQSAR	http://www.scimatics.com/jsp/qsar/QSARIS.jsp
Sarchitect	http://www.strandls.com/sarchitect/index.html
SYBYL-X	http://tripos.com/index.php?family=modules,SimplePage&page=SYBYL-X
Tsar™	http://www.accelrys.com/products/tsar/tsar.html
Unscrambler X	http://www.camo.com/rt/Products/Unscrambler/unscrambler.html
V-Life MDS	http://www.vlifesciences.com/products/VLifeMDS/Product_VLifeMDS.php

2.5 OVERVIEW AND CONCLUSION

The selection of suitable descriptors from a large pool of diverse classes of descriptors plays a major role in the development of acceptable and robust predictive QSAR models. To some extent, this depends upon the end point to be modeled. The experience of the QSAR researcher also helps in choosing suitable descriptors. An important aspect is to choose the relevant descriptors considering the problem at hand. Again, one must consider, before developing QSAR models, how the descriptors have been calculated and whether calculations can be reproduced. Lower-dimensional parameters like zero-dimensional (0D), 1D, or 2D are easily computable and are used alone or in combination with other higher-dimensional descriptors for successful model development. But due to the complexity of the endpoint and advancement of QSAR studies, as well as the requirements of mechanistic interpretation of the activity/property/toxicity of chemicals, the use of higher-dimensional parameters is increasing every day.

REFERENCES

[1] Katritzky AR, Fara DC, Petrukhin RO, Tatham DB, Maran U, Lomaka A, et al. The present utility and future potential for medicinal chemistry of QSAR/QSPR with whole molecule descriptors. Curr Top Med Chem 2002;2:1333–56.

[2] Guha R, Willighagen EA. Survey of quantitative descriptions of molecular structure. Curr Top Med Chem 2012;12(18):1946–56.

[3] van de Waterbeemd H, Carter RE, Grassy G, Kubinyi H, Martin YC, Tute MS, et al. Glossary of terms used in computational drug design (IUPAC recommendations 1997). Ann Rep Med Chem 1998;33:397–409.

[4] Randic M. On characterization of chemical structure. J Chem Inf Comput Sci 1997;37:672–87.

[5] Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim, Germany: Wiley-VCH; 2000.

[6] Kohavi R, John G. Wrappers for feature subset selection. Artif Intell 1997;97:273–324.

[7] Dehmer M, Varmuza K, Borgert S, Emmert-Streib F. On entropy-based molecular descriptors: statistical analysis of real and synthetic chemical structures. J Chem Inf Model 2009;49:1655–63.

[8] Willems T, Rycroft C, Kazi M, Meza J, Haranczyk M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. Microporous Mesoporous Mater 2012;149(1):134–41.

[9] Mackay A. Descriptors for complex inorganic structures. Croat Chem Acta 1984;57:725–36.

[10] Segall M, Champness E, Obrezanova O, Leeding C. Beyond profiling: using ADMET models to guide decisions. Chem Biodivers 2009;6(11):2144–51.

[11] Livingstone DJ. The characterization of chemical structures using molecular properties. A survey. J Chem Inf Comput Sci 2000;40:195–209.

[12] Taylor PJ. In: Hansch C, Sammes PG, Taylor JB, editors. Comprehensive medicinal chemistry. vol. 4. Quantitative drug design. The rational design, mechanistic study and therapeutic applications of chemical compounds. Oxford: Pergamon Press; 1991. pp. 241–94.

[13] Rekker R. In the hydrophobic fragmental constant. Amsterdam, the Netherlands: Elsevier; 1977.

[14] Hansch C, Leo A, Hoekman D. In exploring QSAR vol. 2: hydrophobic, electronic and steric constants. Washington, DC: ACS; 1995.

[15] Selassie CD, Mekapati SB, Verma RP. QSAR: then and now. Curr Top Med Chem 2002;2(12):1357–79.

[16] Taft RW. In: Newman MS, editor. Steric effects in organic chemistry. New York, NY: John Wiley & Sons; 1956. p. 556.

[17] Hansch C, Leo A. Substituent constants for correlation analysis in chemistry and biology. New York, NY: Wiley; 1979.

[18] Charton M. Steric effects. IV. E1 and E2 eliminations. J Am Chem Soc 1975;97:6159–61.

[19] Verloop A. The STERIMOL approach to drug design. New York, NY: Marcel Dekker; 1987.

[20] García-Domenech R, Gálvez J, de Julián-Ortiz JV, Pogliani L. Some new trends in chemical graph theory. Chem Rev 2008;108:1127–69.

[21] Broto P, Moreau G, Vandycke C. Molecular structure: perception, autocorrelation descriptor and SAR studies. System of atomic contributions for the calculation of the *n*-octanol/water partition coefficients. Eur J Med Chem Chim Ther 1984;19:71–8.

[22] Wold S, Geladi P, Esbensen K, Ohman J. Multiway principal components and PLS analysis. J Chemom 1987;1:41–56.

[23] Estrada E, Ivanciuc O, Gutman I, Gutierreza A, Rodríguez L. Extended wiener indices. A new set of descriptors for quantitative structure–property studies. New J Chem 1998;22:819–23.

[24] Balaban AT. Chemical graphs. Theor Chim Acta 1979;53(4):355–75.

[25] Bonchev D, Balaban AT, Mekenyan O. Generalization of the graph center concept and derived topological centric indexes. J Chem Inf Comput Sci 1980;20:106–13.

[26] Wiener H. Structural determination of paraffin boiling points. J Am Chem Soc 1947;69:17–20.

[27] Randic M. Characterization of molecular branching. J Am Chem Soc 1975;97:6609–15.

[28] Balaban AT. Distance connectivity index. Chem Phys Lett 1982;89:399–404.

[29] Kier LB, Hall LH. Derivation and significance of valence molecular connectivity. J Pharm Sci 1981; 70(6):583–90.

[30] Hall LH, Kier LB. The electrotopological state: an atom index for QSAR. Quant Struct-Act Relat 1991;10:43–51.

[31] Hall LH, Kier LB. The E-state as the basis for molecular structure space definition and structure similarity. J Chem Inf Comput Sci 2000;30:784–91.

[32] Bonchev D, Trinajstić N. Overall molecular descriptors. 3. Overall Zagreb indices. SAR QSAR Environ Res 2001;12(1–2):213–35.

[33] Kier LB. In: Rouvray DH, editor. Computational chemical graph theory. New York, NY: Nova Science Publishers; 1990. pp. 152–74.

[34] Kier LB. A shape index from molecular graphs. Quant Struct-Act Relat 1985;4(3):109–16.

[35] Kier LB. Shape indexes of orders one and three from molecular graphs. Quant Struct-Act Relat 1986;5(1):1–7.

[36] Estrada E. Spectral moments of the edge adjacency matrix in molecular graphs. 1. Definition and applications to the prediction of physical properties of alkanes. J Chem Inf Comput Sci 1996;36:844–9.

[37] Roy K, Das RN. On Extended Topochemical Atom (ETA) indices for QSPR studies. In: Castro EA, Hagi AK, editors. Advanced methods and applications in chemoinformatics: research progress and new applications. Hershey, PA: IGI Global; 2011. pp. 380–411.

[38] Roy K, Ghosh G. Introduction of Extended Topochemical Atom (ETA) indices in the Valence Electron Mobile (VEM) environment as tools for QSAR/QSPR studies. Internet Electron J Mol Des 2003;2(9):599–620.

[39] Mitra I, Roy K, Saha A. QSAR of anti-lipid peroxidative activity of substituted benzodioxoles using chemometric tools. J Comput Chem 2009;30(16):2712–22.

[40] Roy K, Popelier PLA. Predictive QSPR modeling of acidic dissociation constant (pK_a) of phenols in different solvents. J Phys Org Chem 2009;22(3):186–96.

[41] Roy K, Popelier PLA. Exploring predictive QSAR models using Quantum Topological Molecular Similarity (QTMS) descriptors for toxicity of nitroaromatics to *Saccharomyces cerevisiae*. QSAR Comb Sci 2008;27(8):1006–12.

[42] Popelier PLA. Quantum molecular similarity. 1. BCP space. J Phys Chem A 1999;103(15): 2883–90.

[43] Bader RFW, Preston HJT. The kinetic energy of molecular charge distributions and molecular stability. Int J Quantum Chem 1969;3(3):327–47.

[44] Howard ST, Lamarche O. Description of covalent bond orders using the charge density topology. J Phys Org Chem 2003;16(2):133–41.

[45] Bader RFW, Slee TS, Cremer D, Kraka E. Description of conjugation and hyperconjugation in terms of electron distributions. J Am Chem Soc 1983;105(15):5061–8.

[46] Rohrbaugh RH, Jurs PC. Description of molecular shape applied in studies of structure/activity and structure/property relationships. Anal Chim Acta 1987;199:99–109.

[47] Stanton DT, Jurs PC. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure–property relationship studies. Anal Chem 1990;62: 2323–9.

[48] Hill TL. Introduction to statistical thermodynamics. Reading, MA: Addison-Wesley; 1960.

[49] Bonchev D. In: Bawden DD, editor. Information theoretic indices for characterization of chemical structures. Chemometrics series, vol. 5. New York, NY: Research Studies Press Ltd.; 1983.

[50] Bonchev D, Mekenyan O, Trinajstic N. Isomer discrimination by topological information approach. J Comput Chem 1981;2(2):127–48.

[51] Katritzky AR, Gordeeva EV. Traditional topological indices vs. electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. J Chem Inf Comput Sci 1993;33(6):835–57.

[52] Hopfinger AJ, Tokarsi JS. In: Charifson PS, editor. Practical applications of computer-aided drug design. New York, NY: Marcel Dekker; 1997. pp. 105–64.