Chapter 1
QSAR/QSPR Modeling: Introduction

Abstract Development of predictive quantitative structure–activity relationship (QSAR) models plays a significant role in the design of purpose-specific fine chemicals including pharmaceuticals. Considering the wide application of different types of chemicals in human life, QSAR modeling is a useful tool for prediction of biological activity, physicochemical property, and toxicological responses of untested chemical compounds. Descriptors play a crucial role in the development of any QSAR model since they represent quantitatively the encoded chemical information. They not only help in the derivation of a mathematical correlation between the chemical structure information and the response of interest, but also enable exploration of the mechanistic aspect involved in a biochemical process. QSAR analysis is now widely employed as a rational tool for the prediction and design of chemicals of health benefits, industrial/laboratory process, or household applications.

Keywords Descriptors · Physicochemical · Electronic · Structural · Topological · Quantum chemical

1.1 Introduction

Chemistry plays an important role in defining the behavioral manifestations of chemical compounds. Development of suitable techniques which allow modification of the chemical features of molecules is very useful not only in the field of chemistry but also in other branches of natural sciences. Quantitative structure–-activity relationship (QSAR) modeling is one such technique that allows the interdisciplinary exploration of knowledge on compounds covering the aspects of chemistry, physics, biology, and toxicology. It provides a formalism for developing mathematical correlation between the chemical features and the behavioral manifestations of (structurally) similar compounds. The entire technique is defined on the basis of a strong mathematical algorithm, and it provides a reasonable basis for

K. Roy et al., *A Primer on QSAR/QSPR Modeling*,
SpringerBriefs in Molecular Science, DOI 10.1007/978-3-319-17281-1_1

establishing a predictive correlation model. Apart from providing a mathematical correlation, QSAR technique also enables the exploration of chemical features encoded within descriptors. Descriptors being the quantitative numbers represent attributes of the chemicals and aid in the establishment of a mathematical correlation. Hence, different types of descriptors play a significant role in the identification as well as analysis of the chemical basis involved in a process under consideration. The descriptors also allow the user to modify or 'fine-tune' the existing chemical behavior into a desired one by suitable changes in chemical structures. Furthermore, such analysis employs chemical information from relatively small number of chemicals in deriving a mathematical correlation while allows the prediction of the same response for a large number of chemicals. This particular characteristic is highly important when dealing with biological (or toxicological) data that involve ethical issues related to animal experiment. The QSAR technique proves to be a valuable alternative method in this perspective and is encouraged for the design and development of biologically active molecules as well as in predictive toxicology analysis. The QSAR formalism is also widely employed to serve different purposes of material science toward the design and development of purpose-specific novel and/or alternative chemicals. It may be very interesting to note that historically the earliest inception for the ideology of QSAR modeling emerged from the simple concept of a correlation between response and chemical nature of molecules which remains the same even today after various developments and nourishments in the QSAR algorithmic basis. Broadly, the two main purposes of QSAR can be identified as the development of a mathematical equation or model and the explanation of the modeled chemical features as encoded in descriptors. Presently, development of predictive QSAR models on various endpoints is proposed by different international authorities as a reliable tool of exploring chemical knowledge following a rational basis [1].

1.2 What Is QSAR/QSPR Modeling?

1.2.1 Definition and Formalism

QSAR modeling on a set of structurally related chemicals refers to the development of a mathematical correlation between a chemical response and quantitative chemical attributes defining the features of the analyzed molecules. Hence, such study attempts to establish a mathematical formalism between the behavior of a chemical, i.e., chemical response and a set of quantitative chemical attributes which may be extracted from the chemical structures using suitable experimental or theoretical means. The naming of the study depends upon the nature of the response (also known as 'endpoint') being modeled giving rise to three major classes, namely quantitative structure–property/activity/toxicity relationship (QSPR/QSAR/QSTR) studies considering the modeling of physicochemical property, biological

activity, and toxicological data, respectively. The nomenclature can also be employed to define some more specific endpoints such as quantitative structure–-cytotoxicity relationship to denote modeling of cytotoxicity of chemicals. On the other hand, QSPR, i.e., quantitative structure–property relationship modeling, can be employed to designate all such related techniques as any type of biological and toxicological as well as physicochemical behavior may be considered as the 'property' of a given chemical. However, we shall use the term 'QSAR' to denote all such studies. Since a mathematical relationship is developed, such studies allow the prediction of molecular behavior for new chemicals or even hypothetical molecules. Therefore, the basic formalism of QSAR technique can be mathematically represented as follows:

$$\text{Biological activity} = f(\text{Chemical attributes}) \tag{1.1}$$

The basic ideology for the phrase 'chemical attribute' is to denote the features that define the behavioral manifestation, i.e., response of the analyzed chemical compounds. In other words, the chemical attributes are the fundamental information of the chemicals which control the response under study. Since the aim was to develop a mathematical correlation, these features or attributes are precise quantitative chemical information that might be derived using an experimental analysis or suitable theoretical algorithm that diagnoses chemistry of the molecules. Sometimes, information obtained from both the theoretical as well as experimental basis is employed. It is often observed that the behavioral manifestation of any chemical species can be explained by its physicochemical properties which represent the intrinsic molecular nature such as melting point, boiling point, and surface tension. Hence, the chemical attributes in Eq. (1.1) is often described in terms of the information derived directly from the chemical structure and the physicochemical information usually derived using experimental techniques leading to the following expression [1].

$$\text{Response} = f(\text{chemical structure}, \text{physicochemical property}) \tag{1.2}$$

Considering the employment of a series of chemical information in presence/absence of physicochemical features, the QSAR equation for a specific response can be mathematically stated as follows:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \cdots + a_nX_n \tag{1.3}$$

Since we are talking in terms of a mathematical correlation, such equations are better explained in terms of variables. Here, Y is the dependent variable representing the response being modeled, i.e., activity/property/toxicity while $X_1, X_2, \ldots, X_n$ are the independent variables denoting different structural features or physicochemical properties in the form of numerical quantities or descriptors and $a_1, a_2, \ldots, a_n$ are the contributions of individual descriptors to the response with a_0 being a constant. Hence, we can see that the physicochemical properties can not only be employed as a

dependent or response variable giving a structure–property relationship, i.e., QSPR, but they might also be used as independent or predictor variables. QSAR studies may even employ one response parameter, e.g., activity/toxicity as predictor variable for the modeling of another type of activity/toxicity endpoint. Such studies are named as quantitative activity–activity relationship (QAAR) or quantitative toxicity–toxicity relationship (QTTR) or quantitative property–property relationship (QPPR) modeling, as appropriate. It will be interesting to note that although the modeled response should be quantitative in order to develop a regression model, it might also be categorical entities which may be used for development of classification models. However, the predictor variables in QSAR modeling should always be quantitative.

The QSAR analysis is principally aimed at quantification of chemical information followed by developing a suitable interpretative relationship addressing a given response. The extracted chemical or physicochemical information can be utilized for modification of chemical structures leading to the 'fine-tuning' of the properties and biological response, e.g., decreased lipophilicity, enhanced activity, and reduced toxicological manifestation. Thus, mathematics here serves as a tool for deriving a suitable relationship which is then exploited as per the requirement of the designer [2]. On a much broader perspective, QSAR studies encompasses avenues of chemistry and physics accounting for intrinsic molecular nature, mathematics and statistics for modeling and calculation, and biology to encompass the involved biochemical interaction. Thus, predictive mathematical models are developed exploring the knowledge of chemistry and biology in a rational way to meet the desired need of the chemicals. Different concepts and perspectives of mathematics are tacitly used in order to derive predictive QSAR models which may be used for prediction of endpoint data of a large number of untested chemicals. It might be envisaged that the role of mathematics in QSAR analysis is to provide an abstract backbone for developing a characteristic correlation between chemistry and biology of the investigated chemicals.

The QSAR study can be visualized to comprise of three simple steps, namely (a) data preparation, (b) data processing, and (c) data interpretation for a set of chemicals. The quantitative data are obtained from two major components, namely the response or endpoint to be addressed and the predictor or independent variables (i.e., X variables) defining the chemical attributes. The response data can be activity (e.g., anti-malarial, anti-oxidant, anti-arrhythmic, anti-HIV, and anti-cancer), property (e.g., aqueous solubility, *n*-octanol/water partition coefficient, melting point, surface tension, critical micelle concentration value, and chromatographic retention), or toxicological (e.g., organ- or disease-specific acute/chronic toxicity outcomes such as carcinogenicity, skin-irritation, genotoxicity, and hepatotoxicity as well as toxicity toward environment in terms of death of specific indicator organisms such as *Tetrahymena*, daphnids, bacteria, fungi, and fish) behavior of chemical compounds. The first step, i.e., the preparation of data involves arrangement and conversion of the data in a suitable form. The response data for various biological and toxicological endpoints are usually obtained in two forms, namely 'dose-fixed response' pattern where the dose or concentration of a chemical required to produce a desired fixed response is measured and 'response-fixed dose'

pattern in which the response elicited by a chemical at a fixed dose (concentration) is opted for. An example of the first pattern may be EC_{50} (effective concentration in 50 % population), IC_{50} (concentration required for inhibition of 50 % population), LD_{50} (the dose required to kill half of the total population), etc. Since response values for these analyses being obtained from multiple assays at different dose or concentration levels of chemicals, these (i.e., doses required to elicit a fixed response) are preferably used as the independent variable (Y) in QSAR studies. Hence, a model can be developed from the information of varying concentrations of chemicals required to exhibit a fixed biological (or toxicological) response. One important treatment of the response variable is its logarithmic transformation allowing conversion of a wide range of response data (activity/property/toxicity) into a smaller scale. Another reason for this logarithmic data conversion is that biological/toxicological data give a parabolic curve for the dose–response relationship while the corresponding log dose–response relationship for the same data yields a sigmoidal curve that bears a linear middle portion rendering the modeling easy. It might be noted that the unit for the concentration of chemicals is expressed in molar terms, i.e., M, mM, μM, and nM and hence a chemical eliciting a fixed response at a lower concentration (C) than others actually possesses higher activity or toxicity profile. Hence, activity or toxicity profile of chemicals bears an inverse relationship with their concentration. For all practical purposes, an inverse of the concentration term is usually employed for modeling biological or toxicological data, i.e., log1/C or −logC. A data set of chemicals subjected to QSAR analysis is also expected to possess a sufficiently wide range of response data spanning at least 3–4 log units. Furthermore, all the compounds employed for a specific modeling operation are supposed to have a same mechanism of action toward the chosen response. The quantitative data for the predictor variables are obtained from experimental observations usually comprising of different physicochemical measures as well as theoretical calculations. The theoretical computation involves consideration of chemical theories that might be appended with a suitable encoding algorithm. Finally, a data matrix is prepared in which rows present different chemicals in the data set while the response variable and several independent predictor variables are presented in columns. Following the preparation of the data, the modeler needs to process it toward the goal of developing a mathematical equation or model. It may be noted that the data-processing step usually includes several pretreatment operations prior to model development such as removal of inter-correlated features and division of data set which have been discussed in later parts of this book. The data matrix comprising of response and descriptors can be subjected to linear as well as nonlinear model development in combination with a suitable feature selection algorithm. Multiple linear regression (MLR) and partial least squares (PLS) are the representative techniques for the development of linear correlation models while genetic algorithm (GA), stepwise algorithm, etc. can serve methods for variable selection (i.e., feature selection). The nonlinear modeling approaches include artificial neural network (ANN), support vector machine (SVM) and so on. As we can see that the data-processing step including model development involves handling a significant amount of data, such studies should be

associated with proper statistical tests. QSAR studies employ computation of several statistical measures and metrics to characterize the quality, stability, and validation of the models. The final operation, i.e., the interpretation of the developed model, is very crucial and it requires a thorough knowledge on the biochemical aspects of the molecules toward the response being modeled. It might be noted that QSAR modeling eventually attempts to establish a chemical basis for specific phenomena such as activity, property, or toxicity by the development of a suitable correlation equation or model. Since, all the chemicals in a data set are assumed to act via same mode of action with respect to a specific response, establishing a mechanistic foundation opens two doors: (a) prediction of the response of existing untested or new chemicals and (b) design and development of completely new chemicals possessing the desired activity/property/toxicity profile. The incorporation of a mathematical algorithm makes the QSAR technique a sound and rational tool [1, 2]. Figure 1.1 presents a simple overview of the QSAR formalism.

Encoding of the chemical features in QSAR analysis is done using a suitable mathematical algorithm. The aim was to perform a definite diagnosis of chemical structural features followed by the derivation of quantitative numbers also known as 'descriptors.' These descriptors carry explicit structural information and are used to establish a correlation with a response of interest. Hence, in a simple terminology, descriptors provide the basis for quantitative depiction of chemical structure, i.e.,

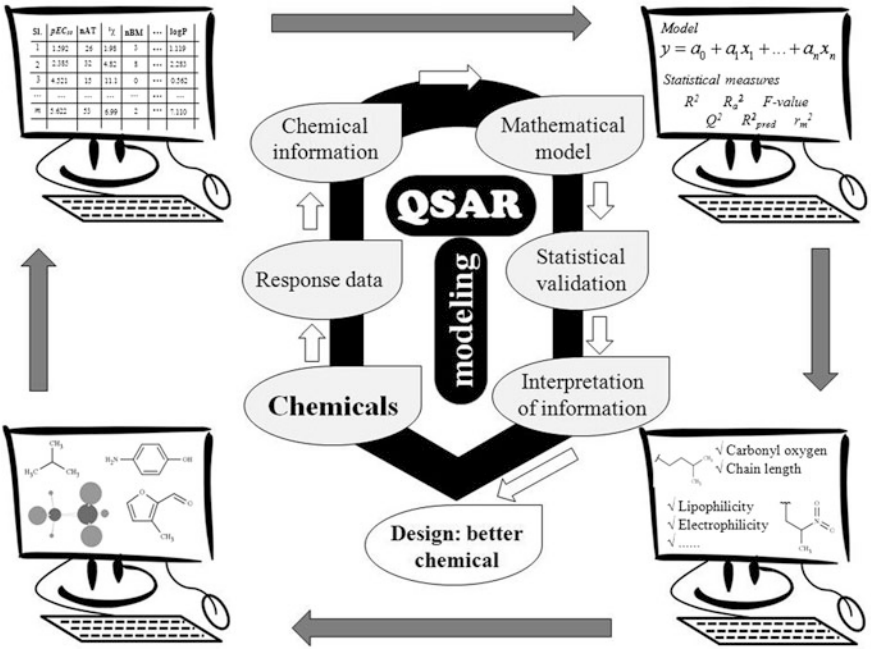

Fig. 1.1 A simple schematic overview of the formalism of QSAR

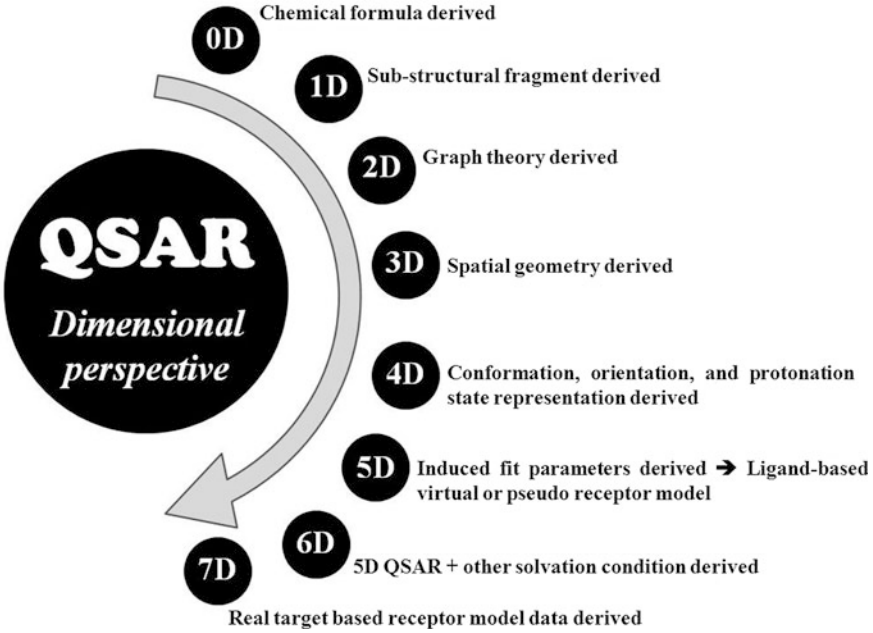

Fig. 1.2 The dimensional perspective of QSAR technique

quantitative numbers derived from a suitable mathematical operation of chemical information. Now, considering the mathematical basis involved in the quantification of chemical information, descriptors can present the dimension of the corresponding QSAR analysis. Since, the extraction of chemical information involves several hypothetical assumptions, QSAR study can be overviewed from a dimensional perspective. In Fig. 1.2, we have outlined QSAR techniques obtained using varying dimensional chemical information. However, based on the mathematical algorithm involved for developing a quantitative correlation, QSAR analysis can be conveniently classified into regression and classification types. The former type of analysis explicitly involves quantitative response values while in case of the classification analysis, one can perform classification of the data into predefined groups or classes. Figure 1.3 shows the mentioned QSAR methods with representative examples in each case.

1.2.2 Objectives of QSAR: Key Features

The principal objective of any QSAR analysis lies in the rational development of a mathematical model accompanied with the exploration of the chemical information involved therein. Such modeling always uses comparatively less amount of data of

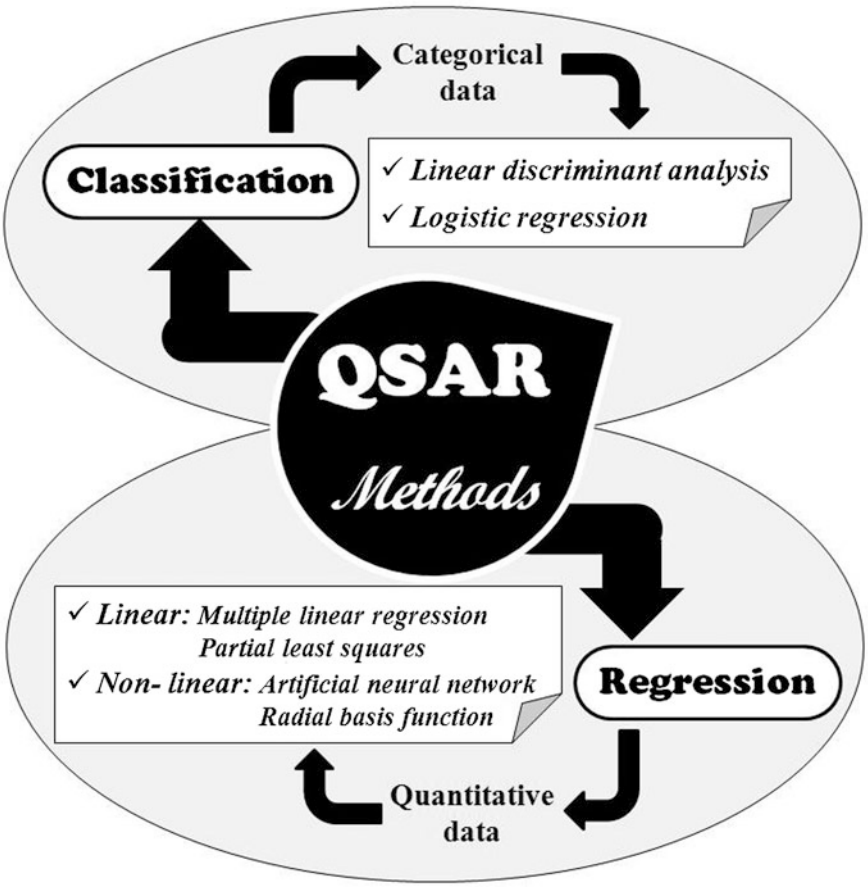

Fig. 1.3 Types of QSAR analysis based on the employed mathematical algorithm for developing correlation

chemical response and allows the prediction for a relatively large number of compounds. This provides an opportunity for this technique to be utilized in various fields. Table 1.1 briefly summarizes the potential key features of the QSAR formalism with an overview of the corresponding applications thereof.

1.2.3 Background

Chemistry serves an essential role for the interdisciplinary exploration of knowledge on the behavioral manifestation of chemicals. Different types of chemicals influence the lives of the human being covering the aspects of industrial use,

Table 1.1 An overview of the key features of the QSAR formalism

Sl. No.	The key objectives and related implications	Brief description
1	Prediction of given response: activity/property/toxicity	A mathematical model is developed with the aim of predicting response of structurally similar chemicals. Usually the prediction is performed for chemicals not included in developing the models. Such chemicals are termed as test set or external set chemicals. Usually a chemical applicability domain is developed using the modeling set (training set) and the prediction of any untested or a new chemical lying within the domain is considered reliable
2	Reduction and replacement of experimental (laboratory) animals	A QSAR study reduces animal experimentation during the preclinical stages of development of drugs since it uses limited chemical response data. The same advantage is also applied in predictive toxicology modeling. Based on the '3R' concept of Russell and Burch, namely replacement, reduction, and refinement of animal experiment in scientific studies, QSAR appears to provide a valuable alternative solution to such ethical issue. Authoritative bodies such as ECVAM, REACH regulation of European Union, office of toxic substances of US-EPA, and OECD propose the use of QSAR as studies alternative to in vivo experiment
3	Virtual screening of library data	Since QSAR leads to the development of an explicit mathematical equation, it can be employed for the screening of chemical library comprising of a large number of compounds. The information derived from descriptors can be utilized as reasonable filtering conditions toward the selection of desired compounds. Examples of some commonly used chemical library are ZINC, DUD benchmark, PubChem, ChemBank, ChEMBL, DrugBank, and Inter-bioscreen.
4	Diagnosis of mechanism	The nature of descriptive information encoded by the descriptors plays a crucial role in this perspective. Establishment of a probable mechanistic interpretation involves defined knowledge on the endpoint especially if it is biological or toxicological. The extracted chemical information is correlated with the corresponding response of interest considering the coefficient of the variables.
5	Categorization of data	A classification algorithm of QSAR allows discrimination of chemicals into groups when the response data are categorical. Such operation is primarily important in the assessment of chemical

(continued)

Table 1.1 (continued)

Sl. No.	The key objectives and related implications	Brief description
		toxicity where categorization of data into different levels of hazard such as high, low, and moderate seem useful
6	Optimization of lead molecules	One of the principle objectives of the QSAR study is the design of purpose-specific chemicals with desired response value. This principle is highly useful during the structural optimization of 'lead' molecules in a drug-designing project in order to get molecules with desired properties. QSAR studies along with other in silico methods can be suitably used toward the successful design and development of drug molecules
7	Structural refinement of synthetic target molecules	It is possible to incorporate the findings of previous QSAR observations during the structural modification process. In a study, Hansch depicted the use of prior knowledge of lipophilicity in eliminating the CNS side effect of the drug Sulmazole [Hansch C. Drug Inf J 1984;18:115–22]

laboratory and institutional applications, as well as household consumption. Hence, it has been a goal of the scientific community to study and search for the information that defines the behavior of the chemicals. The structure–activity relationship emerged as a notion for establishing a link between the chemical structures and their elicited response in a quantitative manner. It will be interesting to note that development of different chemical principles has assisted in the development of QSAR studies. Before going into the details of historical development of the QSAR paradigm, we would like to discuss a few basic dogma of chemistry of compounds. Chemicals are governed by different types of forces and energies that control their physicochemical behavior and the elicited response thereof. Attractive and repulsive forces are the resultant outcomes of intra- and inter-molecular bonding energies of chemicals under the influence of their electronic orbital interactions. All forces must be in an energetically favorable state of balance for the initiation of any kind of molecular interaction. It is to be noted that the attraction (cohesion between similar entities or adhesion between different entities) and repulsion forces might operate simultaneously during a molecular interaction. Considering the biological (and toxicological) response elicited by chemicals, different types of forces play a crucial role for instituting interaction between a chemical and a biomolecule. The physicochemical nature of compounds can be described by three principle phenomena, namely hydrophobic, steric, and electronic effects while various bonding interactions include covalent bond, hydrogen bond, ionic interaction, and dipolar interaction. All these forces and interactions function accordingly when a chemical interferes with a biological system and thereby elicits suitable response. Table 1.2 presents an overview of the mentioned forces and bonding interactions [3, 4].

The name of Mendeleev may be cited as one of the earliest scientists in the field of chemistry who used the concept of chemical correlation in 1870s by formulating the rule of eight. It is believed that the ideology of QSAR emerged in the field of toxicology and later it was supported by experiments in physical organic chemistry. With the progress of time, the concept of chemical correlation became firm in presence of strong mathematical formalisms and with advancements of chemical and physical principles. Cros observed the toxicity of primary alcohols to be correlated with their aqueous solubility in 1863 which implicates that the initial fundamental basis for QSAR modeling was emerged from the toxicological study. However, Crum-Brown and Fraser are considered to be the pioneer in the realm of QSAR modeling who represented physiological action in terms of 'chemical constitution' using a mathematical expression (Eq. 2.1) in 1868, although the phrase 'chemical constitution' was not a well-explained concept at that time.

$$\phi = f(C) \tag{1.4}$$

This observation was followed by Richardson who observed a proportional relationship of narcotic effects of primary alcohols with their molecular weight in 1869. Reynolds and Richet were the next to observe the correlational behavior of chemical nature with corresponding physiological response. Further confidence to the mathematical relationship proposed by Crum-Brown and Fraser was added by Meyer, Overton, and Baum though their work of correlating biological potency of narcotic substances with olive oil/water partition coefficient. It will be noteworthy to mention here that following extended studies, Overton depicted a proportionate mechanistic relationship between increased chain length of the studied compounds and their narcotic behavior in tadpole. Furthermore, he reported different toxicological outcomes of morphine in human and tadpoles and considered a change in the structure of protein in the studied genus. At the beginning of the twentieth century, Traube [5] observed surface tension of chemicals to be related with their narcotic potency which was later modified by Seidell [6] by depicting a similar correlation when solubility and partition coefficient measures were considered [7]. Hence, we can see that the initial observations of QSAR analysis stemmed from toxicological studies and the correlating parameters were principally physicochemical attributes. In 1933, Ferguson added a thermodynamic basis to it by proposing the narcosis behavior to be linked with the relative saturation of the substance in the applied phase. The next notable development was the exploration of the ionization of chemical species. Albert et al. [8] reported his decisive work on the ionization and shape of aminoacridine compounds in correlating their bacteriostatic potential [7]. This was followed by Bell and Roblin who also performed similar studies on sulfonamides in 1942. The study of developing quantitative descriptors for mathematical correlation models was put into light by Hammett who introduced the decisive electronic substituent constant measure Hammett sigma (σ) in relating the relative reaction rate of *meta*- and *para*-substituted benzoic acid derivatives. This study represents an essential foothold since it was the first

Table 1.2 An overview of the various forces and bonding interactions involved

Type of force/bonding interaction	Brief description
The physicochemical effects	
Hydrophobic effect	Important for eliciting activity/toxicity in biological systems. Also crucial in estimating process efficiency in case of industrial chemicals. Measurement of *n*-octanol/water partition coefficient gives a good measure of this feature in the biological system. Computationally derived measures, viz. CLOGP, MLOGP, AlogP98, etc. can be determined for the whole molecules, while it is also possible to compute hydrophobicity contributions of individual fragments
Electronic effect	This includes different types of dispersion forces, charge transfer complex formation, ionic interaction, inductive effect, hydrogen bonding, polarization effect, acid-base catalytic property, etc. and facilitates interaction with biological receptor systems
Steric effect	Such effects correlate with the spatial arrangement of molecules in the three-dimensional space. These are important in monitoring binding of chemicals to biological receptor cavity
The bonding interactions	
Covalent bond	Such bonding presents the strongest interaction (in the biological system) formed by shared electrons between each of the two participating atoms. In the context of biological receptors, formation of covalent bond between receptor and ligand (chemical) depicts irreversible binding. It has a strength of 50–150 kcal/mol
Ionic bond	This involves electrostatic attraction force between two oppositely charged ions. It bears a strength of 5–10 kcal/mol and hence much less stronger than covalent bond in the biological system. The decrease in strength of the bond is proportional to the squared distance between the participating atoms
Hydrogen bond (H-bond)	This is a weak bonding force with a strength of 2–5 kcal/mol. Such bond formation takes place between a hydrogen atom attached to a strongly electronegative atom and another atom of higher electronegativity. Atoms such as O, N, S, and F can take part in such bond formation in an intra- (same molecule) as well as an inter-molecular (different molecule) fashion. Hydrogen bonds play a very important role in exploring binding of a ligand to its biological receptor. H-bonds stabilize the structure of DNA and hence control its functional characteristics. A H-bond is highly directional at the donor atom (i.e., donating atomic H) and such bonding is controlled by the orbital spatial distribution of the acceptor site (i.e., accepting atomic H) along with a dipolar orientation of the donor group
Hydrophobic force	

(continued)

Table 1.2 (continued)

Type of force/bonding interaction	Brief description
	It measures the dislikeness of molecules toward water. Involvement of such force leads to a favorable change in entropy of a system. Addition of a chemical possessing a nonpolar group in water limits the free Brownian movement of water molecules which is overcome by the reduction of contact of water with the nonpolar interface causing aggregation of the chemical. Addition of a $–CH_2–$ group imparts a strength of 0.37 kcal/mol to a molecule due to hydrophobic force
van der Waals interaction	Such force is relatively weak with a strength varying from 0.5 to 1 kcal/mol and is nonionic in nature. Presence of electronegative atoms such as O and N in molecule cause drawl of the electron cloud toward themselves leading to the formation of a partial positive (δ+) and a partial negative (δ−) charge, i.e., dispersion of charges within the molecule. Two such participating molecules establish a weak bonding interaction. The involved forces are characterized as Keesom force, Debye force, London force, etc. depending on the nature of the established dipole interaction
Pi–pi (π–π) stacking interaction	This is a special type of non-covalent attractive force taking place between unsaturated systems such as arenes. Here, two planar molecules lead to the formation of a stacked geometric complex involving a non-covalent bonding interaction such that the solvent exposed surface area of the complex is minimized. Usually three stacked geometries are identified, viz. parallel-displaced, T-shaped edge-to-face, and eclipsed face-to-face
Charge transfer complex	This is an electron-donor and electron-acceptor complex formed between Lewis bases and Lewis acids. Charge transfer complexes are identified by specific absorption band in the UV-visible range which is different than both the donor and acceptor moieties
Orbital overlapping interaction	Overlapping of pi orbitals leads to the formation of dipole–dipole force of attraction. The π-orbital electron cloud in an aromatic system imparts a negative charge above and below the ring while keeping the equatorial hydrogen atoms positively charged and thereby allowing a dipole–dipole-type interaction between two aromatic systems due to overlap of orbital electron densities. Presence of a lone electron pair containing substituent in aromatic system can influence such bonding
Ion-dipole and ion-induced dipole interaction	Such forces have been found to influence the aqueous solubility of crystalline systems considering water as a dipolar molecule. Here, the cationic and anionic parts of a molecule, respectively, interact with the partially negatively charged oxygen and partially positively charged hydrogen atom of water molecule

established linear free energy relationship (LFER) in the QSAR modeling paradigm as shown by the following equations:

$$\log(k_X/k_H) = \rho \cdot \sigma_X \tag{1.5}$$

$$\log(K_X/K_H) = \rho \cdot \sigma_X \tag{1.6}$$

here k_H and k_X are the rate constant terms for unsubstituted and substituted benzoic acid derivatives, respectively, while K_H and K_X denote their respective equilibrium constants. σ_X represents the Hammett electronic constant of the substituent *X,* and ρ is the reaction constant term. Since ionization constant terms have been employed to depict σ, the above-mentioned equations are related to the popular free energy equation (see below) and termed as the LFER model.

$$\Delta G^0 = -RT \ln K \tag{1.7}$$

In Eq. (1.7), G denotes Gibbs free energy change, R is the ideal gas constant, and T is the ideal temperature (in Kelvin). Taft devised the first steric descriptor, i.e., the Taft steric parameter E_S in this LFER formalism defining the rates of base- and acid-catalyzed hydrolysis of aliphatic esters and provided an option for separating the effects of polar, steric, and resonance contributions. The next pioneering contribution was the development of Hansch equation in the early 1960s. Corwin Hansch is also credited the title of the 'Father of modern QSAR' who performed studies on plant growth regulators using relative hydrophobicity measure of substituent (π). The linear form of the equation was devised by Fujita and Hansch by the incorporation of Hammett constant term. A general form of the equation is presented below which has undergone several modifications in subsequent times.

$$\log 1/\mathrm{C} = k_1\pi + k_2\sigma + k_3E_s + k_0 \tag{1.8}$$

Here, k_0 represents a constant and k_1, k_2, and k_3 are the coefficient terms of the respective equation variables. Considering the 'random walk process' by drug molecules inside a biological system, Hansch formulated a parabolic relationship which was later extended by incorporating electronic and steric parameters. Both the equations are presented below:

$$\log 1/\mathrm{C} = -a(\log P)^2 + b\log P + \text{constant} \tag{1.9}$$

$$\log 1/\mathrm{C} = -a(\log P)^2 + b\log P + \rho\sigma + \delta E_s + \text{constant} \tag{1.10}$$

Free and Wilson instituted another approach of QSAR model development on a series of congeneric chemicals by using summed contribution of the parent moiety and structural fragments to represent biological activity.

$$\mathrm{BA} = \sum a_i x_i + \mu \tag{1.11}$$

Here, μ represents the contribution of the parent moiety while a_i denotes the contribution of individual structural fragments with the indicator variable x_i showing their presence ($x_i = 1$) or absence ($x_i = 0$). This equation was later modified by Fujita and Ban who implemented logarithmic activity term to keep the response variable at same level with other free energy terminologies. Considering the scope of this chapter, we will not go into an exhaustive discussion on historical avenues of QSAR modeling. The mentioned discoveries have been pioneering ones and we can see that how the idealism of correlation of 'chemical constituents' with response became a mathematical relationship through the journey involving physicochemical and thermodynamic concepts [7]. Following the development of the LFER model, the mathematical basis for QSAR was well established. Mathematical principles have much more profound impact on theoretical chemistry including QSAR analysis. In the late 1940s, studies on 'chemical graph theory' that involves concepts of mathematics and chemistry led to the development of quantitative descriptors on a purely theoretical basis. Wiener and Platt were the first to develop graph theory-based quantitative topological variables in 1947 known as Wiener index and Platt index, respectively, and reported predictive QSPR models on boiling points of hydrocarbons. This study opened a complete new possibility in the field of theoretical chemistry especially with reference to QSAR formalism that subsequently led to the developments of minimum topological difference (MTD) method of Simon, connectivity index parameters by Randić, Kier and Hall, and many more. This graph theoretic depiction of chemical structures were purely on two-dimensional basis and simultaneous studies on three-dimensional molecular geometry also led to the development of different three-dimensional attributes. Presently, several hundreds to thousands of algorithms are presented to encode molecular features and generate quantitative descriptors employing varying dimensionality, which can be used for QSAR modeling using various statistical methods. However, it will be interesting to note that the sole objective of all such methods and techniques remains the same that initially started with the journey of finding a clue to correlate response with 'chemical constitution' which was a mere composition considered at that time [1, 2, 7]. Figure 1.4 summarizes the pioneering achievements that led to the historical evolution of the QSAR formalism.

1.2.4 Importances of QSAR

Although the development of predictive QSAR/QSPR/QSTR models appears to be a relatively simple task, it has got enormous applications in serving the need of scientific fraternity. It has always been a matter of curiosity that how it is possible for different chemical agents to exert different response profile, and sometimes it is rather astonishing that even the same chemical can elicit different biological actions. Hence, the chemical features appear to be very crucial in determining behavior of chemicals. QSAR techniques can provide several advantages in terms of model predictivity and utilization of limited experimental resources, employing less

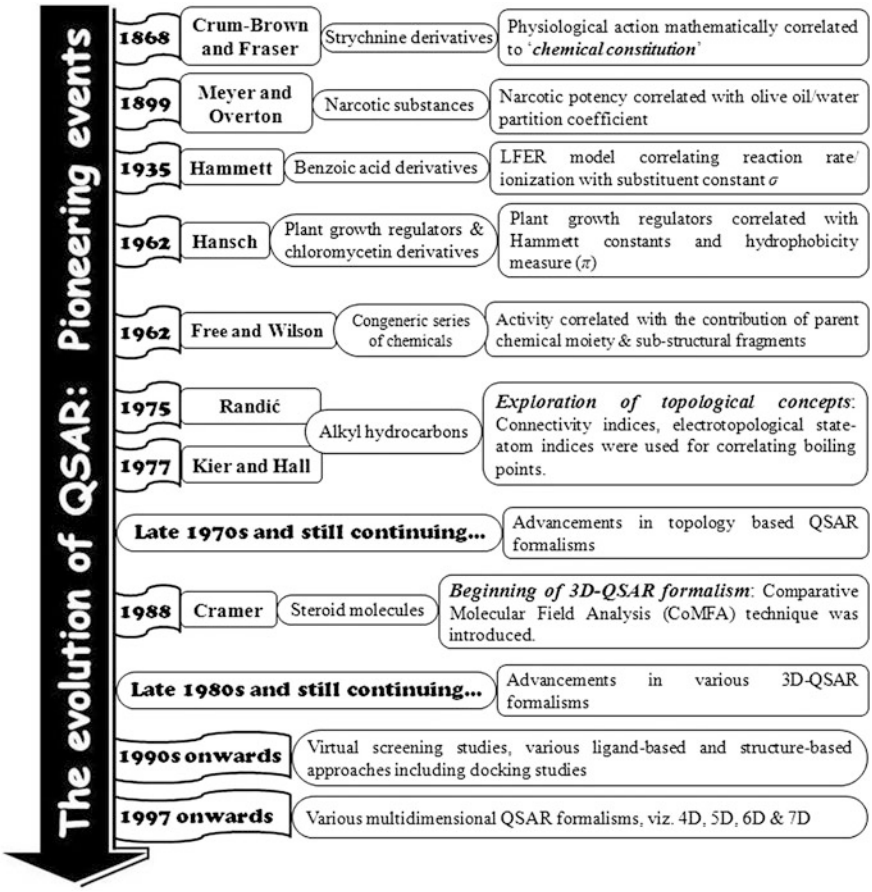

Fig. 1.4 A summary of the pioneering discoveries that led to the gradual evolution of the QSAR study

computational time. Such features encourage the use of QSAR and related techniques in costly research programs such as drug-discovery and development where it can provide valuable information by aiding rational designing strategy with minimal cost involvement. Furthermore, since the QSAR technique can allow the prediction of a chemical response of relatively large number of compounds (within the chemical domain) by using response data of limited number of chemicals, it is widely employed in predictive toxicology analysis for the assessment of chemical hazards. Figure 1.5 depicts an overview of the representative advantages provided by QSAR modeling studies. It may be noted that QSAR helps in achieving efficient, effective, safe, and environmentally benign chemicals and processes thereof and thereby facilitates a 'sustainable chemical' process [2].

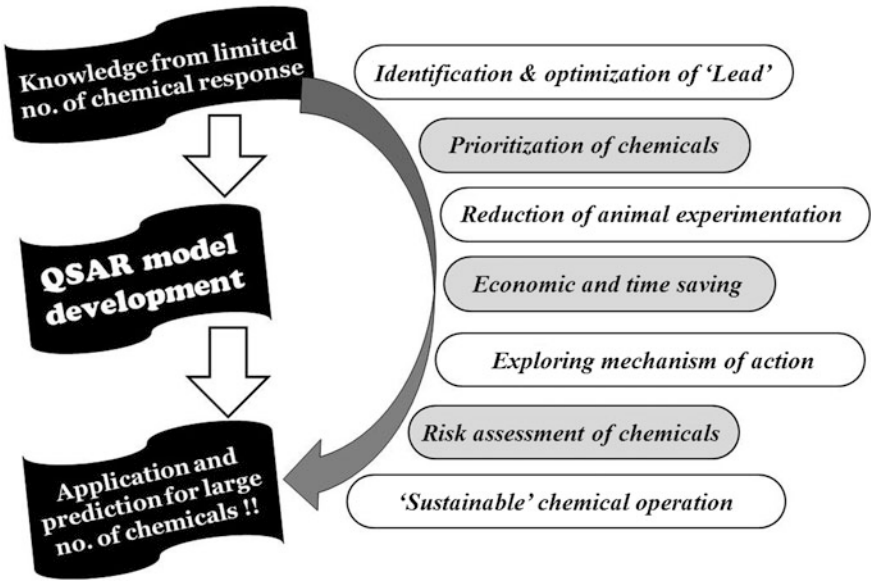

Fig. 1.5 The major advantages obtained from QSAR modeling analysis

1.2.5 QSAR and Regulatory Perspectives

The idealism of developing predictive models using the QSAR techniques is being acknowledged and prescribed by several international regulatory bodies. The following aspects are addressed by different regulatory bodies with the aim of performing risk assessment of chemicals.

1. Assessment of chemical hazard: It comprises identification as well as dose—response characterization of hazard, including classification and labeling of the chemicals.
2. Assessment of exposure.
3. Assessment of hazard and exposure.
4. Identification of persistent, bioaccumulative, and toxic (PBT) as well as very persistent and very bioaccumulative (vPvB) chemicals.

It is obvious that determination of chemical toxicity involves a sound amount of animal experiments in order to generate reliable chemical response data. Hence, it is one of the prime objectives of any hazard assessment strategy to search for suitable alternative method that will reduce animal experimentation. QSAR plays a significant role in this context since it employs comparatively less amount of response data and can predict the same for a large number of chemicals. The QSAR technique complies with the '3R' principle of Russell and Burch, namely replacement, reduction, and refinement of animals in biological experiments and aids in

regulatory assessment by performing prioritization of chemicals as well as filling of data gaps. Furthermore, modeling of categorical data (if present) becomes an important aspect here since the toxicological response of chemicals can be categorized into several groups or classes and hence designating different levels of hazards, viz. high, moderate, low, etc. The regulatory agencies which purport the use of QSAR as a valid alternative strategy to animal experiment include the European Centre for the Validation of Alternative Methods (ECVAM) of the European Union, the Office of Toxic Substances of the US Environmental Protection Agency (US-EPA), the Agency for Toxic Substances and Disease Registry (ATSDR), and the Council for International Organizations of Medical Sciences. The European Commission introduced the REACH (Registration, Evaluation, Authorization, and Restriction of Chemicals) regulations in 2006 with an aim of performing systemic evaluation of toxicological hazard of existing as well as new chemicals (imported or produced) and identified QSAR as an alternative method for toxicity testing of animals. The organization of economic cooperation and development (OECD) proposed a set of five point seminal guidelines in 2004 for the proper development and validation of predictive QSAR models by its member countries [9]. With the passage of time, QSAR studies have become an essential part of regulatory assessment on a global perspective, and various countries have developed their own 'expert systems' for determining chemical hazards. Expert systems are the computational applications providing a subject-matter expertise to non-experts by the use of definite logical reasoning. Different expert systems contain models on toxicological endpoints that are prepared and maintained by professional personnel as trusted systems with a suitable user interface such that any unknown or new chemical can easily be tested of its toxicity or categorical-hazard using the existing knowledge-base. Table 1.3 gives a representative overview of some commonly used QSAR expert systems.

1.2.6 Applications of QSAR

Chemicals represent an indispensable part of human necessity considering varying applications spanning from laboratory to industrial processes as well as household usage. QSAR presents a suitable option in the rational monitoring of activity/property/toxicity of chemicals and hence is useful in a wide variety of applications. Since fine-tuning of the behavioral nature of chemicals gives fruitful results for a significantly large class of chemicals involving pharmaceuticals, agrochemicals, perfumeries, analytical reagents, solvents, surface modifying agents, etc., the application area and possibility of the QSAR technique is enormous. In a global perspective, the chemicals modeled using the QSAR method can be overviewed in three major types, namely chemicals of health benefits (drugs, pharmaceuticals, food ingredients, etc.), chemicals involved in industrial/laboratory processes (solvents, reagents, etc.), and the chemicals posing hazardous outcome [persistent organic pollutants (POPs), toxins, xenobiotics, carcinogents, volatile organic

Table 1.3 A representative overview of some QSAR expert systems

Expert system	Web-address	Expert system	Web-address
Open source systems (free)		*Commercial systems (paid)*	
QSAR TOOLBOX (OECD)	http://www.qsartoolbox.org/	Derek Nexus	http://www.lhasalimited.org/products/derek-nexus.htm
Lazar	http://lazar.in-silico.de/predict	HazardExpert	http://www.compudrug.com/hazardexpertpro
Toxtree	http://ihcp.jrc.ec.europa.eu/our_labs/eurl-ecvam/laboratories-research/predictive_toxicology/qsar_tools/toxtree	The BfR Decision Support System (DSS)	http://www.tandfonline.com/doi/pdf/10.1080/10629360701304014
VEGA	http://www.vega-qsar.eu/	TOPKAT	http://www.sciencedirect.com/science/article/pii/0027510794901252
DEMETRA	http://www.demetra-tox.net/	MCASE and CASE Ultra	http://www.multicase.com/
EPI Suite™	http://www.epa.gov/opptintr/exposure/pubs/episuite.htm	Leadscope	http://www.leadscope.com/
TEST	http://www.epa.gov/nrmrl/std/qsar/qsar.html	TerraQSAR™	http://www.terrabase-inc.com/
OncoLogic™	http://www.epa.gov/oppt/sf/pubs/oncologic.htm	ACD/Percepta	http://www.acdlabs.com/products/percepta/physchem_adme_tox/
		MolCode Toolbox	http://www.molcode.com/
		TIMES	http://oasis-lmc.org/products/software/times.aspx

compounds (VOCs), etc.]. In Fig. 1.6, we have attempted to divide the employment of QSAR application in three broad areas, namely drug designing, material science, and predictive toxicology. Some potential areas of material science which can be addressed by employing predictive QSAR modeling have been depicted in Fig. 1.7, while Fig. 1.8 shows some representative endpoints addressed by the QSAR technique in the sphere of assessing predictive toxicology. It might be interesting to note that apart from modeling biological activity and toxicity endpoints, the drug-designing paradigm involves modeling of ADME which aims to monitor the pharmacokinetic profile of drug candidates prior to its synthesis and thereby enhancing the efficacy of the designed compounds inside biological system.

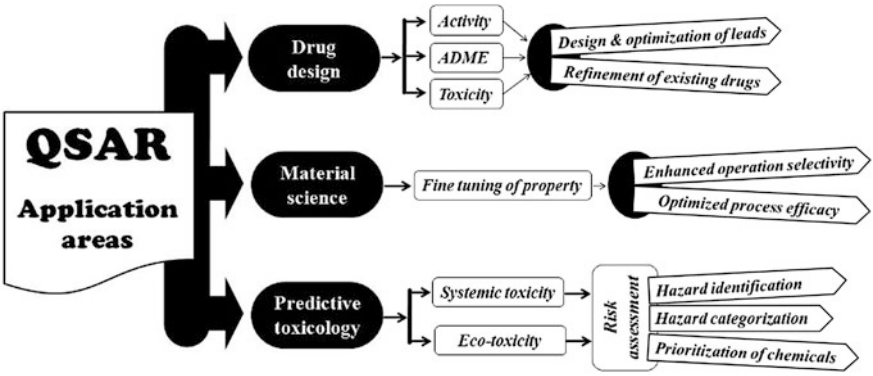

Fig. 1.6 The broad application areas addressed by QSAR modeling studies

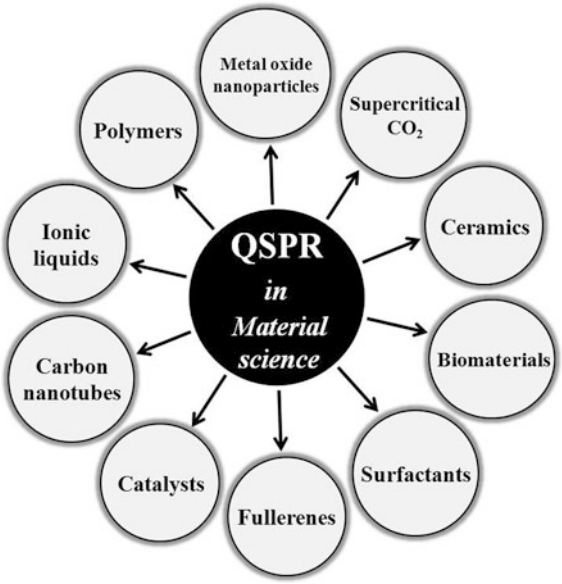

Fig. 1.7 Some representative endpoints addressed by QSAR analysis in the field of material science

Assessment of toxicity of chemicals principally involves two options, namely assessment of systemic toxicity as well as the monitoring of ecotoxicological hazard. Drugs and pharmaceuticals are capable of posing toxicity to the specific organ system, e.g., hepatotoxicity, cardiovascular toxicity, and nephrotoxicity, while they can also be of serious concern in an environmental perspective since wastewater stream containing even trace amount of such compounds can lead to damage in the ecosystem. Physiologically based pharmacokinetic (PBPK) modeling is another potential area that involves modeling of chemicals such as VOCs using

Fig. 1.8 Some representative endpoints addressed by QSAR study in the realm of predictive toxicology analysis

physicochemical (log$P$) as well as biochemical parameters (Michaelis constant K_m, maximal velocity V_{max}, hepatic clearance, etc.).

Hence, we can see that the simple ideology of QSAR, i.e., development of a suitable mathematical correlation between the chemical attributes and a response of interest, can be of significant application to serve the human community. Considering the rising health hazard issues and other environmental damage, modern technologies are aimed toward the establishment of a 'sustainable' and 'green' ecosystem that deals with chemical processes that ensure environmental benevolence in terms of efficiency, effectiveness, and safety concerns. QSAR plays an encouraging role in achieving this environmental greenness through the design and development of process-specific chemicals with reduced (or no) hazardous outcomes.

Drug design and development remain the utmost important area addressed by the QSAR formalism. The challenge faced in this perspective is quite higher since the development of a drug molecule is a time consuming as well as costly procedure. Furthermore, the rate of success is also very low since the chance of rejection is very high at any stage of the development paradigm. Figure 1.9 presents an overview of the steps involved in the development of a drug molecule starting from its initial developmental stage. QSAR study can speed up this discovery process by providing rational information on the chemistry of the investigational molecules covering the issues of its contribution to pharmacological behavior, ADME property as well as the toxic outcomes. QSAR can provide valuable information at the stages of design and development and preclinical study, thereby facilitating the outcomes of clinical research and the subsequent approval process. It may be noted that since biomolecular activity involves complex interaction involving

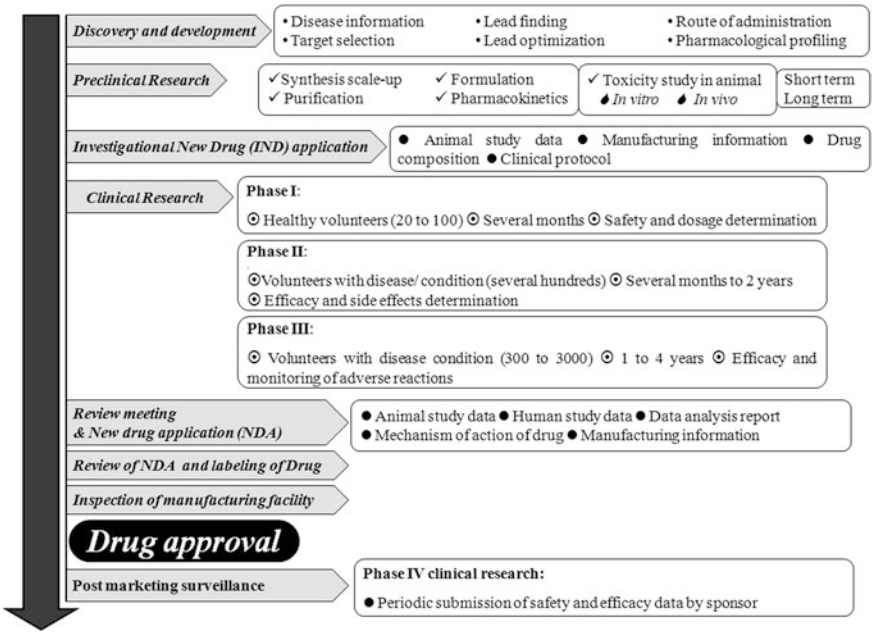

Fig. 1.9 Different phases involved in the development of a drug

Fig. 1.10 Interplay of different in silico techniques with predictive QSAR modeling study

ligand-receptor attributes inside the living system, the development of potential lead molecules would certainly utilize some other techniques as well, namely molecular docking, pharmacophore modeling cheminformatics, and virtual screening along with the QSAR technique. Such techniques are useful in establishing a suitable biochemical correlation for the discovery of drug candidates and can also be applied to other fields as well like toxicophore analysis. Figure 1.10 shows the interplay among various in silico techniques including the QSAR algorithm successfully deployed toward the design of target molecules.

Application of the QSAR technique in combination with other in silico methods has been very fruitful in the drug-discovery paradigm, and some representative examples of such designed drug molecules which were later approved by the US Food and Drug Administration (US-FDA) as drug entities are presented in Fig. 1.11.

Fig. 1.11 Representative examples of drugs designed and developed using different in silico techniques including QSAR modeling analysis. Under individual drugs, the information shown includes the disease indication, the year of US-FDA approval, the proprietary name, and the manufacturing company, respectively

1.3 What Are Descriptors?

1.3.1 Definition

A QSAR model can be expressed as a simple mathematical equation which can correlate the properties (physicochemical/biological/toxicological) of molecules employing diverse computationally or experimentally derived quantitative parameters termed as '*descriptors.*' The descriptors are correlated with the experimental properties (response) using a variety of chemometric tools in order to obtain a statistically significant QSAR model. Molecular descriptors are the '*terms that characterize specific information of 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.*' The developed equation should provide a significant insight into the essential structural requisites of the molecules which contribute to the biological response of the studied molecules [10]. In other words, the response of a chemical can be mathematically presented as the function of descriptors (Eq. 1.12).

$$\begin{aligned}&\text{Response (activity/property/toxicity)}\\&= f(\text{Molecular information extracted using chemical structure or physicochemical property})\\&= f(\text{Descriptors})\end{aligned} \tag{1.12}$$

An ideal descriptor should possess the following features for the construction of a reliable QSAR model:

1. A descriptor should be relevant to a broad class of compounds.
2. A descriptor must be correlated with the studied biological responses while illustrating insignificant correlation with other descriptors.
3. Calculation of the descriptor should be fast and independent of experimental properties.
4. A descriptor should produce different values for structurally dissimilar molecules, even if the structural differences are little.
5. A descriptor should possess physical interpretability to determine the query features for the studied compounds.

A schematic illustration is presented in Fig. 1.12 to depict the steps how a chemical structure is employed to compute descriptors and then utilized in QSAR model development.

1.3.2 Types of Descriptors

Descriptors can be of different types depending on the method of their computation or determination: physicochemical (hydrophobic, steric, or electronic), structural

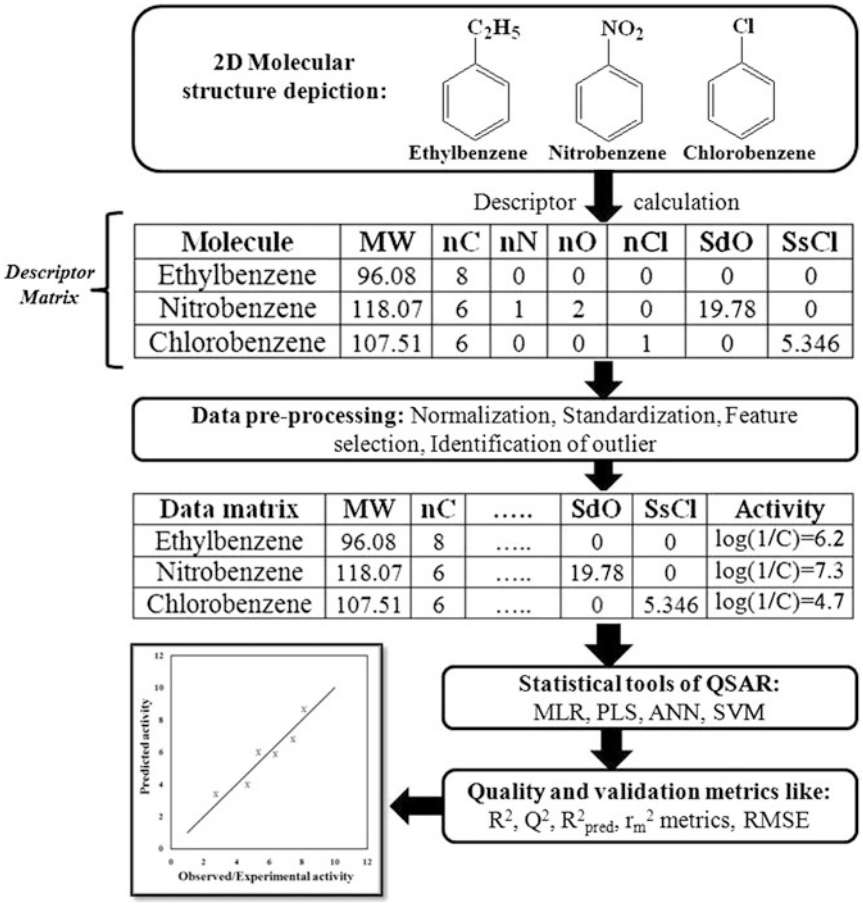

Fig. 1.12 Schematic illustration to show how chemical structures are employed to compute descriptors and QSAR model development

(frequency of occurrence of a substructure), topological, electronic (molecular orbital calculations), geometric (molecular surface area calculation), or simple indicator parameters (dummy variables). In a broader perspective, descriptors (specifically, physicochemical descriptors) can be classified into two major groups: (1) substituent constants and (2) whole molecular descriptors [11, 12]. *Substituent constants* are basically physicochemical descriptors which are designed on the basis of factors, which govern the physicochemical properties of chemical entities. *Whole molecular descriptors* are expansions of the substituent constant approach, but many of them are also derived from experimental approaches.

Descriptors may also be classified based on dimensions. Table 1.4 gives a useful illustration of commonly used molecular descriptors based on dimensions. It is

Table 1.4 Different descriptors employed in the QSAR study based on dimension

Dimension of descriptors	Parameters
0D-descriptors	Constitutional indices, molecular property, atom and bond count
1D-descriptors	Fragment counts, fingerprints
2D-descriptors	Topological parameters, structural parameters, physicochemical parameters including thermodynamic descriptors
3D-descriptors	Electronic parameters, spatial parameters, molecular shape analysis parameters, molecular field analysis parameters and receptor surface analysis parameters

interesting to point out that we have confined our discussion here from 0D- to 3D-descriptors only, though higher dimensional descriptors are also available.

1.3.2.1 2D-Descriptors

Topological

Topological descriptors are calculated based on the graphical representation of molecules and thus they neither require estimation of any physicochemical properties nor need the rigorous calculations involved in the derivation of the quantum chemical descriptors. The structure representation of the molecule depends on its 2D-graphical topology indicating the position of the individual atoms and the bonded connections among them. The formulation of these descriptors is based upon the characterization of chemical structure by graph theory. The graph theoretic determination of the molecular structure involves vertices symbolizing atoms and the covalent bonds representing the edges [13]. In Table 1.5, we have presented the most commonly used topological descriptors along with their formal mathematical definitions briefly, due to their widespread use in QSAR model development.

Structural Parameters

Detailed list of structural descriptors [11] is given in Table 1.6.

Physicochemical Parameters

Physicochemical parameters are designed on the basis of factors, which govern the physical and chemical properties of chemical entities. Due to change in physicochemical properties, absorption, distribution, transport, metabolism, and elimination, behavior of bioactive chemical entities may be changed. The important physicochemical factors affecting bioactivity of drugs and chemical include hydrophobicity, electronic, and steric character of the whole molecules and also the

Table 1.5 A representative overview of topological descriptors used in QSAR model development

Descriptors type	Mathematical definition
Balaban J index	$J = \frac{M}{\mu+1}\sum_{\text{all edges}} \left(\delta_i\delta_j\right)^{-0.5}$ where M is the number of edges, μ represents cyclomatic number and δ_i (or δ_j) can be defined as: $\delta_i = \sum_{j=1}\delta_{ij}$
Bond/edge connectivity indices	$\in = \sum_{l=1}^{P_2}\left[\delta(e_i)\delta\left(e_j\right)\right]_l^{-0.5}$ where $\delta(e)$ corresponds to edge degree and is summed (l) over all the p_2 adjacent edges.
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: $I_i = \frac{\left[(2/N)^2\delta^v+1\right]}{\delta}$ and $\Delta I_i = \sum_{j\neq 1}\frac{\left(I_i - I_j\right)}{r_{ij}^2}$ where N is the principal quantum number and r_{ij} being the topological distance between atoms i and j
Extended bond/edge connectivity indices	${}^m\in_t = \sum_s\prod_i\left[\delta(e_i)\right]_s^{-0.5}$ 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
Extended topochemical atom (ETA) indices	Some basic ETA indices definitions are given below $\alpha = \frac{Z-Z^v}{Z^v}\cdot\frac{1}{\mathrm{PN}-1}$, $\beta = \Sigma x\sigma + \Sigma y\pi + \delta$, $\gamma_i = \frac{\alpha_i}{\beta_i}$, $[\eta]_i = \sum_{j\neq i}\left[\frac{\gamma_i\gamma_j}{r_{ij}^2}\right]^{0.5}$, $\varepsilon = -\alpha + 0.3\times Z^v$, $\psi = \frac{\alpha}{\varepsilon}$ 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
Kappa shape indices	${}^1\kappa = 2\frac{{}^1P_{\max}{}^1P_{\min}}{\left({}^1P_i\right)^2}$; ${}^2\kappa = 2\frac{{}^2P_{\max}^2P_{\min}}{\left({}^2P_i\right)^2}$; ${}^3\kappa = 2\frac{{}^3P_{\max}^3P_{\min}}{\left({}^3P_i\right)^2}$ where, the numbers of one, two, and three path lengths are denoted by ${}^1P_i$, ${}^2P_i$ and ${}^3P_i$, 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: ${}^1P_{\max} = (A(A-1))/2$; ${}^1P_{\min} = (A-1)$ ${}^1\kappa = \frac{A(A-1)^2}{\left({}^1P_i\right)^2}$; ${}^2\kappa = \frac{(A-1)(A-2)^2}{\left({}^2P_i\right)^2}$; ${}^3\kappa = \frac{(A-1)(A-3)^2}{\left({}^3P_i\right)^2}$ for odd value of A and, ${}^3\kappa = \frac{(A-2)^2(A-3)}{\left({}^3P_i\right)^2}$ for even value of A
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_{Csp}^3 are the covalent radii of the atom x and sp^3 hybridized carbon atom, respectively. The corresponding alpha-modified *kappa* shape indices are defined

(continued)

Table 1.5 (continued)

Descriptors type	Mathematical definition
	below: ${}^{1}\kappa_{\alpha} = \frac{(A+\alpha)(A+\alpha-1)^2}{({}^{1}P_i+\alpha)^2}$; ${}^{2}\kappa_{\alpha} = \frac{(A+\alpha-1)(A+\alpha-2)^2}{({}^{2}P_i+\alpha)^2}$; ${}^{3}\kappa_{\alpha} = \frac{(A+\alpha-1)(A+\alpha-3)^2}{({}^{3}P_i+\alpha)^2}$ for odd A values and ${}^{3}\kappa_{\alpha} = \frac{(A+\alpha-2)^2(A+\alpha-3)}{({}^{3}P_i+\alpha)^2}$ for even A values
Molecular connectivity index	${}^{m}\chi_t = \sum_{j=1}^{n_m} {}^{m}S_j$ where, n_m represents the number of t type subgraphs of order m. The term ${}^{m}S_j$ may be defined as follows: ${}^{m}S_j = \prod_{i=1}^{m+1} (\delta_i)_j^{-0.5}$ and δ_i for the ith atoms may be defined as: $\delta_i = \sigma_i - h_j$, 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
Randic branching index (χ)	$\chi = \sum_{\text{all edges}} \left(\delta_i\delta_j\right)^{-0.5}$ where, δ_i and δ_j represent the number of other non-hydrogen atoms bonded to atoms (vertices) i and j, respectively, forming an edge ij
Subgraph count index	It is the number of sub-graphs 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 sub-graphs are divided into three types on the basis of path, cluster, and ring (SC_3_P, SC_3_C, SC_3_CH)
Valence molecular connectivity index	${}^{m}\chi_t^{v} = \sum_{j=1}^{n_m} {}^{m}S_j^{v}$ Here, the corresponding term δ^{v} is defined as: $\delta_i^{v} = \frac{\left(Z_i^{v}-h\right)}{\left(Z-Z_i^{v}-1\right)}$, where Z and Z^{v} are the atomic number and the total number of valence electron, respectively, for the ith vertex
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
Zagreb group indices	$\text{Zagreb} = \sum_i \delta_i^2$ where δ_i is the valency of vertex atom i

Table 1.6 Structural parameters used in the development of QSAR models

Parameters	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 which are considered to be meaningful for molecular mechanics. All terminal H atoms are ignored
H-bond donor	It counts the number of groups or moieties capable of donating hydrogen bonds
H-bond acceptor	This descriptor calculates the number of hydrogen-bond acceptors present in the molecule

substituents present in the molecules [14–16]. Some formal definitions of physicochemical descriptors commonly used as predictor variables in QSAR analysis are shown in Table 1.7.

Indicator Variables

Indicator variables have been employed in QSAR models due to their simplicity. Substructure-based descriptors can be easily employed as indicator variables. Two sets of compounds which differ from each other only by a substructure existing in one set but not the other can be studied as an entire set when using an indicator variable. The major limitation of this variable is that this approach should only be employed when the two sets of compounds are identical in every respect, except for the substructure being coded with the indicator variable.

Thermodynamic Descriptors

The most commonly used thermodynamic descriptors [11] in QSAR models are described in Table 1.8.

1.3.2.2 3D-Descriptors

Electronic Parameters

Electronic descriptors are defined in terms of atomic charges and are used to describe electronic aspects both of the whole molecule and of particular regions, such as atoms, bonds, and molecular fragments. Electrical 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 [11]. The electronic descriptors used in the present studies are summarized in Table 1.9.

Spatial Parameters

Spatial parameters comprise a series of descriptors calculated based on the spatial arrangement of the molecules and the surface occupied by them. The list of spatial descriptors [11] is summarized in Table 1.10.

Table 1.7 Formal definitions of most commonly used physicochemical descriptors in QSAR analysis

Parameter	Definitions
Parameters defining hydrophobic nature	
Partition coefficient	$\log P = \log K_{o/w} = \log \frac{[C]_{n-\text{octanol}}}{[C]_{\text{water}}}$ where C is the concentration of a solute in the respective mentioned phase (water or *n*-octanol). Usually, compounds having log P value more or less than 1 are considered to be hydrophobic and hydrophilic, respectively.
Hydrophobicity constant (π)	$\pi_X = \log P_X - \log P_H$ where P_X and P_H are the partition coefficient values of the compound with and without specific substituent, respectively. Positive value of π of a given substituent imparts lipophilic character to a molecule and vice versa
Parameters defining electronic nature	
Hammett substituent constant (σ)	$\sigma_X = \log(K_X/K_H)$ where X is a substituent, and K_X and K_H are the equilibrium or dissociation constant with and without the substituent, respectively. Two parameters, namely σ_m and σ_p are widely used representing the respective values for meta and para substituents in an aromatic system
Acid dissociation constant	Acid dissociation constant can be explained by following equation: $K_a = \frac{[A^-][H^+]}{[HA]}$ where A^- is the conjugate base of acid HA and H^+ is the proton. The negative logarithmic function (pK_a) is used for the modeling purpose and can be defined as: $pK_a = -\log_{10} K_a$. It is usually determined using the famous Henderson Hasselbalch equation: $pK_a = pH - \log \frac{[A^-]}{[HA]}$ where, pH is the negative logarithmic concentration of H^+ ion, i.e., $pH = -\log[H^+]$
Parameters defining steric nature	
Taft's steric factor (E_s)	$E_s = \log k_X - \log k_0$ where k_0 and k_X are the rate constants of hydrolysis of an organic compound without having and having substituent X, respectively. The parameter E_s gives a measure of intramolecular steric effect of substituents
Charton's steric parameter (v) 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 the Charton's steric parameter (υ_X). Taft also pointed out that E_s varies parallel to the atom group radius. The Charton's steric parameter can be defined as: $\upsilon_X = r_X - r_H = r_X - 1.20$ where, r_X and r_H are the minimum van der Waals radii of the substituent and hydrogen, respectively

(continued)

Table 1.7 (continued)

Parameter	Definitions
Molar refractivity	$\mathrm{MR} = \left(\frac{n^2-1}{n^2+2}\right) \times \frac{\mathrm{MW}}{\rho}$ where n represents refractive index, molecular weight is denoted by MW, and ρ is the density of the molecule. Molar refractivity provides a measure of volume occupied by an atom or a group
Verloop STERIMOL parameters	Verloop and coworkers developed STERIMOL parameters, which are a set of five descriptors (L, B1, B2, B3, and B4) in order 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 an angle of 90^{o} with each other. The large number of parameters required to define each substituent, and the large number of compounds necessary to incorporate all the parameters into 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
Parachor	An important whole molecular parameter defining the steric nature is parachor which can be explained by following equation $\mathrm{PA} = \gamma^{1/4} \cdot \frac{\mathrm{MW}}{\rho_L - \rho_V}$ where γ is the surface tension of the liquid, MW is the molecular weight, and ρ_L and ρ_V are the respective densities of the liquid and vapor state. Parachor depends on molecule volume

Table 1.8 Thermodynamic parameters used in the development of QSAR models

Descriptor	Description
AlogP	Log of the partition coefficient using Ghose & Crippen's method
AlogP98	The AlogP98 descriptor is an implementation of the atom-type-based AlogP method
Alogp_atypes	The 120 atom types defined in the calculation of AlogP98 are available as descriptors. Each AlogP98 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

Table 1.9 Electronic descriptors employed in the construction of QSAR models

Parameters	Explanations
Sum of atomic polarizabilities	It is the summation of atomic polarizabilities (A_i). The polarizabilities are calculated as follows: $P_a = \sum_i A_i$ The coefficient, A, is used for calculation of molecular mechanics
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, Z) of the dipole moment are calculated. It is determined by using partial atomic charges and atomic coordinates
Highest occupied molecular orbital (HOMO) energy	It is the highest energy level in the molecule that contains electrons. 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	It is the lowest energy level in the molecule that contains no electrons. 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)	It 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

Molecular Shape Analysis (MSA) Descriptors

The MSA descriptors are used to determine the molecular shape commonality [11]. Most commonly used MSA descriptors are following: difference volume (DIFFV), common overlap steric volume (COSV), common overlap volume ratio (Fo), non-common overlap steric volume (NCOSV), and root mean square to shape reference (ShapeRMS). A detailed explanation of these MSA descriptors is provided in Chap. 3.

Table 1.10 Spatial parameters used in the development of QSAR models

Parameters	Explanation
Radius of gyration (RadOfGyration)	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. This can be calculated as follows: $\text{RadofGyration} = \sqrt{\left(\sum \frac{\left(x_i^2+y_i^2+z_i^2\right)}{N}\right)}$ here, N is the number of atoms and x, y, z are the atomic coordinates relative to the center of mass
Jurs descriptors	The 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
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, i.e., *XY*, *YZ,* and *XZ*. Descriptors depend not only on conformation but also on the orientation of molecule. Molecules are rotated to align principal moments of inertia with *X*, *Y,* and *Z* axes
Molecular surface area (area)	It 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
Density	This 3D-descriptor 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
Principal moment of inertia (PMI)	The moments of inertia are computed for a series of straight lines through the center of mass. These are associated with the principal axes of the ellipsoid. If all three moments are equal, the molecule is considered to be a symmetrical top
Molecular volume (Vm)	This 3D-descriptor is the volume inside the contact surface. It is related to binding and transport

Molecular Field Analysis (MFA) Parameters

The MFA formalism computes probe interaction energies on a rectangular grid around a collection 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^+ or CH_3) and a molecular model at a series of points defined by a rectangular grid. Fields of molecules are represented using grids in MFA and each energy associated with an MFA grid point can serve as input for the calculation of a QSAR [17].

Table 1.11 List of software tools and online platforms for computation of molecular descriptors

Software/online platform	Weblink
Cerius^2	http://accelrys.com/
CODESSA PRO	http://www.codessa-pro.com/index.htm
Discovery studio	http://accelrys.com/
DRAGON	http://www.talete.mi.it/products/dragon_description.htm
E-Dragon at VCCLAB	http://www.vcclab.org/lab/edragon/
GRID	http://www.moldiscovery.com/soft_grid.php
JME Molecular Editor	http://www.molinspiration.com/jme/index.html
Linux4Chemistry	http://www.linux4chemistry.info/
MOE	http://www.chemcomp.com/software.htm
MOLCONN-Z	http://www.edusoft-lc.com/molconn/
MOLE db	http://michem.disat.unimib.it/mole_db/
MOLGEN-QSPR	http://www.molgen.de/?src=documents/molgenqspr.html
OCHEM	https://ochem.eu/home/show.do
OpenBabel	http://openbabel.org/
PaDEL-Descriptor	http://padel.nus.edu.sg/software/padeldescriptor/
PCLIENT	http://www.vcclab.org/lab/pclient/
QSARModel	http://www.molcode.com/
QuaSAR	http://www.chemcomp.com/feature/qsar.htm
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

Receptor Surface Analysis (RSA) Parameters

The energies of interaction between the receptor surface model and each molecular model can be used as descriptors for generating QSARs [17]. The surface points that organize as triangle meshes in the construction of the RSA store these properties as associated scalar values. Receptor surface models provide compact, quantitative descriptors, which capture three-dimensional information of interaction energies in terms of steric and electrostatic fields at each surface point. A detailed explanation of these RSA descriptors is provided in Chap. 3.

1.3.3 Software Tools and Online Platforms

QSAR is gaining popularity among the researchers with the development of new and advanced software tools and online platforms which allow them to determine the molecular structural features responsible for compounds activity/property/toxicity. Table 1.11 shows a representative list of most commonly employed software tools and online platforms for the generation of descriptors from molecular structures.

1.4 Conclusion

Development of techniques to fine-tune and modify the chemistry of compounds provides an enormous opportunity toward the development of purpose-specific chemicals. The search for the answer to the query how different chemicals elicit different responses and even the same chemical shows varying behavioral features has led to the exploration of different chemical attributes. Predictive QSAR modeling technique provides an option for developing a mathematical basis for the elicited chemical responses. Since the generated basis is highly rational on the ground of chemical information, such techniques are widely employed to maneuver the needs of the industry as well as academia. The drug-discovery paradigm involving costly and time-consuming steps can be easily rationalized and put under suitable basis using QSAR and other suitable in silico techniques in the preclinical research programmes. The QSAR technique also enables optimization of chemical operations by enhancing the selectivity of various process chemicals. Furthermore, the QSAR technique has profound applications in the risk assessment paradigm considering the minimal engagement of ethical issues related to animal experiment while using regression or classification-based predictive mathematical models. Among other components, descriptors present one of the crucial elements of the QSAR formalism. The ultimate diagnosis of chemical features is preserved in the form of quantitative numbers as descriptors which enables the identification of mechanism of action of a given biochemical process and any modification thereof. It should be noted that ‘no’ single descriptor can provide any universal solution to chemical problems. Sometimes, the nature of endpoints becomes the determining parameter in choosing suitable descriptors. Although a wide variety of descriptors are available for use, the goal of a modeler should be toward the use or development of descriptors which are easily computable giving an explicit amount of chemical information. Hence, one of the major goals of the modeler should not only be directed toward the development of a good mathematical correlation between response and descriptors, but it should also provide a suitable explanation of the result, i.e., a mechanistic overview such that the QSAR formalism can be used as a rational chemical designing tool instead a ‘black-box’ method of deriving a mathematical correlation involving a series of abstract mathematical algorithms.

References

1. Todeschini R, Consonni V, Gramatica P (2009) Chemometrics in QSAR. In: Brown S, Tauler R, Walczak R (eds) Comprehensive chemometrics, vol 4. Elsevier, Oxford, pp 129–172
2. Tute MS (1990) History and objectives of quantitative drug design. In: Hansch C, Sammes PG, Taylor JB (eds) Comprehensive medicinal chemistry, vol 4. Pergamon Press, Oxford, pp 1–31
3. Sinko PJ (ed) (2011) Martin's physical pharmacy and pharmaceutical sciences, 6th edn. Lippincott Williams & Wilkins, Baltimore
4. Daniels TC, Jorgensen EC (1982) Physicochemical properties in relation to biological action. In: Doerge RF (ed) Wilson and Gisvold's textbook of organic medicinal and pharmaceutical chemistry, 8th edn. J.B. Lippincott Co., Pennsylvania
5. Traube J (1904) Theorie der Osmose and Narkose. Pflüg Arch Physiol 105:541–558
6. Seidell A (1912) A new bromine method for the determination of thymol, salicylates, and similar compounds. Am Chem J 47:508−526
7. Selassie CD (2003) History of quantitative structure-activity relationships. In: Abraham DJ (ed) Burger's medicinal chemistry and drug discovery, vol 1., Drug DiscoveryWiley, New York, pp 1–48
8. Albert A, Rubbo SD, Goldacre R (1941) Correlation of basicity and antiseptic action in an acridine series. Nature 147:332−333
9. Fjodorova N, Novich M, Vrachko M, Smirnov V, Kharchevnikova N, Zholdakova Z, Novikov S, Skvortsova N, Filimonov D, Poroikov V, Benfenati E (2008) Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. J Environ Sci Health Part C Environ Carcinog Ecotoxicol Rev 26:201–236
10. Guha R, Willighagen E (2012) A survey of quantitative descriptions of molecular structure. Curr Top Med Chem 12:1946–1956
11. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
12. Livingstone DJ (2000) The characterization of chemical structures using molecular properties. A survey. J Chem Inf Comput Sci 40:195–209
13. Roy K, Das RN (2014) A review on principles, theory and practices of 2D-QSAR. Current Drug Metabol 15:346–379
14. Taylor PJ (1991) Quantitative drug design. the rational design, mechanistic study and therapeutic applications of chemical compounds. In: Hansch C, Sammes PG, Taylor JB (eds) Comprehensive medicinal chemistry, vol 4. Pergamon Press, Oxford; pp 241–294
15. Rekker R (1977) The hydrophobic fragmental constant. Elsevier, Amsterdam
16. Hansch C, Leo A, Hoekman D (1995) Exploring QSAR vol 2: hydrophobic, electronic and steric constants. ACS, Washington DC
17. Hopfinger AJ, Tokarsi JS (1997) In: Charifson PS (ed) Practical applications of computer-aided drug design. Marcel Dekker, New York, pp 105–164