Chapter-5

Quantitative Structure Activity Relationship (QSAR) *Study*

Chapter 5.

Quantitative Structure Activity Relationship (QSAR) Study

5.1. Introduction

Over the past three decades, the center of gravity (the intellectual focus) of medicinal chemistry has shifted dramatically from, how to make a molecule, to what molecule to make The challenge now is gathering of information to make decisions regarding the use of resources in drug design The bio-response of a compound can be the result of several types of interactions between the bioactive compound and the receptor, such as hydrophobic and electrostatic forces, hydrogen bonding and electron donor/acceptor complex These interactions are closely related to physicochemical and structural properties of its component molecules Therefore it is possible to predict or explain the biological behavior of molecules from their physicochemical properties

Drug design is an iterative process which begins with a compound that displays an interesting biological profile and ends with optimizing both the activity profile for the molecule and its chemical synthesis The process is initiated when the chemist conceives a hypothesis which relates the chemical features of the molecule (or series of molecules) to the biological activity Without a detailed understanding of the biochemical processes responsible for activity, the hypothesis generally is refined by examining structural similarities and differences for active and inactive molecules Compounds are selected for synthesis which maximize the presence of functional groups or features believed to be responsible for activity The combinatorial possibilities of this strategy for even simple systems can be explosive The alternative to this labor intensive approach to compound optimization is to develop a theory that quantitatively relates variations in biological activity to changes in molecular descriptors which can easily be obtained for each compound. The basic assumption underlying this field of research called quantitative structure-activity relationship (QSAR) is that the structure of a molecule determines its performance

QSAR represent an attempt to correlate structural or property descriptors of compounds with the biological activities. This paradigm can be expressed by P = f(S), where P is any physical, agrochemical, biomedical, toxicological or environmental activity of interest and S may represent either an empirical property of the total molecular structure, a relevant substructure fragment or a theoretical structural descriptor (or a set of descriptors) quantifying some aspects of molecular

structure A QSAR can then be utilized to help guide chemical synthesis Activities used in QSAR include chemical measurements and biological assays The physicochemical descriptors, which include parameters to account for hydrophobicity, topology, electronic properties, and steric effects, are determined empirically or, more recently, by computational methods.

5.2. Historical development of QSAR

The first approach to develop quantitative relationships which described activity as a function of chemical structure relied on the principles of thermodynamics The free-energy terms ΔE, ΔH and ΔS were represented by a series of parameters which could be derived for a given molecule. Electronic effects such as electron donating and withdrawing tendencies, partial atomic charges and electrostatic field densities were defined by Hammett sigma (σ) values, resonance parameters (R values), inductive parameters (F values) and Taft substituent values (ρ^*, σ^*, E_s). Steric effects such as molecular volume and surface area were represented by values calculated for Molar Refractivity and the Taft steric parameter Enthalpic effects were calculated using partition coefficients (log P) or the hydrophobic parameter (π) which was derived from the partition coefficient In addition, an assortment of structural indices was used to describe the presence of specific functional groups at positions within the molecule.

QSAR studies date back to the 19th century In 1863, A F A Cros at the University of Strasbourg observed that toxicity of alcohols to mammals increased as the water solubility of the alcohols decreased [724] More than a century ago in 1868, Crum-Brown and Fraser expressed the idea that the physiological action of a substance was a function of its chemical composition and constitution.[725] They published a relationship between biological activity and the chemical structure, which is considered to be the first formulation of a QSAR model The physiological activity 'Φ' was expressed as a function of the chemical structure C.

$$\Phi = f(C)$$

A few decades later, in 1893, Richet showed that the cytotoxicities of a diverse set of simple organic molecules were inversely related to their corresponding water

solubilities [726] At the turn of the 20th century, Hans Horst Meyer of the University of Marburg and Charles Ernest Overton of the University of Zurich, working independently, noted that the toxicity of organic compounds depended on their lipophilicity (olive oil/water partition coefficients).[727] In 1939 Ferguson introduced a thermodynamic generalization to the correlation of depressant action with the relative saturation of volatile compounds in the vehicle in which they were administered [728] The extensive work of Albert, and Bell and Roblin established the importance of ionization of bases and weak acids in bacteriostatic activity [729-731]

Meanwhile in 1930's, on the physical organic front, great strides were being made in the delineation of substituent effects on organic reactions, led by the seminal work of Hammett, which gave rise to the "sigma-rho" culture.[732,733] Hammett correlated electronic properties of organic acids and bases with their equilibrium constants and reactivity Taft devised a way for separating polar, steric, and resonance effects and introduced the first steric parameter, E_S [734] The contributions of Hammett and Taft together laid the mechanistic basis for the development of the QSAR paradigm by Hansch and Fujita by study on the structure-activity relationships of plant growth regulators and their dependency on Hammett constants and hydrophobicity [735] They later on combined hydrophobic constants with Hammett's electronic constants to yield the linear Hansch equation and its many extended forms like Hansch parabolic equation and Kubinyi bilinear model.[736,737]

$\text{Log } 1/C = a \cdot \sigma + b \cdot \pi + c\,k$	Linear form
$\text{Log } 1/C = a \cdot \log P - b \cdot (\log P)^2 + c \cdot \sigma + k$	Non linear form
$\text{Log } 1/C = a \cdot \log P - b \cdot \log(\beta \cdot P + 1) + k$	Kubinyi bilinear model

where, C = Concentration required to produce a standard response

Log P = Partition coefficient between octanol and water

σ = Hammet substituent parameter

π = Relative hydrophobicity of substituents

a, b, c, k = Model coefficients

In general, Hansch type studies are performed on compounds which contained a common template (usually a rigid one such as an aromatic ring) with structural variation limited to functional group changes at specific sites

While there are limits to the Hansch approach, it permitted complex biological systems to be modeled successfully using simple parameters. The approach has been used successfully to predict substituent effects in a wide number of biological assays The main problem with the approach was the large number of compounds which were required to adequately explore all structural combinations Further, the analysis methods did not lend themselves to the consideration of conformational effects. Several authors have published articles which provide additional background on the Hansch approach [738,739]

Alternative approaches to compound design have been suggested which avoid the combinatorial problem found in Hansch type analyses Free and Wilson used a series of substituent constants which related biological activity to the presence of a specific functional group at a specific location on the parent molecule [740] The relationship between biological activity and the presence or absence of a substituent was then expressed by the following equation

$$\text{Biological activity} = A + \sum_i \sum_j G_{ij} X_{ij}$$

where, A was defined as the average contribution of the parent molecule, G_{ij} the contribution to activity of a functional group i in the j^{th} position and X_{ij} the presence (1.0) or absence (0 0) of the functional group i in the j^{th} position

The procedure used the equation above to build a matrix for the series and represented this matrix as a series of equations Substituent constants then were derived for every functional group at every position Statistical tests were used to test the importance of the constants If the models were shown to be valid, the model was used to predict activity values for compounds which had not been prepared. In general, while a large number of compounds are required to explore the effects of multiple substitution patterns, the Free-Wilson approach substantially reduces the number of analogs required. However, the method demands that the effects of substituents are additive

Limitations in this approach led to the more sophisticated Fujita-Ban equation that used the logarithm of activity, which brought the activity parameter in line with other free energy-related terms [741]

$$\text{Log biological activity} = A + \sum_i \sum_j G_{ij} X_{ij}$$

Variations on this activity based approach have been extended by Klopman et al [742] and Blake et al.[743]

In 1972, John Topliss published a paper which detailed methodology to automate the Hansch approach [744] The method assumed that the lead compound of interest contained at least one phenyl ring which could serve as the template for functional group modifications The first modification to the template was preparation of the *para*-chloro derivative to examine lipophilicity. Additional substitution patterns were then made sequentially in an attempt to explore and optimize the relationship between activity and the hydrophobic and electronic character of the molecule. While the Topliss approach is easy to follow, it has several drawbacks The primary problems are that the procedure is not applicable to all types of studies and that there is a high degree of risk associated with its use (it essentially ignores the possibility of interactions between substituents as it changes one substituent at a time)

Topological methods have also been used to address the relationships between molecular structure and biological activity. The Minimum Topological Difference (MTD) method of Simon[745] and the extensive studies on molecular connectivity by Kier and Hall have contributed to the development of quantitative structure property/activity relationships.[746,747] Recently, these electrotopological indices that encode significant structured information on the topological state of atoms and fragments as well as their valence electron content have been applied to biological and toxicity data [748] Other developments in QSAR include approaches such as HQSAR (Hologram QSAR), Inverse QSAR, and Binary QSAR [749-753]

5.3. Molecular descriptors (parameters) used in QSAR

Parameters are of critical importance in determining the types of intermolecular forces those underlay drug-receptor interactions Molecular descriptors are numerical values obtained by the quantification of various structural and physicochemical

characteristics of the molecule. It is envisaged that molecular descriptors quantify these attributes so as to determine the behavior of the molecule and the way the molecule interacts with a physiological system. Since the exact mechanism of drug activity is unknown in many cases, it is desirable to start with descriptors spanning as many attributes of the molecules as possible and then assess their ability to predict the desired activity/property The three major types of parameters that were initially suggested and still hold way are hydrophobic, electronic and steric parameters

In the historical development of QSAR, no other parameter has generated more interest, excitement, and controversy than hydrophobicity (or lipophilicity) Molecular recognition depends strongly on hydrophobic interactions between ligands and receptors and thus very useful for bioactivity predictions Transport and distribution processes within biological systems are, to a large extent, controlled by lipophilicity of the system components The highly hydrophobic interior of a bilayer membrane enables or facilitates the passage of lipophilic substances and prevents the free diffusion of polar molecules except water in and out of cells The gain in entropy appears as the critical driving force for hydrophobic interactions that are primarily governed by the repulsion of hydrophobic solutes from the solvent water and the limited but important capacity of water to maintain its network of hydrogen bonds Lipophilicity is usually measured by determining the equilibrium concentration of the compound in two immiscible polar/non-polar liquids and expressed as the logarithm of the partition coefficient The octanol-water partition coefficient (log $P_{o/w}$) is the parameter most widely used to measure hydrophobicity because it has been shown that this partition system is a good model for many biological processes. Partition coefficients deal with neutral species, whereas distribution ratios incorporate concentrations of charged and/or polymeric species as well

However, experimental log Po/w measurements are time-consuming and are limited to a certain range, e.g $-3 < \log P_{o/w} < 3$ It is possible to obtain log $P_{o/w}$ values from fragmental contributions of the different atoms using computational calculations. Chromatographic methods have been also been used successfully to assess lipophilicity of organic compounds. Thus, the retention factor in gradient reversed phase high pressure liquid chromatography (RP-HPLC) with pure aqueous mobile phases (log k'_w) is commonly used as a lipophilic descriptor. Several advantages are

attributed to log k'_w A priori, it reflects polar/no polar partitioning in a manner similar to shake-flask measurements and it is dependent on the solute structure and polar functionalities However, it is difficult to measure directly, because of the prohibitively long retention times of organic solutes in pure water as mobile phase

On the other hand, electrostatic properties play a crucial role in the receptor-bioactive compound recognition process Extensive studies using electronic parameters reveal that electronic attributes of molecules are intimately related to their chemical reactivities and biological activities A great number of electrostatic descriptors have been described Basically, semiempirical and *ab initio* methods have to be distinguished in the electrostatic descriptors determinations The most frequently used descriptor includes frontier orbital electron densities, Mulliken's population charge distribution, superdelocalizabilities, dipole moments and others (Table 5 1) In this sense, some electrostatic properties i e electron affinity, LUMO's and HOMO's energies can be related to experimental properties such as redox potentials Furthermore, the latter is known to be related to bio-reduction or bio-oxidation processes of bioactive compounds The electrochemical response can be experimentally studied, i e by polarographic or by cyclic voltammetry techniques

The quantitation of steric effects is complex at best and challenging in all other situations, particularly at the molecular level. An added level of confusion comes into play when attempts are made to delineate size and shape Nevertheless, sterics are of overwhelming importance in ligand-receptor interactions as well as in transport phenomena in cellular systems The first steric parameter to be quantified and used in QSAR studies was Taft's constant E_s Other streic parameters considered in the QSAR studies are molecular weight (MW), molar refraction (MR) etc. as mentioned in Table 5 1

Table 5.1 Molecular Descriptors used in QSAR

Type	Description	Parameters
Hydrophobic parameters (Thermo-dynamic Descriptors)	Describes energy of molecules and their conversions	Partition coefficient (log P) Hansch's substitution constant (π) Hydrophobic fragmental constant (f, f') Distribution coefficient (log D) Apparent log P (fixed pH) (log P', log P_{app}) Capacity factor in HPLC (log k', log k'_w) Solubility parameter (log S)
Electronic Descriptors	Describe the electron orientation and charge	Hammett constants (σ, σ^+, σ^-) Taft's inductive (polar) constants (σ^*) Swain and Lupton field parameter (F) Swain and Lupton resonance parameter (R) Ionization constant (pK_a, ΔpK_a) Chemical shifts · IR, NMR
Quantum chemical Descriptors	Descriptors that are calculated using semi-empirical methods that are likely to be more accurate	Atomic net charge (Q^σ, Q^π) Superdelocalizability (S^N, S^E, S^R) Energy of highest occupied molecular orbital (E_{HOMO}) Energy of lowest unoccupied molecular orbital (E_{LUMO}) Electrostatic potential (V_r) Polarizability (POL)
Steric Descriptors	Describe the molecules' solvent-accessible surface areas (SASA) and their charges	Taft's steric parameter (E_s) Molar volume (MV) van der Waals radius (r_v) van der Waals volume (V_w) Molar refractivity (MR) Parachor (P_r) STERIMOL parameters (L, B_{1-4})
Topological Descriptors	Based from graph/structure concepts and geometric features such as shape, size, and branching	Molecular connectivity indices (χ) Valence molecular connectivity indices (χ^v) Kier's shape indices (κ)

5.4. Taxonomy of QSAR

Computational chemistry represents molecular structures as numerical models and simulates their behavior with the equations of quantum and classical physics. Available programs enable scientists to generate and present molecular data including geometries, energies and associated properties (electronic, spectroscopic and bulk) The usual paradigm for displaying and manipulating these data is a table in which compounds are defined by individual rows and molecular properties (or descriptors) are defined by the associated columns

On the basis of the origin of molecular descriptors used in calculations, QSAR methods can be divided into several groups [754] First group is based on a relatively small number (usually many times smaller than the number of compounds in a data set) of physicochemical properties and parameters describing, for example, hydrophobic, steric, and electrostatic effects Usually, these descriptors are used as independent variables in multiple regression approaches These methods are typically referred to as Hansch analysis Second type of methods is based on quantitative characteristics of molecular graphs (molecular topological descriptors) Because molecular graphs or structural formulas are "two-dimensional (2D)", these methods are referred to as 2D QSAR Most of the 2D QSAR methods are based on graph theoretical indices, for example, molecular connectivity indices, molecular shape indices, topological and electrotopological state indices and atom-pair descriptors Sometimes, topological descriptors are also combined with physicochemical properties of molecules The third group of methods is based on descriptors derived from spatial representation of molecular structures Correspondingly, these methods are referred to as three-dimensional or 3D QSAR, they have become increasingly popular with the development of fast and accurate computational methods for generating 3D conformations and alignments of chemical structures The early examples of 3D QSAR include molecular shape analysis (MSA)[755], distance geometry[756,757], and Voronoi techniques.[758] The first method uses shape descriptors and multiple linear regression analysis, whereas the latter methods apply atomic refractivity as structural descriptors and the solution of mathematical inequalities to obtain the quantitative relationships. Two original 3D QSAR methods, CoMFA (Comparative Molecular Field Analysis)[759-761] and GRID,[762] were developed almost simultaneously in the mid- to late-1980s. Since the introduction, CoMFA and

COMSiA (Comparative Molecular Similarity Indices Analysis)[763,764] approaches, perhaps two of the most popular examples of 3D QSAR, have rapidly become popular in 3D QSAR. These techniques have elegantly combined the power of 3D molecular modeling and partial least-square (PLS) optimization technique[765,766] and found wide applications in medicinal chemistry and toxicity analysis. Both are based on the assumption that changes in binding affinities of ligands are related to changes in molecular properties, represented by fields

Descriptors in the case of CoMFA[760,767] and CoMFA-like methods such as COMSiA, Comparative binding energy (COMBINE)[768], and quantitative similarity-activity relationship (QsiAR)[769] represent electrostatic, steric, and hydrophobic field values in the grid points surrounding molecules. They differ only in the implementation of the fields. In COMSiA, five different similarity fields are calculated steric, electrostatic, hydrophobic, hydrogen bond donor, and hydrogen bond acceptor. These fields were selected to cover the major contributions to ligand binding. Similarity indices are calculated at regularly spaced grid points for the pre-aligned molecules In a standard CoMFA procedure, all molecules under investigation are first structurally aligned, and the steric and electrostatic fields around them are sampled with probe atoms, usually sp^3 carbon with a +1 charge, on a rectangular grid that encompasses aligned molecules

The 4D QSAR methodology is an extension of the 3D QSAR methodology developed by Hopfinger et al.[770] which considers conformational information as the fourth dimension Data sets for 4D QSAR include all possible conformations, orientations and, in some cases, protonation states. In most cases, binding between a ligand and site arises from weak interactions, such as hydrogen bonds formed between proton donors and acceptors Covalent interactions, where bonds are broken and reformed, tend to be less important Similar to the CoMFA method, 4D QSAR starts by defining a set of grid points on which molecular properties will be evaluated In addition to the grid points, the method performs conformational ensemble sampling and uses the information obtained to evaluate grid cell occupancies. These occupancies are then used to evaluate interaction pharmacophore elements (IPE's) The IPE's together with the molecular properties are then used to develop a predictive model. The 4D

QSAR models use a genetic algorithm that selects the most bioactive conformation, which integrates into the best model

Additional dimensions further hone predictive powers For instance, 5D QSAR allows the model to include induced fit, which occurs during the binding of many ligands [771] In this case, the ligand changes the protein's shape, bringing the active parts into proximity with the substrate Similarly the additional dimension in 6D QSAR[772] simultaneously considers various salvation models, which is when solute and solvent molecules combine using relatively weak covalent bonds, to screen for adverse effects *in silico*. The "dimensional" approach is not the only novel model that improves QSAR analysis (COMBINE) analysis yields predictive QSARs as well as mechanistic insights Finally, QSAR methods can also be classified by the type of the correlation methods used in model development Linear methods include linear regression or Multiple Linear Regression (MLR), PLS, or principal component regression (PCR), whereas nonlinear methods can be exemplified, for example, by *k*-Nearest Neighbors (*k*NN)[773,774] and artificial neural networks (ANN)[775] methods

Table 5.2 Defining Dimensions in QSAR

Dimension	Definition
1D-QSAR	Affinity correlates with pKa, logP, etc.
2D-QSAR	Affinity correlates with a structural pattern (e g , chemical connectivity).
3D-QSAR	Affinity correlates with the three-dimensional structure
4D-QSAR	As with 3D, but with multiple representations of ligand conformation/orientation
5D-QSAR	As with 4D, but with multiple representations of induced-fit scenarios.
6D-QSAR	As with 5D, but with multiple representations of solvation models.

Different QSAR methods have their own strengths and weaknesses For example, 3D QSAR methods generally result in the diagrams of important molecular fields that can be easily interpreted in terms of specific steric and electrostatic interactions important for the ligand binding to their receptor. However, the need to align structures in 3D, which is time-consuming and subjective, precludes the use of 3D QSAR techniques for the analysis of large data sets On the other hand, 2D QSAR methods are much faster and more amenable to automation because they require no

conformational search and structural alignment. Thus, 2D methods are best suited for the analysis of large numbers of compounds and computational screening of molecular databases; however, the interpretation of the resulting models in familiar chemical terms is frequently difficult, if not impossible.

5.5. Tools and techniques of QSAR

Drugs exert their biological effects by participating in a series of events which include transport, binding with the receptor and metabolism to an inactive species Since the interaction mechanisms between the molecule and the putative receptor are unknown in most cases (i e., no bound crystal structures), one is reduced to making inferences from properties which can easily be obtained (molecular properties and descriptors) to explain these interactions for known molecules Once the relationship is defined, it can be used to aid in the prediction of new or unknown molecules The relationship is usually a mathematical expression derived by statistical and related techniques The parameters describing physicochemical properties are used as independent variables and the biological activities are dependent variables

5.5.1. Biological Parameters

In QSAR analysis, it is imperative that the biological data be both accurate and precise to develop a meaningful model It must be realized that any resulting QSAR model that is developed is only as valid statistically as the data that led to its development The equilibrium constants and rate constants that are used extensively in physical organic chemistry and medicinal chemistry are related to change in free energy values (ΔG) Percentage activities (e g % inhibition of growth at certain concentrations) are not appropriate biological endpoints because of the nonlinear characteristic of dose-response relationships These types of endpoints may be transformed to equieffective molar doses. Only equilibnum and rate constants pass muster in terms of the free-energy relationships or influence on QSAR studies. Biological data are usually expressed on a logarithmic scale (log 1/C) because of the linear relationship between response and log dose in the midregion of the log dose-response curve Inverse logarithms for activity are used so that higher values are obtained for more effective analogs

Biological data should pertain to an aspect of biological/biochemical function that can be measured Because there is considerable variation in biological responses, test samples should be run in duplicate or preferably triplicate, except in whole animal studies where assay conditions (e g., plasma concentrations of a drug) preclude such measurements It is also important to design a set of molecules that will yield a range of values in terms of biological activities Generally, the larger the range (>2 log units) in activity, the easier it is to generate a predictive QSAR This kind of equation is more forgiving in terms of errors of measurement In the case of isolated receptors, the endpoint is clear-cut and the critical step is evident But in more complex systems, such as cellular systems or whole animals, many localized steps could be involved in the random-walk process and the eventual interaction with a target Usually the observed biological activity is reflective of the slow step or the rate-determining step Various types of biological data have been used in QSAR analysis A few common endpoints are outlined in Table 5 3

Table 5.3 Types of Biological Data Utilized in QSAR Analysis

Source of Activity	Biological Parameters
1. Isolated receptors	
Rate constants	Log k_{cat}, Log k_{uncat}, Log k
Michaelis-Menten constants	Log $1/K_m$
Inhibition constants	Log $1/K_i$
Affinity data	pA_2, pA_1
2. Cellular systems	
Inhibition constants	Log $1/IC_{50}$
Cross resistance	Log *CR*
In vitro biological data	Log *1/C*
Mutagenicity states	Log TA_{98}
3. *"In vivo"* systems	
Bioconcentration factor	Log *BCF*
In vivo reaction rates	Log *I* (Induction)
Pharmacodynamic rates	Log *T* (total clearance)

5.5.2. Statistical methods

The most widely used mathematical technique in QSAR analysis is MLR analysis Regression analysis is a powerful means for establishing a correlation between independent variables and a dependent variable such as biological activity

$$Y_i = b + aZ_i + E$$

Certain assumptions are made with regard to this procedure

1 The independent variables, which in this case usually include the physicochemical parameters, are measured without error. Unfortunately, this is not always the case, although the error in these variables is small compared to that in the dependent variable.
2 For any given value of X, the Y values are independent and follow a normal distribution The error term E_i possesses a normal distribution with a mean of zero
3 The expected mean value for the variable Y, for all values of X, lies on a straight line
4 The variance around the regression line is constant The "best" straight line for model $Y_i = b + aZ_i + E$ is drawn through the data points, such that the sum of the squares of the vertical distances from the points to the line is minimized. Y represents the value of the observed data point and Y_{calc} is the predicted value on the line.

The correlation coefficient r is a measure of quality of fit of the model. It constitutes the variance in the data In an ideal situation one would want the correlation coefficient to be equal to or approach 1, but in reality because of the complexity of biological data, any value above 0 90 is adequate The standard deviation is an absolute measure of the quality of fit Ideally *s* should approach zero, but in experimental situations, this is not so It should be small but it cannot have a value lower than the standard deviation of the experimental data. The magnitude of *s* may be attributed to some experimental error in the data as well as imperfections in the biological model. A larger data set and a smaller number of variables generally lead to lower values of *s* The *F* value is often used as a measure of the level of statistical significance of the regression model A larger value of *F* implies a more significant

correlation has been reached. The confidence intervals of the coefficients in the equation reveal the significance of each regression term in the equation

To obtain a statistically sound QSAR, it is important that certain caveats be kept in mind. One needs to be cognizant about collinearity between variables and chance correlations Use of a correlation matrix ensures that variables of significance and/or interest are orthogonal to each other With the rapid proliferation of parameters, caution must be exercised in amassing too many variables for a QSAR analysis Outliers in QSAR model generation present their own problems. If they are badly fit by the model (off by more than 2 standard deviations), they should be dropped from the data set, although their elimination should be noted and addressed Their aberrant behavior may be attributed to inaccuracies in the testing procedure (usually dilution errors) or unusual behavior They often provide valuable information in terms of the mechanistic interpretation of a QSAR model They could be participating in some intermolecular interaction that is not available to other members of the data set or have a drastic change in mechanism.

In some cases, the types of biological data, the choice of descriptors, and the class of optimization methods are closely related and mutually inclusive. For instance, multiple linear regression can be applied only when a relatively small number of molecular descriptors are used (at least five to six times smaller than the total number of compounds) and the target property is characterized by a continuous range of values The use of multiple descriptors makes it impossible to use MLR because of a high chance of spurious correlation[776] and requires the use of partial least squares or nonlinear optimization techniques

5.5.3. Compound selection

In setting up to run a QSAR analysis, compound selection is an important angle that needs to be addressed One of the earliest manual methods was an approach devised by Craig,[777] which involves 2D plots of important physicochemical properties Care is taken to select substituents from all four quadrants of the plot The Topliss operational scheme allows one to start with two compounds and construct a potency tree that grows branches as the substituent set is expanded in a stepwise fashion.[21]

Other methods of manual substituent selection include the Fibonacci search method, sequential simplex strategy, and parameter focusing by Magee [778-780]

One of the earliest computer-based and statistical selection methods, cluster analysis was devised by Hansch to accelerate the process and diversity of the substituents [736] Newer methodologies include D-optimal designs, which focus on the use of det (X'X), the variance- covariance matrix The determinant of this matrix yields a single number, which is maximized for compounds expressing maximum variance and minimum covariance [781-783] A combination of fractional factorial design in tandem with a principal property approach has proven useful in QSAR [784] Extensions of this approach using multivariate design have shown promise in environmental QSAR with nonspecific responses, where the clusters overlap and a cluster-based design approach has to be used [785] With strongly clustered data containing several classes of compounds, a new strategy involving local multivariate designs within each cluster is described. The chosen compounds from the local designs are grouped together in the overall training set that is representative of all clusters [786]

Table 5.4 2D QSAR methods

1	Free energy models	Hansch analysis (Linear Free Energy Relationship, LFER)
2	Mathematical models	Free Wilson analysis
		Fujita-Ban modification
3.	Other statistical methods	Discriminant analysis (DA)
		Principle Component Analysis (PCA)
		Cluster Analysis (CA)
		Combine Multivariate Analysis (CMA)
		Factor Analysis (FA)
4	Pattern recognization	
5	Topological methods	
6	Quantum mechanical methods	

5.6. QSAR study of a series of 2,3-dihydroimidazo[1,2-c]pyrimidines by Hansch model

5.6.1. General procedures

Optimization of structures

All the imidazo[1,2-*c*]pynmidine structures were built on workspace of ChemDraw Ultra (Version 11 0, CambridgeSoft Corporation) and the descriptor calculations were done in the molecular package TSAR 3D (version 3 3 for Windows, Accelrys Software Inc) The calculation of the quantum-chemical descriptors for different compounds were performed using a semiempincal molecular orbital method (AM1), starting from standard bond lengths and bond angles. All geometries were fully optimized by minimizing the energy with respect to geometrical variables without symmetry constraints, using a 0 01 kcal/mol gradient. Most stable structure for each compound was generated and used to calculate various physicochemical descriptors like thermodynamic, steric and electronic values of descriptors (Table 5 7)

Descriptors calculation, QSAR model development and validation

The calculation of molecular descriptors of 2,3-dihydroimidazo[1,2-*c*]pyrimidines as well as the regression analyses were carried out using the molecular package TSAR 3D (version 3 3 for Windows, Accelrys Software Inc.) All the calculated descriptors were considered as independent variable and biological activity as dependent variable Various QSAR models were generated by stepwise regression (MLR). Statistical measures used were n = number of compounds, r = correlation coefficient, r^2 = squared correlation coefficient, F-test (Fischer's value) for statistical significance, SEE = standard error of estimation, q^2 = cross validated correlation coefficient and correlation matrix to show mutual correlation among the parameters The predictive ability of the generated correlations was evaluated by cross validation method employing a 'leave-one-out' scheme

The squared correlation coefficient (or coefficient of multiple determination) r^2 is a relative measure of fit by the regression equation Correspondingly, it represents the part of the variation in the observed data that is explained by the regression. The correlation coefficient values closer to 1.0 represent the better fit of the regression. The F-test reflects the ratio of the variance explained by the model and the variance due to the error in the regression High values of the F-test indicate that the model is

statistically significant Standard deviation is measured by the error mean square, which expresses the variation of the residuals or the variation about the regression line Thus standard deviation is an absolute measure of quality of fit and should have a low value for the regression to be significant.

5.6.2. Results and discussion

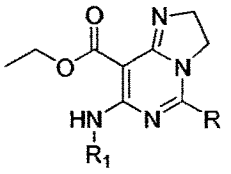
To identify the chemical structural features required for antimycobacterial activity of 2,3-dihydroimidazo[1,2-*c*]pyrimidines, QSAR study was performed as a part of our composite program of rational drug design In multivariate statistics, it is common to define three types of outliers.

1 X/Y relation outliers are substances for which the relationship between the descriptors (X variables) and the dependent variables (Y variables) is not the same as in the (rest of the) training data.
2. X outliers in which a substance is an X outlier if the molecular descriptors for this substance do not lie in the same range as the (rest of the) training data
3. Y outliers are only defined for training or test samples They are substances for which the reference value of response is invalid

A set of 13 compounds out of thirty 2,3-dihydroimidazo[1,2-*c*]pyrimidines exhibiting significant antimycobacterial activity was used to develop the QSAR models separately In light of the above guidelines, 3 compounds of imidazopyrimidine series were considered as outliers because their residual values were higher in comparison to the other compounds included in the present study. Compounds were divided into training and test sets consisting of 13 and 3 molecules, respectively The training set has been used for QSAR model development and the test set was used to validate the developed QSAR models. The activity data have been given as minimum inhibitory concentration (MIC) values and converted to micromolar units and then further to -log scale (pMIC) and subsequently used as the response variable (a dependent variable) for the QSAR analysis

The pMIC values, along with the structure of compounds used in regression analysis, are presented in Table 5 5

Table 5.5 Antimycobacterial activity of imidazo[1,2-*c*]pyrimidines



No	Comp	R_2	R_6	pMIC
Training Set				
1	IF 03	H	$C_6H_5CH_2$	7 174
2	IF 04	H	C_6H_5	7.153
3	IF 07	H	4-CH_3-C_6H_4	7 776
4	IF 09	H	4-CH_3O- C_6H_4	8.196
5	IF 10	H	2-F-C_6H_4	7.782
6	IF 11	H	4-F- C_6H_4	7.782
7	IF 15	H	4-Cl-C_6H_4	7.805
8	IF 16	H	4-Br-C_6H_4	6 861
9	IB 05	C_6H_5	2-CH_3-C_6H_4	6.573
10	IB 06	C_6H_5	4-CH_3-C_6H_4	6 573
11	IB 08	C_6H_5	4-CH_3O-C_6H_4	7.893
12	IB 13	C_6H_5	4-Cl-C_6H_4	7 296
13	IB 14	C_6H_5	4-Br- C_6H_4	6 643
Test Set				
14	IF 12	H	3-CF_3-C_6H_4	6 246
15	IB 04	C_6H_5	C_6H_5	6 475
16	IB 10	C_6H_5	4-F-C_6H_4	7.277

Various molecular descriptors (Table 5 6) calculated theoretically are presented in Table 5 7 and their correlation with the biological activity was established using Hansch LFER model.

Table 5.6 Molecular descriptors calculated for QSAR study

Sr No	QSAR descriptor	Type
1	Log P	Lipophilic
2	Lipole	Lipophilic
3	Lipole components	Lipophilic
4	Energy of highest occupied molecular orbital (HOMO)	Quantum chemical
5	Energy of lowest unoccupied molecular orbital (LUMO)	Quantum chemical
6	Zero order molecular connectivity indices ($^{0}\chi$)	Topological
7	First order molecular connectivity indices ($^{1}\chi$)	Topological
8	Second order molecular connectivity indices ($^{2}\chi$)	Topological
9	Third order molecular connectivity indices ($^{3}\chi$)	Topological
10	Valence zero order molecular connectivity indices ($^{0}\chi^{v}$)	Topological
11	Valence first order molecular connectivity indices ($^{1}\chi^{v}$)	Topological
12	Valence second order molecular connectivity indices ($^{2}\chi^{v}$)	Topological
13	Valence third order molecular connectivity indices ($^{3}\chi^{v}$)	Topological
14	Kier's zero order shape indices (κ_0)	Topological
15	Kier's first order shape indices (κ_1)	Topological
16	Kier's second order shape indices (κ_2)	Topological
17	Kier's third order shape indices (κ_3)	Topological
18	Ionization potential (IP)	Electronic
19	Total Dipole (TD)	Electronic
20	Dipole moment (μ)	Electronic
21	Dipole Moment -X Axis (DX)	Electronic
22	Dipole Moment -Y Axis (DY)	Electronic
23	Dipole Moment -Y Axis (DZ)	Electronic
24	Molar refractivity (MR)	Polarizability
25	Mean Polarizability (POL)	Polarizability

Table 5.7 Values of molecular descriptors used in MLR analysis

No	Comp	log P	Lipole	HOMO	LUMO	IP	TD	$^{1}\chi^{v}$	$^{2}\chi^{v}$	MR	POL
1	IF 03	2 398	8 459	-8 356	-0 399	8 356	4 067	7 341	4 927	83 038	39 378
2	IF 04	2 304	8 550	-8 434	-0 620	8 434	4 318	6 884	4 578	78 203	37 464
3	IF 07	2 771	9 517	-8 385	-0 602	8 385	4 491	7 295	5 078	83 244	39 875
4	IF 09	2 051	3 825	-8 341	-0 585	8 341	5 252	7 407	4 941	84 666	40 083
5	IF 10	2 443	8 725	-8 500	-0 705	8 500	5 568	6 990	4 692	78 419	36 854
6	IF 11	2 443	9 331	-8 518	-0 740	8 518	4 310	6 984	4 719	78 419	36 937
7	IF 15	2 822	11 941	-8 532	-0 739	8 532	4 335	7 393	5 191	83 008	38 065
8	IF 16	3 095	13 369	-8 559	-0 781	8 559	4 375	7 795	5 656	85 826	38 458
9	IB 05	4 399	14 048	-8 324	-0 749	8 324	5 424	9 393	6 542	108 008	52 464
10	IB 06	4 399	13 767	-8 304	-0 707	8 304	5 903	9 387	6 594	108 008	52 149
11	IB 08	3 679	9 785	-8 291	-0 736	8 291	6 642	9 499	6 456	109 430	52 450
12	IB 13	4 450	11 367	-8 847	-0 916	8 847	6 918	9 484	6 706	107 771	49 722
13	IB 14	4 723	11 987	-8 864	-0 932	8 864	6 992	9 887	7 171	110 590	50 060

In the first step the statistical method of stepwise multiple regression procedure, based on the forward selection method, was applied for variable selection with the aim to obtain the best regression equation (in such a way that variables that show little increment or are redundant in the explanation of the dependent variable were rejected).

The generated equations are presented in Table 5 8. Only those models were selected which had shown fairly good correlation of the physicochemical properties with the antimycobacterial activity of the test compounds

Preliminary statistical surveys of QSAR equations suggest that the influence of individual parameters is less on the biological activity as indicated by low values of r (equations 1 – 3). However, log P and MR individually were found to have reasonable correlation with biological activity (equation 1 and 3) but correlation of log P with biological activity was found more significant than MR as indicated by F-value (equation 1)

Table 5.8 QSAR Equations for the antimycobacterial activity of imidazo[1,2-*c*]pyrimidines

No	Equation	n	r	r^2	*s*	*F*
1	[(-0 3882)(±0 1303) log P] + 8 6(±0 4379)	13	0 668	0 447	0 437	8 872 (4 96)
2	[(-0 1022)(±0 1559) TD] + 7 886(±0 838)	13	0 194	0 038	0 576	0 43 (4 96)
3	[(-0 0208)(±0 0105) MR] + 9 2647(±0 9792)	13	0 512	0 263	0 504	3 917 (4 96)
4	[(-0 7161)(±0 16) log P] + [(0 3916)(±0 1452) TD] + 7 5926(±0 5115)	13	0 824	0 68	0 348	10 604 (4 10)
5	[(-0 9317) + (±0 3568) log P] + [(0 0404)(±0 0249) MR] + 6 6298(±1 2826)	13	0 749	0 562	0 408	6 405 (4 10)
6	[(0 4156)(±0 2267) TD] + [(-0 0476)(±0 0175) MR] + 9 5433(±0 9014)	13	0 669	0.448	0 457	4 059 (4 10)
7	[(-0 8418)(±0 3214) log P] + [(0 3453)(±0 182) TD] + [(0 0122)(±0 0267) MR] + 7 1156(±1 171)	13	0 828	0 687	0 363	6 580 (3 71)

We extended our study for multi-parameter correlations as they are permitted for a data set of 13 compounds in accordance with the lower limit of rule of thumb According to this rule for a QSAR model development one should select one parameter for a five compound data set The multicolinearity or auto-correlation between the parameters is indicated by the change in signs of the coefficients, a change in the values of previous coefficient, change of significant variable into insignificant one or an increase in standard error of the estimate on addition of an additional parameter to the model To reduce the risk of finding spurious relationships, no simultaneous highly correlated descriptors were considered ($r < \pm 0\ 75$ in intervariable correlation, Table 5.9).

Among the multi-parametric models, the bi- and tri-parametric models showed improvement in the correlation. While MLR analysis in the biparametric model with TD and MR didn't show any improvement in the correlation (equation 6, r = 0 669), an increase in r-value was observed by the inclusion of the molecular descriptor MR to log P (equation 5, r = 0.749) The addition of the electronic parameter TD to log P in the biparametric model significantly increased the correlation coefficient (equation 4, r = 0 824) improving the predictive power of the QSAR model The correlation was

only slightly increased when log P, TD and MR were included in the tri-parametric model (equation 7, r = 0 828)

From the analysis of various QSAR equations it can be concluded that there is good correlation of biological activity and log P with TD or MR in the biparametric MLR (equations 4 and 5); and with all three descriptors in tri-parametric MLR (Equation 7) A high value of r for equation 4 and 7 explains good correlation between the physicochemical parameters with the antimycobacterial activity and the significance of these equations at 95%. Due to small value of the coefficient for MR in the equation 7 (0.0122) and t_{cal} (0 458) being less than t_{tab} (2.228), it fails at t-test This indicates insignificant contribution of this molecular descriptor towards the biological activity and hence equation 7 has not been considered for prediction. Furthermore, the value of F-test for equation 4 (F_{cal} = 10.604) is greater than the F_{tab} even at 1% significance level suggesting that this QSAR model is significant at 99% probability level (F_{tab} = 7 56).

In order to confirm our results we have estimated the antimycobacterial activity (pMIC) of training and test sets using both the equations (4 and 7) and compared them with the observed values. The data presented in Table 5.9 show that the observed and predicted biological activities by equation 4 are very close to each other, which is evident by low residuals This was further supported by the plot of MLR predicted pMIC values against the observed pMIC values (Figure 5 1).

It is seen that the equation 4 (F = 10 604) is statistically more significant than equation 7 (F = 6 580) at the significance level of 95% Thus, the predictive power of equation 4 is more than the later one and found to be more suitable and validated model for further predictions.

Table 5.9 Observed and predicted antimycobacterial activity of substituted imidazo[1,2-*c*]pyrimidines using the best MLR models

No	Comp	Equation 4			Equation 7		
		Obs pMIC	Calc pMIC	Residual	Obs pMIC	Calc pMIC	Residual
Training Set							
1	IF 03	7 174	6 029	1 145	7 174	7 514	-0 340
2	IF 04	7 153	7 634	-0 481	7 153	7 621	-0 469
3	IF 07	7 776	7.367	0.408	7 776	7 350	0 426
4	IF 09	8 196	8 181	0 016	8 196	8 236	-0.039
5	IF 10	7 782	8.023	-0 242	7 782	7 938	-0 157
6	IF 11	7 782	7 531	0 250	7 782	7 504	0.277
7	IF 15	7 805	7.27	0 535	7 805	7 250	0 555
8	IF 16	6 861	7 089	-0 228	6 861	7 068	-0 207
9	IB 05	6 573	6 567	0 007	6 573	6 603	-0.030
10	IB 06	6 573	6 754	-0 180	6 573	6 769	-0 195
11	IB 08	7 893	7 559	0 333	7 893	7 647	0 245
12	IB 13	7 296	7 115	0 180	7 295	7 074	0 222
13	IB 14	6 643	6 948	-0 305	6 643	6.903	-0 260
Test Set							
14	IF 12	6 246	7 929	-1 683	6.246	7.769	-1 523
15	IB 04	6 475	8 129	-1 655	6 475	8.174	-1 700
16	IB 10	7 277	7.392	-0 115	7 277	7.341	-0 064

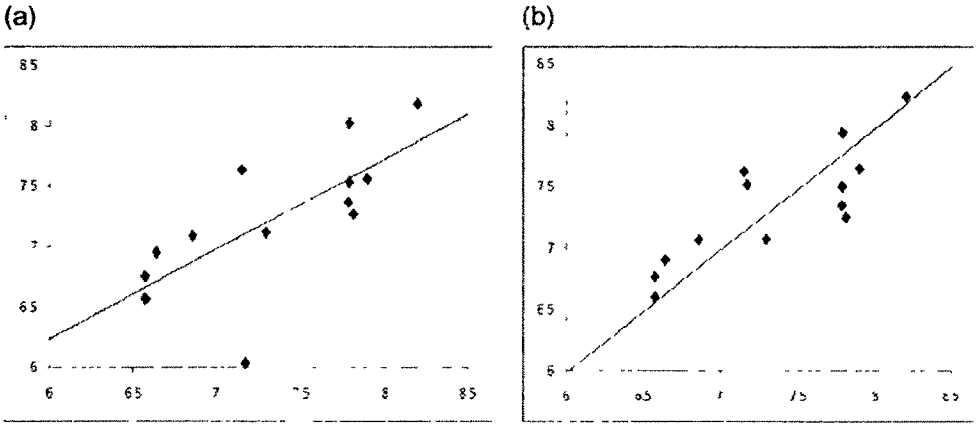


Figure 5.1 Plot of predicted Vs observed pMIC values for the MLR model by (a) equation 4 and (b) equation 7

The large coefficient of lipophilic parameter (log P) in equation 4 indicates its higher influence on the biological activity The negative value suggests that the lipophilicity of the molecule should be decreased and positive TD coefficient explains incorporation of electron withdrawing substitutions Thus, inclusion of electron withdrawing and less lipophilic substituents may enhance the potency of the molecule which is evident from the results of the antimycobacterial activity. The key factor observed was that the phenyl group at C-5 has decreased the potency of imidazo[1,2-*c*]pyrimidines significantly than the unsubstituted one