



# Quantitative Structure-Activity Relationship (QSAR): A Review

**Aniket Lakshaman Jadhav , Sangram Kondiba Nazirkar, Aditya Vilas Khomane**

**Guide Name: Prof. Pandurang Vijapure**

Bachelor Of Pharmacy, Sarsam College Of Pharmacy Palshiwadi,

Dr. Babasaheb Ambedkar Technological University, Lonere, Maharashtra, India

**Abstract:** Quantitative Structure-Activity Relationship (QSAR) is attempt to quantitatively correlate structural and biological property of compound or molecule. It dates back to the nineteenth century and now it has advancement from QSAR to 3D QSAR. QSAR are used in the drug design and medicinal chemistry. In this article we discuss various QSAR model such as Hansch Analysis, Free Wilson Analysis along with various physicochemical properties such as lipophilic parameter, electronic parameter and steric factor. Along with various method and descriptor used in QSAR, 2D QSAR and 3D QSAR. The application and limitation in QSAR study.

**Keywords:** QSAR, Physicochemical property, QSAR descriptors

## 1. INTRODUCTION

The quantitative structure-activity relationship (QSAR) is an attempt to quantitatively correlate structural or property descriptors of compounds with biological activities. The physicochemical descriptor consists of parameter account for constitutional, thermodynamic, fragment constant, conformational, hydrophobicity, topology, electronic properties, hydrogen bond donor, hydrogen bond acceptor, steric effect are determined by the computational method. Structure relates to the property or descriptors of the molecules in QSAR, whereas function corresponds to a biological/biochemical experiment. QSAR have lot of advancement in drug design and drug development. From the protein binding affinity and toxicity determination to rate constants various kind of activities are performed by QSAR. It also includes chemical measurement and biological assay. If biological property determines then it refers as Quantitative structure-property relationship (QSPR) and toxicology determine then it refers as Quantitative structure-toxicity relationship (QSTR).

## 2. HISTORY

The quantitative structure-activity relationship dates back to the nineteenth century. In 1866, “Crum Brown and Fraser” published QSAR equation first time. It considers the first formulation of the quantitative structure-activity relationship. Richart et al (1983) reported that the toxicity of organic compounds inversely follow their water solubility or differ in biological activity due to changes in chemical and physiological properties. The QSAR well defined later by Fujita and Ban in 1970.

## 3. QSAR MODELS

Since the generation of QSAR technique various models in QSAR introduced:

### 3.1. Hansch Analysis:

These is linear free related energy approach divided into two classes:

#### *Linear Models Corwin*

The significance of lipophilicity, defined as the octanol-water partition coefficient ( $P$ ), on biological activity was identified by Hansch in 1969. This parameter is a measure of a compound bioavailability, which decide, how much of the compound reaches the target. The equation is as:

$$\log(1/C) = a \log P + b$$

In these equation ‘ $C$ ’ is molar concentration of compound that produces standard response (e.g., LD50, ED50, IC50, EC50 etc) The correlation improved by combining Hammett’s electronic parameter and Hansch’s measure of lipophilicity by using equation as:

$$\log(1/C) = k_1\pi + k_2\sigma + k_3$$

Where  $\sigma$  is the Hammett substituent parameter and  $\pi$  is analogously to  $\sigma$ . [2]

#### *Non-Linear Models*

The failure of linear equations in circumstances with broad hydrophobicity ranges prompted the development of the Hansch parabolic equation, which includes a  $(\log P)^2$  element in QSAR equation. This can be explained in one of two ways, a word that refers to the fact that multiple membranes must be crossed in order for compounds to pass through to get the desired location, and those that have the most the hydrophobicity of the membranes will become localised initially come across. The Hansch approach correlates changes in chemical structures with changes in lipophilic, electronic, and occasionally steric substituent characteristics in biological reaction. It is expressed mathematically as:

$$\begin{aligned} \log(1/C) &= \Delta G_h + \Delta G_e + \Delta G_s + \text{constant} \\ \log(1/C) &= a \log P - b (\log P)^2 + c\sigma + dE_s + \text{constant} \end{aligned}$$

Where  $\log P$  is logarithm of partition coefficient,  $\sigma$  is Hammett electronic constant and  $E_s$  is Taft steric constant.  $a$ ,  $b$ ,  $c$  and  $d$  are the coefficients determined by multiple regression analysis to fit the biological data. [1,2]

**Advantages:**

- 1) The small organic molecules descriptors ( $\sigma$ ,  $\pi$ ,  $E_s$  etc.) can be used to describe biological systems.
- 2) Predictions are quantitative and measurable statistically.
- 3) Quick and easy.
- 4) Potential extrapolation

**Disadvantages:**

- 1) The compounds are require in large number.
- 2) Use of small molecule descriptors on biological systems are mostly limitations.
- 3) In biological systems, steric factors have a restricted application.
- 4) Drug partial protonation in physiological conditions. [1,2]

**3.2. Free Wilson Analysis / De Novo Approach:**

The Free Wilson technique is a structure-activity relationship model that is based on facts. For each structural trait that differs from a set of randomly chosen compounds, an indicator variable is created.

$$\log(1/C) = \sum a_j X_{ij} + \mu$$

This de novo technique, like standard QSAR, assumes that substituent effects are additive and constant. The biological activity is represented by  $\log(1/C)$ . The third substituent,  $X_j$ , has a value of 1 if it is present (a specific substituent or structural characteristic), and 0 if it is not. The word  $a_j$  represents the total average activity and denotes the contribution of the  $j$ th substituent to biological activity. In each position, the sum of all activity contributions must equal zero. Linear regression analysis is used to solve a set of linear equations that have been formulated.

**Advantages**

1. It is simple to create a table for regression analysis.
2. The insertion and removal of compounds is straightforward and has little impact on the values of other regression coefficients.
3. As a reference compound, any compound can be chosen.
4. A pseudosubstituent is made up of two substituents that always occur together in two distinct locations of the molecule.
5. Problems with singularity are usually avoided.

**Limitations**

1. First and foremost, structural variation is required at at least two different substitution positions. Otherwise, a meaningless group contribution, one for each component would arise.
2. It has the drawback of not providing a decent foundation for interpreting the findings in terms of a drug-receptor interaction.
3. The related group contribution contains the entire experimental error of this one biological data set for every substituent that only appears in the data set, at least to single point determination.

4. In most circumstances, a large number of parameters are required to characterise a small number of compounds, resulting in equations that are statistically insignificant.

### 3.3. Mixed Approach

According to Singer and Purcell, there is a link between Hansch analysis and Free-Wilson analysis (1967). Because the methodologies of Hansch analysis and Free-Wilson analysis are so similar, they can both be employed in the same framework. This is due to their theoretical consistency and numerical activity contribution equivalencies. The mixed approach is the name given to this development, which can be represented by the equation below.[2]

$$\log(1/C) = \sum a_j + \sum c_j \theta_j + \text{Constant}$$

The contribution for each *i*th substituent is denoted by the word  $a_j$ , whereas  $\theta_j$  denotes any physiological or chemical attribute of  $X_j$  is a substitute. Mixed approach based on the following assumption:

1. All of the drugs in the study have the same parent structure.
2. Distinct derivatives have different substitution patterns to be identical
3. The role of substitution in biological activity to be additive
4. Unaffected by the presence or absence of other factors substitution

### 3.4. Other Approaches

In the last two decades, pattern recognition techniques have received a lot of attention. In concept, they are similar to traditional QSAR method and pattern recognition. The number of variables in a pattern recognition system is the only thing that matters. The study is significantly higher than Hansch's analysis. Consistent Multivariate approaches, such as principal component analysis, are used to achieve outcomes. Techniques such as component analysis or soft modelling, for example, SIMCA or PLS analysis are two options.

Many diverse but inherently linked QSAR techniques begin with a hyperstructure, which is a hypothetical molecule that has all structural properties of the compounds being studied. In the step-wise optimization approach, the presence and absence of specific hyperstructure atoms or groups in individual molecules are connected with biological activities. For example, LOCON and LOGANA. [2]

## 4. PHYSICOCHEMICAL PROPERTIES

The physicochemical properties of QSAR include lipophilic parameter, electronic parameter and steric factor or steric effect.

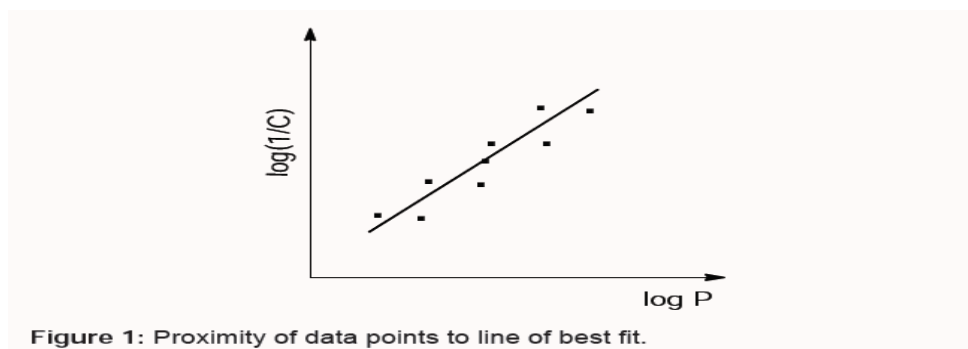
### 4.1. Lipophilic Parameter:

One of the most researched physicochemical properties is lipophilicity. Lipophilicity tests have been there for a long time. In silico lipophilicity technologies that are both reliable and affordable are frequently utilised in drug development. [2,3]

**Partition Coefficient:** The partition coefficient determined lipophilicity of drug and indicate its ability to penetrate cell membrane. It is defined as the ratio between unionized drugs dispersed between the organic and aqueous layers at equilibrium. Drugs having a high partition coefficient can pass through biological membranes. The diffusion of drug molecules over a

rate-controlling membrane or through a matrix system is largely dependent on the partition-coefficient. Drugs with a lower partition coefficient are not highly desirable for oral controlled release formulations and drug with higher partition coefficient are poor for oral controlled formulations. A drug to reach site of action, it pass through a number of biological membrane.  $P$  is parameter used to amount movement of drug through these membranes.

For the type of relationship formed is determined by the substances employed. The range of possible  $P$  values in the case of a short range, regression analysis can be used. A straight line equation can be used to express the results. (Figure 1)



(Fig From: Kumar K et al., 2019)

The linear relationship between partition coefficient of drug and its activity shown by equation;

$$\text{Log}(1/C) = K1 \log P + K2$$

Where,

$C$  is concentration of drug required to produced standard response in given time.

$\log P$  is the logarithm of the molecules partition coefficient between 1-Octanol and water.

$K1$  and  $K2$  are constants.

**Regression analysis:** It is a set of mathematical methods for obtaining mathematical equations that relate various sets of data calculated using theoretical considerations or obtained from experimental work into a suitable computer programme. The relationship between the partition coefficients and activity of a number of related compounds appeared to be linear (Figure 1) these data could be represented in the form of straight line equation ( $y = mx + c$ ). RA (Regression Analysis) calculate the values of  $m$  and  $c$  that gave the line of best fit to the data [3].

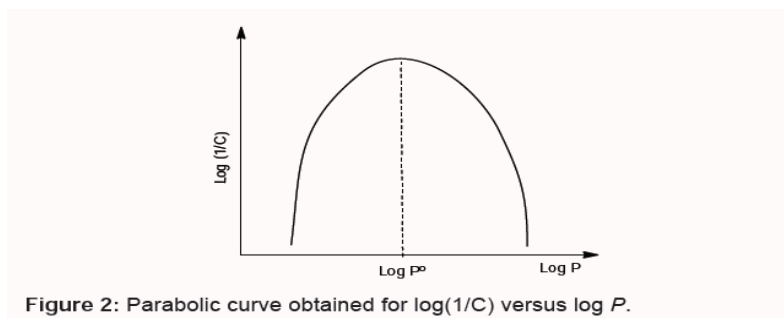


Figure 2: Parabolic curve obtained for  $\log(1/C)$  versus  $\log P$ .

(Fig from: Kumar K et al.,2019)

The graph of  $\log(1/C)$  over greater ranges of  $P$  values (Figure 2) versus  $\log P$  has a maximum value and a parabolic form ( $\log P_0$ ). This occurrence of a maximum value shows that there is an optimal balance between lipid and aqueous solubility for maximum biological activity. The drug will have a problem going through the membrane. Below this, it will be reluctant to leave the membrane, whereas above this, it will be unwilling to leave the membrane. The optimal partition coefficient for biological data is represented by  $\log P_0$  activity.

**Lipophilic substituents constants ( $\pi$ ):** These are also recognized as hydrophobic substituent's constants.

$$\pi = \log P_x - \log P_H$$

$P_x$  and  $P_H$  represent the partition coefficients of a derivative and the parent molecule, respectively. This is a substituent constant that denotes the difference in hydrophobicity between a parent chemical and its substituted analogue. It is frequently replaced with the more general molecular phrase  $\log$  of  $\log K_{ow}$  or  $\log P$ , the 1-octanol/water partition coefficients.

Lipophilic substituent constant can be used as an alternative to the partition coefficient.

**Distribution coefficient:** Lipophilicity is a type of mathematical analysis of pharmacological activity. The value of their distribution coefficients ( $D$ ), which is defined as the ratio of the concentrations of the unionised and ionised compound between an organic solvent and an aqueous medium, is typically used to illustrate the lipophilicity of ionisable compounds. Many chemicals are ionised in the aqueous solution is not taken into consideration in the  $P$  values. The extent to which ionization will also have a substantial impact on absorption and distribution. The distribution of these drugs As a result, in Hansch and other places, E.g., the distribution coefficient of the acid  $HA$  is given by:

$$D = \frac{[HA]_{\text{organic}}}{[H^+]_{\text{(aq)}}/[A^-]_{\text{(aq)}}}$$

Since  $pH$  of the aqueous medium is the deciding factor for the ionization of acids and bases

For acids:  $\log(P/D-1) = pH - pK_a$

For bases;  $\log(P/D-1) = pK_a - pH$

These equations allow the effective lipophilicity of a compound at any  $pH$  to be calculated if the  $pK_a$  and the value of  $P$  for the same solvent system are known.

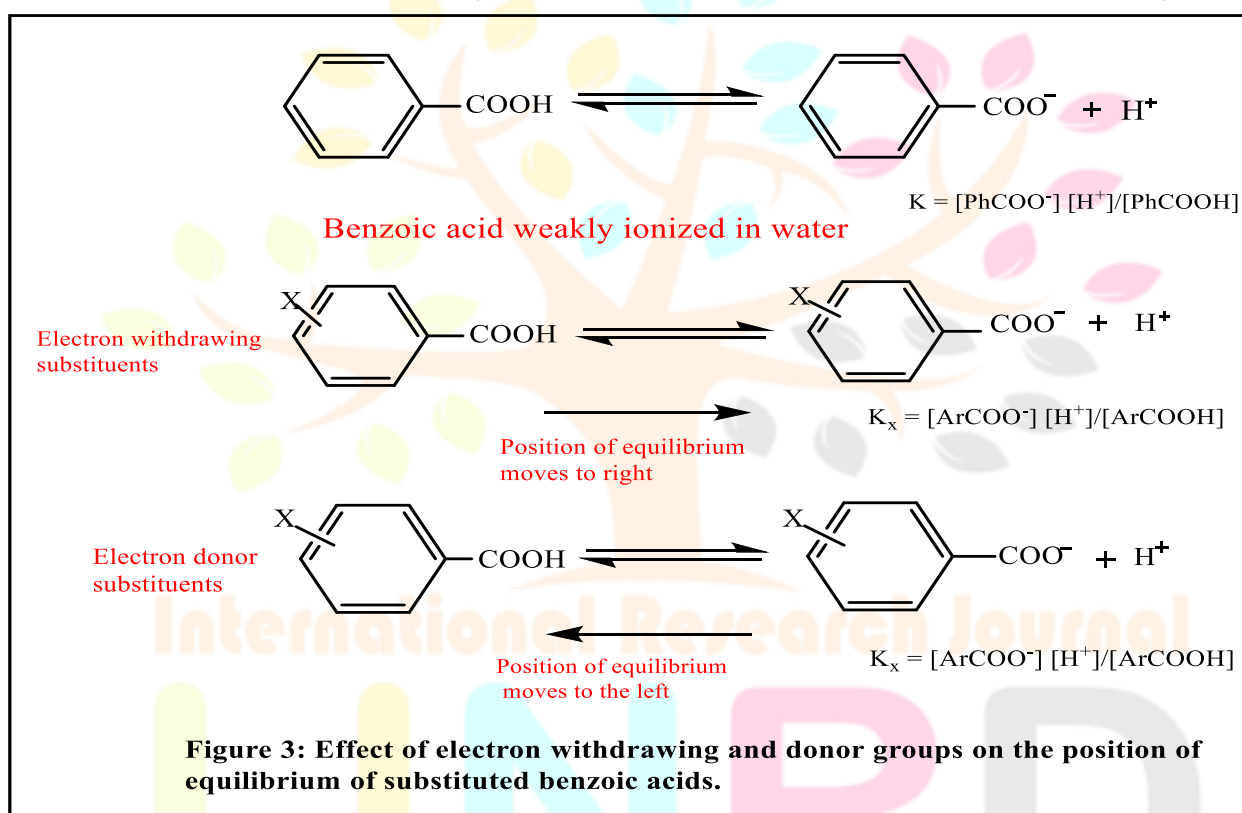
$\log D$  values usually use distribution coefficients.

## 4.2. Electronic Parameter:

The distribution of electrons in a drug molecule will have significant impact on a drug's activity and distribution. A drug normally pass through a number of biological membranes in order to reach its target. Once it gets the site of action, electronic distribution in the drug structure will regulate the type of bonds it makes with the target, which will determine its biological activity.

### The Hammett constant ( $\sigma$ ):

The Hammett equation integrates observed variations in equilibrium or rate constants to systematic changes in the substituents that influence electron donation and withdrawal ability. The electron donating and withdrawing groups characteristics the electronic distribution inside the structure is determined by the molecular structure. Hammett measured the effect of substituents on any reaction by defining an empirical electronic substituent parameter ( $\sigma$ ), which is derived from the acidity constants,  $K_x$  of substituted benzoic acid (Figure 3)



When an electron withdrawing substituent (-X), such as a nitro group, replaces ring hydrogen, it stabilises the carboxylate anion and weakens the O-H bond of carboxyl group. The equilibrium shifts to the right, showing that the substituted molecule is a more stronger acid than benzoic acid ( $K_x > K$ ). In contrast, adding an electron donor substituent (-X) to the ring, such as a methyl group, enhances the acidic OH group while decreasing carboxylate anion stability. This causes the equilibrium move to the chemical is on the left, showing that it is a weaker acid than benzoic acid ( $K > K_x$ ). Hammett used equilibrium constants to study the relationship between the acid strength and structure of aromatic acids. Hammett constants or Hammett substitution constants ( $\sigma_x$ ) are calculated constants for a variety of ring substituents (X) of benzoic acid. Hammett constants ( $\sigma_x$ ) can be defined as:

$$\sigma_x = \log K_x / K$$

$$\text{i.e., } \sigma_x = \log K_x - \log K$$

$$\sigma_x = pK - pK_x \text{ [as } pK_a = -\log K_a]$$

Since  $K \gg K_x$  a negative value for  $\sigma_x$  indicates that the substituent is acting as an electron donor group. In contrast, as  $K < K_x$  a positive value shows that the substituent is acting as an electron withdrawing group. With the position of the substituent in the molecule the value of  $\sigma_x$  varies. Usually, by using the subscripts o, m and p this position is indicated. When a substituent has opposite signs depending on its position on the ring, it signifies it acts as an electron withdrawing group in one case and as an electron donor group in the other. This is possible because the Hammett constant comprises both the mesomeric (resonance) and inductive contributions to the electron distribution.

### 4.3. Steric Factor:

Steric factor is more difficult to measure than the electronic or hydrophobic properties. Several methods are used to determine steric factor are as follows:

#### *Taft's steric factor ( $E_s$ ):*

Taft used the relative rate constants of the acid-catalyzed hydrolysis of substituted methyl ethanoates (Figure 4) to define the steric parameter in 1956. It was discovered that the rates of this hydrolyse were virtually totally determined by steric factors are required. The usage of methyl ethanoate as a solvent. He defined  $E_s$  as a standard; [3]



**Figure 4:  $\alpha$ -substituted methyl ethanoates hydrolysis**

$$E_s = \log K_x - \log K_o$$

Where,

$K_o$  represent the hydrolysis rate of parent ester.

$K_x$  represent the hydrolysis rate of substituted ester

The values for  $E_s$  obtained for a group using the hydrolysis data are applicable to other structures containing that group.

#### *Molar Refractivity (MR):*

It is a measurement of a compound polarisation as well as its volume. The refractive index term is a measure of the polarisability while the  $M/\rho$  term is a measure of the molar volume of the compound. [3]

$$\text{MR} = (n^2 - 1) M / (n^2 + 2) \rho$$

Where, n is the refractive index.

M is the relative mass.

$\rho$  is the density of the compound.

**Verloop steric parameter:**

It involves computer program called sterimol, which estimates steric substituent values (Verloop steric parameters) from the standard Vander waals radii, bond length, bond angles and possible confirmations for the substituent. Unlike  $E_s$ , the Verloop steric parameters can be measured for any substituent. [3,5]

**5. STATISTICAL METHODS USED IN QSAR**

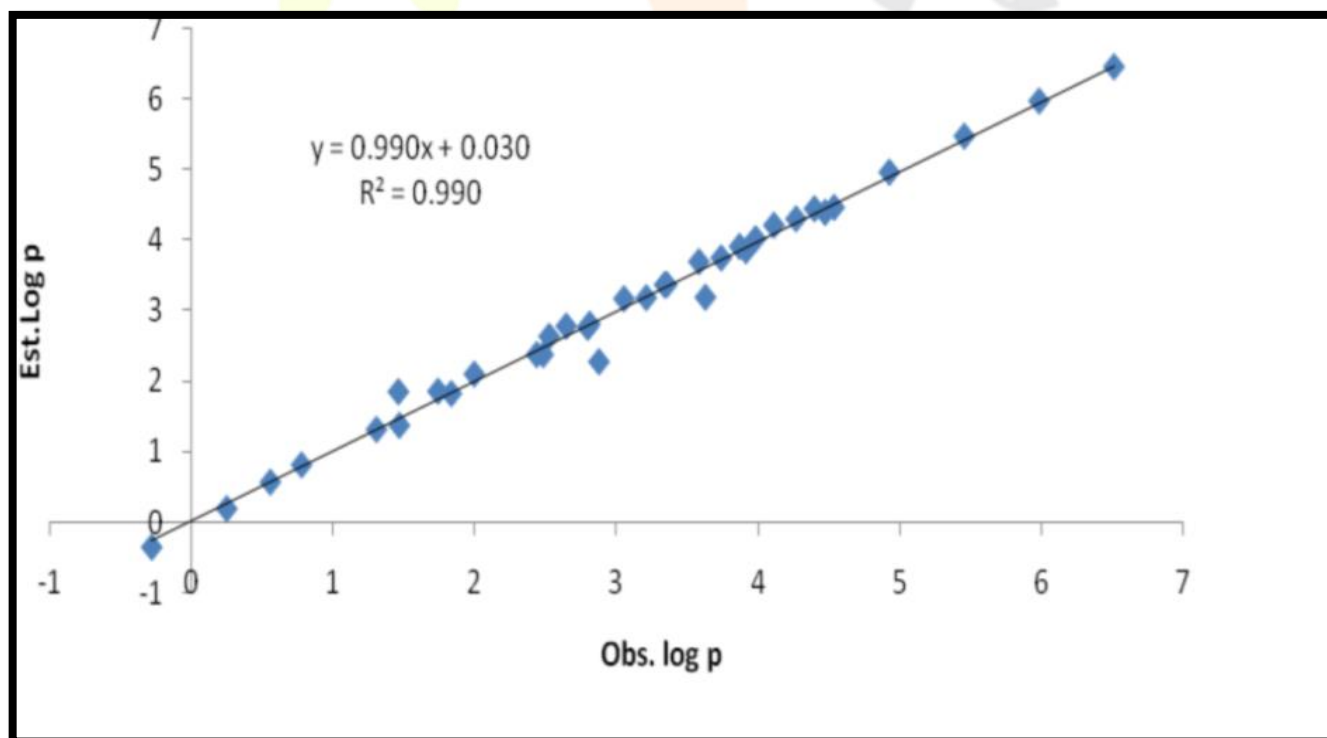
The QSAR statistical methods are classified into two categories depending upon the type of correlation technique used to create a relationship between structural properties and biological activity.

**5.1. Simple Methods**

It consist of multiple linear regression (MLR) and partial least-squares (PLS)

**Multiple Linear Regression (MLR):**

The multiple linear regression approach is used to screen the appropriate descriptor from a big pool of descriptor. In MLR, a linear relationship between the input descriptors and the activity is identified. MLR (multiple linear regressions) is a method for modelling the relationship between two variables by fitting a linear equation to the observed data using two or more explanatory variables and a response variable. The binding affinity and molecular descriptors were correlated using this method. [6,7] Figure 5 shows graphical representation of the observed and calculated activity for any data set using multiple linear regression analysis. (S. Pathan et al, 2016)



**Figure 5: Graphical representation of MLR between observed and calculated biological activity (S. Pathan et al, 2016)**

### ***Partial Least-Squares (PLS):***

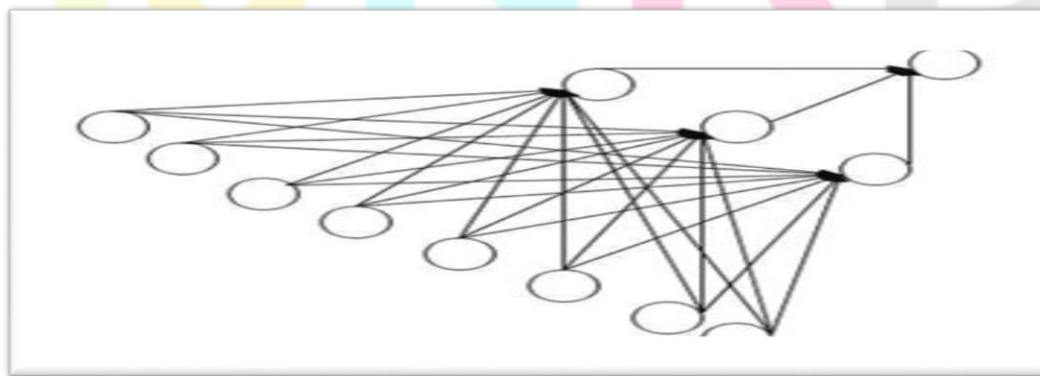
Partial least squares (PLS) can be considered a variation of MLR that transforms the input descriptor and activity space using principal component analysis (PCA) prior to running the linear regression.[8] Because of this, PLS can handle highly correlated input descriptors and is less prone to identifying chance relationships.[9] PLS is a method for building predictive models with a large number of components that are highly collinear. It's employed in a wide range of applied sciences. The conventional algorithm is typically referred to as PLS. The choice of algorithm is determined on the shape of the data matrices. The updating procedure, which employs small updating matrices or an orthogonalization procedure, is one of the computational methods for solving novel algorithms. PLS has been used to monitor and control industrial processes with hundreds of adjustable variables and dozens of outputs. PLS, on the other hand, can be effective when we need to anticipate and there is no practical limit to the number of measured elements. In industrial applications, it's also a popular way for soft modelling.

### **5.2. Non-linear methods**

It include artificial neural networks (ANN). [6,7] and Random Forest Method

#### ***Artificial Neural Network (ANN):***

Artificial neural networks have become increasingly popular in recent years. Quantitative structure–activity relationships (QSAR) between a set of chemical descriptors acquired from the MLR and observed activity can be predicted using neural networks (ANN). It's a generalisation models of biological systems in mathematics The ability to construct is the most significant feature of the ANN using data from experimental measurements of the problem domain to create a model of the problem with the help of in the realm of drug design. Artificial neural networks (ANN) have been used to handle a variety of problems, development of pharmaceutical processes and products. Neural networks are model-free mapping devices that can capture complicated nonlinear relationships in the underlying data that standard QSAR techniques often miss. However, neural networks are known to be unstable, in the sense that tiny changes might cause them to fail. Changes in training data and/or training parameters can have a significant impact on the outcome. The capacity of the generated models to interpret. Fig 6 show schematic representation of ANN.



**Figure 6: Schematic representation of ANN**

***Random Forest Method:***

The Random Forest method is a relatively new machine learning algorithm that has rapidly become the industry standard for generating global statistics models based on QSAR. A Random Forest model is comprised of a large number of independent decision or regression trees (usually 100–500). Bagging is a term used to describe a process. Each tree in the forest is created using a separate bootstrap sample of the training data in this procedure. N compounds have been chosen for the sample substitution for the original dataset. By only evaluating a portion of each tree's descriptors, a second source of randomization is introduced split node, As a result of these two sources of unpredictability, Each tree represents a distinct aspect of the average predictions across the forest based on input data. Trees consistently offer accurate predictions.[10,11,12,13] Random Forests provide useful methods for assessing the relative importance of the input descriptors, and the variance of the predictions across the trees provides excellent estimations of expected prediction errors. [14]

**6. MOLECULAR DESCRIPTORS**

Molecular descriptors convert a compound structure into a set of numerical or binary values that indicate numerous molecular attributes that are considered to be significant for the compound function describing the activity. Based on the requirement on information regarding the molecule's 3D orientation and conformation, two major families of descriptors can be separated.

**6.1. 2D QSAR Descriptors:**

The descriptors used in 2D-QSAR have common property of being independent from the 3D-orientation of compound. The descriptors measure entities constituting the molecule through its topological and geometrical properties to computed, electrostatic and quantum-chemical descriptors or advanced fragment-counting methods.

***Constitutional Descriptors***

Constitutional descriptors describe a molecule's properties about the elements that make up its structure. These descriptors are quick and simple to calculate. Examples of constitutional descriptors include molecular weight, the total number of atoms in the molecule, and numbers of atoms of different identity. A variety of bond characteristics are also considered, such as the total number of single, double, triple, or aromatic type bonds, as well as the aromatic ring.

***Electrostatic and Quantum-Chemical Descriptors***

Electrostatic descriptors collect information about a molecule's electronic nature. Descriptors containing information on atomic net and partial charges. The descriptors with highest negative, positive charge and molecular polarizability are informative. Solvent-accessible either negatively or positively charged atomic surface areas have also been utilized as a source of data. Modeling intermolecular electrostatic descriptors bonding of hydrogen solvent-accessible either negatively or positively charged atomic surface areas have also been employed as a source of data. Energies of highest occupied and the lowest unoccupied molecular orbital form useful quantum chemical descriptors as derivative quantities like absolute hardness.

### ***Topological Descriptors***

Topological descriptors treat the compound's structure as a graph, with atoms acting as vertices and covalent bonds acting as edges. Many indices for measuring molecular connectivity have been developed based on this method, starting with the Wiener index, which counts the total number of bonds in the shortest pathways between all pairs of non-hydrogen atoms. Other topological descriptors contain Randic indices  $x$ , well-defined as sum of geometric averages of edge degrees

of atoms within paths of given lengths, Balaban's J index and Shultz index. Information about valence electrons can be included in topological descriptors, e.g. Kier and Hall indices  $x^v$  or Galvez topological charge indices. The Topological Sub-Structural Molecular Design (TOSS-MODE/TOPS-MODE) [15,16] rely on spectral moments of bond adjacency matrix amended with information on for e.g. bond polarizability. The atom type electrotopological (E-state) indices [17,18] use electronic and topological organization to define the intrinsic atom state and the perturbations of this state induced by other atoms.[19]

### ***Geometrical Descriptors***

Geometrical descriptors are based on the spatial arrangement of the atoms that make up a molecule. These descriptors include molecule surface information obtained from atomic Vander Waals areas and their intersections. Atomic van der Waals volumes can be used to calculate molecular volume. The information about the spatial arrangement of the atoms in a molecule is also captured by principal moments of inertia and gravitational indices. Also used shadow areas obtained by projecting the molecule to its two major axes. The total solvent-accessible surface area is another geometrical descriptor.

### ***Fragment-Based Descriptors and Molecular Fingerprints***

Substructural motifs-based descriptors are frequently employed, especially for quick screening of very large databases. Bits are used to create BCI fingerprints describing the presence or absence of particular elements in a molecule fragments, including atoms and their immediate surroundings ring-based fragments, atom pairs and sequences. The basic set of 166 MDL Keys uses a similar method. Other MDL Keys variations, however, are also accessible including extended or compact sets of keys. The latter are the product of specialised pruning methods or elimination procedures, such as FRED/SKEYS (Fast Random Elimination of Descriptors/Substructure Keys). The Hologram QSAR (HQSAR) technique, which was recently presented, is based on counting the number of occurrences of certain substructural routes of functional groups. By removing the reliance on a pre-defined list of sub-structure motifs.

Daylight fingerprints are a natural extension of fragment-based descriptors. Each molecule's fingerprint is a series of bits. However, a structural motif in the molecule does not equate to a single bit, but rather to a sequence of bits that is added to the fingerprint using a logical "or" operation according to a hashing function. Because of the enormous number of conceivable patterns and the finite length of a bit string, bits in distinct patterns may overlap. As a result, the existence of a bit or multiple bits in a fingerprint cannot be taken as confirmation of the pattern's presence. If one of the bits corresponding to a certain pattern is not set, the pattern is

guaranteed to be absent from the molecule. This enables for the quick identification of molecules that lack particular structural patterns.

## 6.2. 3D QSAR Descriptors:

The 3D-QSAR method is substantially more computationally difficult than the 2D-QSAR method. In general, obtaining numerical descriptors of the complex structure includes many steps. The conformation of the chemical must first be determined using either experimental data or molecular mechanics, and then modified via minimising energy. The conformers in the dataset must then be aligned in space evenly. Finally, several descriptors are computed for the space with submerged conformer. There have also been some approaches created that independent on compound alignment.

### 6.2.1. Alignment-Dependent 3D QSAR Descriptors

The set of approaches that need molecular alignment prior to descriptor calculation is totally dependent on receptor knowledge for the modelled ligand. The alignment can be directed by analysing the receptor-ligand complexes if such data is available. Otherwise, for superimposing the structures in space, only computational approaches must be applied.

#### *Comparative Molecular Field Analysis:*

The electrostatic (Coulombic) and steric (van der Waals) energy fields defined by the investigated chemical are used in Comparative Molecular Field Analysis (CoMFA). The aligned molecule is then placed on a three-dimensional grid. A probe atom with unit charge is placed in each position of the grid lattice, and the energy field potentials (Coulomb and Lennard-Jones) are determined. They are then used as descriptors in subsequent analysis, which is usually done with partial least squares regression. This analysis identifies structure regions that are favourably and adversely connected to the activity at hand.

#### *Comparative Molecular Similarity Indices Analysis*

In the aspect of atom probing throughout the regular grid lattice in which the molecules are embedded, the Comparative Molecular Similarity Indices (CoMSIA) is similar to CoMFA. The probe atom's resemblance to the studied molecule is estimated. CoMSIA, in contrast to CoMFA, employs a different potential function, the Gaussian-type function. The probing atom's steric, electrostatic, and hydrophobic characteristics are then calculated, resulting in a property of unit hydrophobicity. The use of a Gaussian-type potential function rather than Lennard-Jones or Coulombic functions allows for more precise information in grid locations within the molecule. Due to the nature of the potential functions and the arbitrary cut-offs that must be used in CoMFA, unacceptably large values are achieved in these points.

### 6.2.2. Alignment-Independent 3D QSAR Descriptors

The descriptors that are invariant to molecular rotation and translation in space are another type of 3D descriptor. As a result, no compound superposition is necessary.

#### *Comparative Molecular Moment Analysis*

Second-order moments of the mass and charge distributions are used in Comparative Molecular Moment Analysis (CoMMA). The moments are related to the mass centre and the dipole centre.

Principal moments of inertia, magnitudes of dipole moments, and principal quadrupole moments are among the CoMMA descriptors.

Descriptors connecting charge to mass distributions are also defined, such as the magnitudes of dipole projections upon primary moments of inertia and the distance between the centre of mass and the centre of dipole.

### ***Weighted Holistic Invariant Molecular Descriptors***

The invariant information is provided by the Weighted Holistic Invariant Molecular (WHIM) and Molecular Surface WHIM descriptors, which use principal component analysis (PCA) on the centred co-ordinates of the atoms that make up the molecule. As a result, the molecule is transformed into the space that captures the maximum variation. Several statistics, including as variance, proportions, symmetry, and kurtosis, are calculated and used as directional descriptors in this space. Non-directional descriptors are created by combining the directional descriptors. A chemical property can weight each atom's contribution, resulting in various principal components representing variance within the property. Mass, van der Waals volume, atomic electronegativity, atomic polarizability, Kier and Hall electrotopological index, and molecule electrostatic potential can all be used to weigh the atoms.

### ***VolSurf***

The VolSurf method relies on particular probes investigating the grid around the molecule, such as hydrophobic interactions or hydrogen bond acceptor or donor groups. The descriptors based on volumes or surfaces of 3D contours, determined by the same value of the probe molecule interaction energy, are computed using the lattice boxes that result. Different molecular characteristics can be quantified using various probes and energy cut-off values. Molecular volume and surface, as well as hydrophobic and hydrophilic regions, are examples. It is also possible to compute derivative values such as molecular globularity or factors linking the surface of hydrophobic or hydrophilic regions to the surface of the entire molecule.

### ***Grid-Independent Descriptors***

The Grid-Independent Descriptors (GRIND) were created to address the interpretability issues that plague alignment-independent descriptors. It works in a similar way as VolSurf in that it probes the grid with specialised probes. The locations with the highest favourable interaction energies are chosen, assuming that the distances between them are considerable. The probe-based energy are then represented in a form that is independent of the molecule's configuration. The distances between the nodes in the grid are discretized into a series of bins to achieve this. The nodes with the highest product of energies are kept for each distance bin, and the value of the product serves as the numerical descriptor.

## **7. APPLICATION OF QSAR**

Qualitative Structure Activity Relationship (QSAR) main application in drug design and medicinal chemistry are as follows:[7]

- I. To rationalize the development of novel lead compounds that have higher biological activity.

- II. Before synthesis, identify the hazardous compounds and toxicity of the therapeutic molecule. The toxicity of environmental species and other biological systems will be reduced as a result of this.
- III. Pharmacological and pesticidal activity optimization
- IV. The process of identifying and selecting a molecule in order to achieve the best biological responses and pharmacokinetic properties.
- V. To determine the role of numerous qualities in the creation of a therapeutic molecule and to determine which properties are better for improving biological activity.

## 8. LIMITATION OF QSAR

QSAR include clearly-defined physio-chemical descriptors and are best suited for the analysis of large number of compounds and computational screening of molecular databases. But still the day-by-day challenges in the field of drug design shows that this one also has some limitations.[7]

- I. Biomolecules are mostly found in elaborate three-dimensional forms, whereas traditional QSAR exclusively deals with two-dimensional structures.
- II. Only a small number of descriptors are taken into account when employing 2D descriptors, which is a shortcoming of the old method.
- III. Regardless of their abundance, there is no representation of stereochemistry or 3D-structure of molecules.
- IV. Because the resultant model lacks predictability, synthesis on behalf of the 2D model is difficult.
- V. The random association, rather than the actual prediction, is better in 2D QSAR models.
- VI. Considering the standard QSAR equation does not directly suggest new compounds to synthesis, it takes a lot of knowledge of substituent constants in physical organic chemistry to build a molecule

## 9. CONCLUSION

Qualitative structure-activity relationship (QSAR) are the technique used in drug design and medicinal chemistry. By using QSAR, we can determine the toxicity and biological activity of chemical compound. There are so many advancement in QSAR that is depend on the descriptor use and their physiochemical property. 3D QSAR are widely used now a days than the QSAR and 2D QSAR.

## 10. REFERENCES

1. Penniston, J. T., Beckett, L., Bentley, D. L., and Hansch, C. (1969) Passive permeation of organic compounds through biological tissue: a nonsteady-state theory. *Mol. Pharmacol.* 5, 333–336.
2. Ankur Vaidya, Sourabh Jain, Shweta Jain, Abhishek Jain and Ram Agrawal (2014), 'Quantitative structure activity relationship: A novel approach of drug design and discovery', JPSP American Scientific Publishers, Vol 1, 219-232 doi:10.1166/jpsp.2014.1024
3. Kapoor Y., Kapoor K., (2019) 'Quantitative structure activity relationship in drug design: An overview', 2<sup>nd</sup> Edition, SF Journal Of Pharmaceutical and Analytical Chemistry, Vol 2, Article 1017.

4. Hansch C., Leo A., Hoekman D., Exploring QSAR. Fundamentals and Applications in Chemistry and Biology, Volume 1. Hydrophobic, Electronic and Steric Constants, Volume 2. American Chemical Society.1995.
5. Tipker J, Verloop A. Use of STERIMOL, MTD, and MTD\* Steric Parameters in Quantitative Structure-Activity Relationships. American Chemical Society. 1984; 255: 279-296.
6. Ojha L.K., Chaturvedi A.M., Bhardwaj A., Thakur M. and Thakur A. Asian Journal of Research in Chemistry. 5(3), 377-382, (2012).
7. Ojha L.K., Sharma R., and Bhawsar M.R., (2013) Modern drug design with advancement in QSAR: A review, ' International Journal of Research in BioSciences' Vol. 2
8. Geladi P, Kowalski BR. Partial least-squares regression: a tutorial. *Anal Chim Acta* 1986, 185:1–17.
9. Clark M, Cramer RD. The probability of chance correlation using partial least-squares (Pls). *Quant Struct-Act Rel* 1993, 12:137–145.
10. Richard A.L and David W. Modern 2D QSAR for drug discovery, WIREs Comput Mol Sci 2014. doi: 10.1002/wcms.1187
11. Breiman L. Random forests. *Mach Learn* 2001, 45:5–32
12. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* 2003, 43:1947–1958
13. Breiman L. Bagging predictors. *Mach Learn* 1996, 24:123–140.
14. Wood DJ, Carlsson L, Eklund M, Norinder U, Stalring J. QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality. *J Comput Aided Mol Des* 2013, 27:203–219.
15. Estrada, E. *J. Chem. Info. Comput. Sci.*, **1996**, 36, 844-849.
16. Estrada, E.; Uriarte, E. *SAR QSAR Environ. Res.*, **2001**, 12, 309-324
17. Hall, L.H.; Kier, L.B. *Quant. Struct.-Act. Relat.*, **1991**, 10, 43-48
18. Hall, L.H.; Kier, L.B. *J. Chem. Inf. Comput. Sci.*, **2000**, 30, 784-791.
19. Arkadiusz Z.D., Tomasz A and Jorge G., Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review, 'Combinatorial Chemistry & High Throughput Screening', Bentham Science Publishers Ltd. 2006, 9, 213-228

Research Through Innovation