Chapter - 1

General Introduction

The formulation of a problem is far more often essential than its solution, which may be merely a matter of mathematical or experimental skill.

—Albert Einstein
1.1. Introduction

1.1.1. QSARs in Drug Discovery

Quantitative structure–activity relationship (QSAR), in simplest terms, is a method for building computational, virtual, atomistic or mathematical models to establish a reliable and statistically significant correlation between structure and function using a chemometric technique. In terms of drug design, structure here refers to the properties or descriptors of the molecules, their substituents or interaction energy fields, function corresponds to an experimental or biochemical endpoint like rate constant, toxicity, binding affinity, or biological activity towards a known or hypothetical macromolecular target, while chemometric method include MLR, PLS, PCA, PCR, ANN, GA etc. The term ‘quantitative structure–property relationship’ (QSPR) is used when some property other than the biological activity is concerned. The fundamental principle underlying the QSAR formalism is that the difference in structural properties is responsible for the variations in biological activities of the compounds.

1.1.2. Objectives of QSAR

Mostly all the QSAR methods focus on the following goals:

To quantitatively correlate and recapitulate the relationships between trends in chemical structure alterations and respective changes in biological endpoint for comprehending which chemical properties are most likely determinants for their biological activities

To optimize the existing leads so as to improve their biological activities

To predict the biological activities of untested and sometimes yet unavailable compounds

1.1.3. Rationale behind QSAR modelling

The extent of reliability in opting for QSAR modelling depends on the type or nature of property being predicted, the stage of the project and the relative ease and cost of compound synthesis and subsequent testing. More often QSAR models provide useful predictions but many times they fail, despite of good statistics generated from internal data used in training. Regardless of all such problems, QSAR becomes a useful alternative because of the following reasons:
• Conventional syntheses methods are expensive and time-consuming
• Biological assays are also too costly, often requiring time, sacrifice of animals, or compounds in their pure forms
• Drug failures due to poor ADMET profiles at later stages of development (or even after commercialization) are exceedingly expensive and painful
• Large number of compounds are available due to combinatorial chemistry and HTS approaches, but estimations are required for prioritization of synthesis and screening

1.2. Evolution of QSAR

1.2.1. Classical QSAR methodologies

QSAR methods originated way back in the nineteenth century. Since then various approaches have been developed gradually and served as a valuable predictive tool, particularly in the design of pharmaceuticals and agrochemicals. The methods have evolved from Hansch and Free-Wilson’s one or two-dimensional linear free-energy relationships, via Crammer’s three-dimensional QSAR to Hopfinger’s fourth and Vedani’s fifth and sixth-dimensions. All the earlier methods developed by Hammett, Fujita, Hansch and Free-Wilson were based on 2D representations of the molecules and are commonly referred to as ‘classical’ QSAR methodologies. In these traditional studies, affinities of ligands to their binding sites, inhibition constants, rate constants, and other biological end points have been correlated with atomic, group or molecular properties related to a structural pattern such as lipophilicity, polarizability, connectivity indices (e.g., Randic indices, Kier and Hall indices), fingerprints, electronic and steric properties (Hansch analysis) or with certain structural features (Free-Wilson analysis). The chronological order in which these classical or traditional QSAR methods have evolved over a period of time is being given in Table 1.01.

Table 1.01. A brief history of earlier QSAR methodologies

<table>
<thead>
<tr>
<th>Author (Year)</th>
<th>Contributions/Postulates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crum-Brown and Fraser</td>
<td>Physiological activities of substances can be correlated with their chemical composition and constitution. But they did not show how to represent the chemical structure in a quantitative manner¹</td>
</tr>
<tr>
<td>(1868)</td>
<td></td>
</tr>
<tr>
<td>Richardson (1868)</td>
<td>Expressed the chemical structure as a function of solubility²</td>
</tr>
<tr>
<td>Mills (1884)</td>
<td>Developed a QSPR model for the prediction of melting and boiling points in homologous series³</td>
</tr>
<tr>
<td>Richet (1893)</td>
<td>Correlated toxicities of a set of alcohols, ethers and ketones with aqueous solubility and showed that their cytotoxicities are inversely</td>
</tr>
<tr>
<td>Author(s)</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Overton and Meyer (1897, 1899)</td>
<td>Correlated partition coefficients of a group of organic compounds with their anaesthetic potencies and concluded that narcotic (depressant) activity is dependent on their lipophilicity.</td>
</tr>
<tr>
<td>Hammett (1935, 1937)</td>
<td>Correlated the effect of substituent addition on benzoic acid with the dissociation constant, postulated electronic sigma-rho constants and established linear free-energy relationship (LFER) principle.</td>
</tr>
<tr>
<td>Ferguson (1939)</td>
<td>Correlated depressant action with the relative saturation of volatile compounds in their vehicle and introduced a thermodynamic generalization to the toxicity.</td>
</tr>
<tr>
<td>Bell and Roblin (1942)</td>
<td>Studied antibacterial activities of a series of sulfanilamides in terms of their ionizations.</td>
</tr>
<tr>
<td>Albert (1948)</td>
<td>Examined the effects of ionization/electron distribution and steric access on the potencies of a multitude of aminoacridines.</td>
</tr>
<tr>
<td>Taft (1952)</td>
<td>Postulated a method for separating polar, steric, and resonance effects and introduced the first steric parameter, Eₜ.</td>
</tr>
<tr>
<td>Hansch and Muir (1962)</td>
<td>Correlated the biological activities of plant growth regulators with Hammett constants and hydrophobicity.</td>
</tr>
<tr>
<td>Hansch and Fujita (1964)</td>
<td>Combined the hydrophobic constants with Hammett's electronic constants to yield linear Hansch equation and its many extended forms.</td>
</tr>
<tr>
<td>Free and Wilson (1964)</td>
<td>Formulated an additive model, where the activity is discretized as a simple sum of contributions from different substituents.</td>
</tr>
<tr>
<td>Hansch (1969)</td>
<td>Developed parabolic Hansch equation for dealing with extended hydrophobicity ranges.</td>
</tr>
<tr>
<td>Fujita and Ban (1971)</td>
<td>Simplified the Free-Wilson equation estimating the activity for the non-substituted compound of the series and postulated Fujita-Ban equation that used the logarithm of activity, which brought the activity parameter in line with other free energy-related terms.</td>
</tr>
<tr>
<td>Kubinyi (1976)</td>
<td>Investigated the transport of drugs via aqueous and lipoidal compartment systems and further refined the parabolic equation of Hansch to develop a superior bilinear (non-linear) QSAR model.</td>
</tr>
<tr>
<td>Hansch and Gao (1997)</td>
<td>Developed Comparative QSAR, as implemented in C-QSAR program.</td>
</tr>
<tr>
<td>Hurst and Heritage (1997)</td>
<td>Developed Hologram QSAR (HQSR), where the structures are converted into all possible fragments, which are assigned specific integers, and then hashed into a fingerprint to form molecular hologram. The bin occupancies of these holograms are used as QSAR descriptors, encoding the chemical and topological information of molecules.</td>
</tr>
<tr>
<td>Cho and workers (1998)</td>
<td>Developed Inverse QSAR, which seeks to find values for the molecular descriptors that possess a desired activity/property value. In other words, it consists of finding the optimum sets of descriptor values best matching a target activity and then generating a focused library of candidate structures from the solution set of descriptor values.</td>
</tr>
<tr>
<td>Labute (1999)</td>
<td>Developed Binary QSAR to handle binary activity measurements from high-throughput screening (e.g., pass/fail or active/inactive), and molecular descriptor vectors as input. A probability distribution for actives and inactives is then determined, based on Bayes' Theorem.</td>
</tr>
</tbody>
</table>
1.2.2. **Limitations of the classical QSAR methodologies**

Classical QSAR methods are much simpler, faster and more amenable to automation than 3D-QSAR approaches. They include clearly-defined physiochemical descriptors and are best suited for the analysis of large number of compounds and computational screening of molecular databases. Though they have been used for decades to correlate and predict the activity of molecules, they suffer from serious limitations in certain situations\(^{24}\) some of which are as follows:

- Only 2D-structures considered
- Unavailability of appropriate physiochemical parameter (\(e.g.,\) numerical descriptors for new or unusual substituents), rendering the compound unfit for inclusion in QSAR analysis
- Insufficient parameters for describing drug-receptor interactions (\(e.g.,\) steric parameter \(E_s\), Hammett constant \(\sigma\) etc.)
- Confined to only few substitutions in a common reference structure (simple variation of aromatic substituents) and works best with a congeneric series
- No representation of stereochemistry or 3D-structure of molecules, regardless of their availability
- Provide no unique solutions
- Higher risk of chance correlations
- High risk of failure due to 'too far outside' predictions
- No graphical output thereby making the interpretation of results in familiar chemical terms, frequently difficult if not impossible
- Requires considerable knowledge of substituent constants in physical organic chemistry to design a molecule, since the equation do not directly suggest new compounds to synthesize

1.3. **The Multidimensionality concept: \(n\)D-QSAR**

The various QSAR methods are often categorized into following classes according to their dimensionality \(i.e.,\) the structural representation in which the descriptor values are derived:

1D-QSAR correlating activity with global molecular properties like \(pK_a\), \(\log P\) etc.

2D-QSAR correlating activity with structural patterns like connectivity indices, 2D-pharmacophores etc., without taking into account the 3D-representation of these properties

3D-QSAR correlating activity with non-covalent interaction fields surrounding the molecules
4D-QSAR additionally including ensemble of ligand configurations in 3D-QSAR
5D-QSAR explicitly representing different induced-fit models in 4D-QSAR
6D-QSAR further incorporating different solvation models in 5D-QSAR

Of particular interest for the biomedical sciences are the 3D-QSAR techniques, which have emerged as a natural extension to the classical Hansch and Free-Wilson approaches and correlate macroscopic target properties with computed atom-based descriptors derived from the spatial (three-dimensional) representation of the molecular structures\textsuperscript{25-29}. A majority of these methods are ligand-based and depend on the calculation of receptor interactions indirectly using probes positioned at intersections of a lattice straddling a three-dimensional region resembling a binding site surrogate. However, the major drawback of 3D-QSAR techniques is that they all are based on various assumptions\textsuperscript{29} which are as follows:

- There is an underlying relationship between molecular structure and biological activity.
- Receptor binding is directly proportional to the biological activity. Differential effects on second messengers or other signaling steps which transpire between receptor binding and experimentally observed response, are not taken into consideration.
- Molecular structure can be measured and represented with a set of numbers usually called descriptors, which encode all physical, chemical and biological properties of the molecule.
- Molecules with common or related structures generally have similar physicochemical properties (the similarity principle), and thus have similar binding modes and consequently comparable biological activities. The reverse also holds true. Also, molecules located in the same region of the descriptor space present similar activity (the neighbourhood principle).
- Structural properties which lead to an observed biological response are most commonly determined by the non-bonding (or non-covalent) forces, mainly steric and electrostatic.
- The observed biological effect is produced by the modeled ligand itself, and not by its metabolite or degradation product.
- The lowest energy conformation of the ligand is its bioactive conformation, and it is this single conformation of the ligand which exerts the binding effects.
- With few exceptions, the geometry of the receptor binding site is considered rigid.
- The loss of translational and rotational degrees of freedom (entropy) upon binding is believed to follow a similar pattern for all the molecules.
- Total number of rotatable bonds is the only method most frequently used to estimate the entropic cost for freezing non-terminal single-bond rotors.
For all the modeled ligands, the protein binding site is assumed to be same.

For all the modeled compounds, the on-off rate is supposed to be similar *i.e.*, the system is considered to be in equilibrium, and kinetic aspects are usually ignored.

Some of the major factors like desolvation energetics, temperature, diffusion, transport, pH, salt concentration *etc.*, which contribute to the overall free energy of binding are difficult to handle, and thus usually ignored.

Resulting QSAR model may represent one of potentially several solutions to the property–activity correlation problem.

In molecular mechanics based 3D-QSAR methods, free energy of binding is largely explained by the enthalpic component (*i.e.*, the internal energy), which is prone to the inherent force field errors.

In order to overcome some of these shortcomings, multidimensional QSARs have emerged as an un-restrained extension of the 3D-QSAR methods$^{30}$. In this arena, Hopfinger and workers for the first time incorporated receptor information in a QSAR analysis, in the form of an ensemble of all possible conformations, orientations and, in some cases, protonation states, to devise the 4D-QSAR methodology.$^{31, 32}$ Similarly, Vedani *et al.* have developed methods beyond the third dimension by accounting for the effect of different conformations as the fourth dimension$^{33}$, multiple representations of induced-fit scenarios as the fifth dimension$^{34}$, and assessment of different solvation models as the sixth dimension$^{35}$, additionally incorporating contributions from the solvent and entropy factors into the analysis.

1.4. Dimensions Explored

The QSAR models have become increasingly sophisticated over the last few years, adding numerous extra "dimensions". Most of the 3D-QSAR methods use some form of the molecular interaction fields/energies as their primary dimensions or descriptors, however there are some which are based on other unique dimensions as stated in Table 1.02.

**Table 1.02. Different QSAR methodologies and the dimensions investigated by them**

<table>
<thead>
<tr>
<th>QSAR Methods</th>
<th>Dimensions</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoMFA</td>
<td>Grid-based steric (with sp$^3$ C-atom using 6-12 Lennard-Jones function) and electrostatic (with H$^+$ probe using Coulomb’s law) fields$^{36}$</td>
</tr>
<tr>
<td>MSA</td>
<td>Molecular shape similarity measures$^{37}$</td>
</tr>
<tr>
<td>GRID</td>
<td>Steric (with sp$^3$ C-atom using 6-4 Lennard-Jones function), electrostatic (with H$^+$ probe using Coulomb’s law), hydrophobic (with DRY/water probes), and Hydrogen-</td>
</tr>
</tbody>
</table>
### The following section describes in detail the manner in which the afore-mentioned dimensions have been derived and processed in their respective methodologies to generate useful QSAR models:

#### 1.4.01. CoMFA

In 1987, Cramer developed one of the most powerful 3D-QSAR methodology, Comparative Molecular Field Analysis (CoMFA)\(^\text{36}\), by incorporating GRID and PLS techniques into DYLOMMS (DYnamic Lattice-Oriented Molecular Modelling System) method\(^\text{36}\), that involves the use of PCA to extract vectors from the molecular interaction fields. A standard CoMFA procedure, as implemented in the Sybyl Software from Tripos Inc.\(^\text{57}\), proceeds by superimposing all the molecules in their probable bioactive conformations (i.e., their supposed mode of interaction with the receptor), and positioning the overlaid set of molecules in the center of a lattice grid with a spacing of 2 Å. This is followed by the computation of
molecular interaction fields as dimensions or descriptors. The most commonly employed fields in CoMFA are steric (van der Waals) and electrostatic (Coulombic) interactions, which are mainly enthalpic in nature. However, many times the entropic effects, in the form of hydrophobic interactions, are also included in the CoMFA analysis. Creativity of the research and the validity of the underlying theory are the major parameters deciding the type of field to be generated and included in a CoMFA model.

The interaction energies are calculated using different probes. The electrostatic energies are computed with H⁺ probe by employing the Coulomb's law, whereas a sp³ hybridized carbon atom with an effective radius of 1.53 Å and a +1.0 charge is used as the probe for calculating the steric energies using the standard (6-12) Lennard-Jones function. The slope of the Lennard-Jones potentials is very steep close to the van der Waals surface, as a result of which the potential energy at lattice points in the proximity of the surface changes significantly. This implies that a trivial difference in the mutual shift or conformational changes of the compounds may result in very large differences in energy values. Moreover, the Lennard-Jones and Coulombic potentials show singularities (unacceptably large values) at the atomic positions. Therefore to avoid all these problems in CoMFA, the cut-off values (± 30 kcal/mol) for steric and electrostatic energies are defined.

Each probe is positioned in turn at every intersection point of the lattice, and the interaction energies between the probe and each of the compounds are calculated using different molecular force fields. The interaction energies or field values are than correlated with the biological activity data using PLS technique, which identifies and extracts the quantitative influence of specific chemical features of molecules on their biological activity. The results are articulated as correlation equations with the number of latent variable terms, each of which is a linear combination of original independent lattice descriptors. For visual understanding, the PLS output is presented in the form of an interactive graphics consisting of coloured contour plots of coefficients of the corresponding field variables at each lattice intersection, and showing the imperative favourable and unfavourable regions in three-dimensional space which are considerably associated with the biological activity.
1.4.02. MSA

Molecular Shape Analysis (MSA) is a ligand-based 3D-QSAR formalism which attempts to merge conformational analysis with the classical Hansch approach. It deals with the quantitative characterization, representation and manipulation of the molecular shape, as a new dimension in the construction of a QSAR model. The methodology begins by subjecting each molecule in the data set to a fixed valence geometry intramolecular conformational analysis with a scan at 30° increments for all torsional angles except for amide N-(C=O) torsion which is scanned at 180° increment. The conformational energies are estimated using a fixed valence geometry molecular mechanics force-field consisting of dispersion/steric, electrostatic, and, if applicable, hydrogen bonding contributions. For each compound, all apparent intramolecular energy minima are identified and recorded, each of which are then used as starting points in rigorous fixed valence geometry energy minimizations. Both apparently as well as rigorously minimized energy conformations are aspirants for the ‘active’ conformation of each analog in the ensuing steps. To identify the active conformation of each analog, the LBA-LCS (loss in biological activity-loss in conformational stability) approach is used; this is based on the identification of stable low-energy intramolecular conformer states common to the active analogs, which is a high-energy, unstable state for the inactive analogs.

A shape reference structure is selected as the mutant shape generated by the common and difference volume combinations realized by multiple compound alignments or active conformations. The potential active conformation of each compound in the data set is pairwise compared and aligned with the shape reference. This is followed by the calculation of various descriptors which measure relative molecular shape similarity. One of the important shape dimensions is the common overlap steric volume (COSV) between pairs of molecules as a function of conformation and relative intermolecular geometry. It actually measures how much steric space a pair of molecules share under a prescribed intermolecular relationship. Two other descriptors are also arbitrarily defined as alternate mathematical representations of COSV that can advantageously be used in developing empirical QSARs; one has the dimensions of area but is not a physical measure of common atomic surface areas between two molecules, and another has the dimensions of length but is not a cumulative measure of distances between the molecules. These pair-wise shape variables can also be amalgamated with the non-shape thermodynamic and electronic descriptors including the terms from the Hansch equation \( (\pi, E_a, \sigma) \) in developing a MSA 3D-QSAR model.
The shape similarity descriptors along with the non-shape variables are eventually correlated with the biological activities of the molecules using the MLR technique, however, other chemometric methods like PLS and GA can also be employed. The MSA results can be graphically represented as a picture of the most active analog placed in its active conformation or as the superimposition of shape descriptors onto the molecular geometry of the most active molecule.

1.4.03. GRID

GRID was the first tailor-made program designed for the medicinal chemist as an alternative to the original CoMFA approach. It calculates and uses the interaction energy fields as dimensions in molecular field analysis and determines the energetically favourable binding sites on molecules of known structure. Though the approach is similar to CoMFA in that it too computes explicit non-bonded (or non-covalent) interactions between a molecule of known three-dimensional structure and a probe (i.e., a small chemical group with user-defined properties) located at the sample positions on a lattice throughout and around the macromolecule, it offers two distinct advantages; first is the use of a 6-4 potential function for calculating the interaction energies, which is smoother than the 6-12 form of the Lennard-Jones type in CoMFA, and second is, the availability of different types of probes. The program in addition of computing the regular steric and electrostatic potentials, also calculates the hydrogen bonding potential using a hydrogen bond donor and acceptor, and the hydrophobic potential using a “DRY probe”. Later on a water probe was added to calculate hydrophobic interactions. Since the water probe is not only electrically neutral but can also donate and accept a hydrogen bond, the energies determined using this probe are supposed to embrace steric and hydrogen bonding interactions also, besides representing the hydrophobic interaction energy like logP due to its molecular surface area. In addition to the water and DRY probes, other probes which are usually employed singly, include the methyl group, the amine NH₂ group, the carboxylate group and the hydroxyl group. Contour surfaces are calculated at various energy levels for each probe for every point on the grid and are displayed graphically along with the protein structure. While negative energy levels of the contours describe regions at which ligand binding should be favoured, positive energy levels normally characterize the shape of the target.
1.4.04. HASL

The Hypothetical Active Site Lattice (HASL) method is an inverse grid-based approach which represents the shapes of the molecules inside an active site as a collection of grid-points\textsuperscript{39}. The methodology begins with the intermediate conversion of the Cartesian coordinates \((x, y, z)\) of a superimposed set of molecules to a 3D-grid consisting of the regularly-spaced points that are:

- arranged orthogonally to each other
- separated by a particular distance termed as the resolution (which determines the number of grid points representing a molecule)
- all sprawl within the van der Waals radii of the atoms constituting the molecule

The resulting framework of points is referred to as the molecular lattice and represents the receptor active site map. The overall lattice dimensions are dependent on the size of the molecules and the resolution chosen. Typically a reference molecule is selected arbitrarily and its user-defined conformations similar in shape and that have been energy-minimized, are used to generate the HASL. The selected conformation of the reference molecule is centered about the origin of a Cartesian coordinate system, and a regular grid with a chosen resolution is then laid over the molecule. All grid points lying within the van der Waals radii of the atoms of the molecule are designated as 'occupied' and form the molecular lattice. The electronic properties of the occupying atoms are distinctly represented by assigning the lattice points a 'HASL-type' value based on the electron density of the atoms, which constitute the fourth dimension of the molecular lattice. The values of +1, -1 and 0 are assigned to the electron-rich (e.g., O, N), electron-poor (e.g., C in C=O) and neutral atoms/substituents, which roughly represent H-bond acceptors, donors, and lipophilic atom types, respectively. Such internal atom type designations allow apparently different structures to be overlaid with an equivalent electronic "sense", (i.e., similar atomic characteristics of two molecules can be superimposed in 3D-space), to obtain a maximum complementarity of space and physiochemical character within that space. Similar to electron density, other user-selected physiochemical properties such as hydrophobicity can also be employed as the fourth dimension. A second molecule is subsequently selected and subjected to the same routine, generating its 4D lattice which is then compared and consequently aligned to that of the stationary reference molecule. In order to optimally align the molecules based on their lattices, a systematic search is performed which involves a stepped progression of
translational and rotational movements, with an intermediate lattice generated at each step, until a perfect match is obtained. The extent of similarity between the two molecules is calculated according to a fitting function, based upon the degree of correspondence between the points of the two lattices, i.e., on the number of points the two lattices have in common. Once the best possible alignment between the two molecular lattices is obtained, those lattice points of the fitted molecule which are not yet in common with the reference molecule are added to create a new composite construct or larger reference lattice containing the information from both the molecules. This fitting and merging process is then repeated to include all the molecules of the training set in the growing HASL, resulting in a reference lattice entailing every point from all the molecular lattices.

In order to determine the activity contributions from different lattice points, initially the experimental activity value of a molecule is homogeneously divided among its all lattice points. For lattice points which are shared by more than one molecule, the partial bioactivity values are, at first, averaged over these points and, afterwards adjusted by an iterative protocol to fit the experimental activity data of the entire training set. This iteratively optimized HASL is then used as a standardized model to predict the activities of untested molecules. The bioactivity of a specific compound is forecasted by summing all the partial activity values at points in common with the composite reference lattice.

1.4.05. **CoMSIA**

Comparative Molecular Similarity Indices Analysis (CoMSIA) was developed by Klebe et al. to overcome certain limitations of CoMFA like imperfections in potential energy functions, use of cut-off limits, fragmented contour maps etc. In CoMSIA, molecular similarity indices calculated from modified SEAL similarity fields are employed as dimensions to simultaneously consider steric, electrostatic, hydrophobic and hydrogen-bonding properties. These indices are estimated indirectly by comparing the similarity of each molecule in the dataset with a common probe atom (having a radius of 1 Å, charge of +1 and hydrophobicity of +1) positioned at the intersections of a surrounding grid/lattice. For computing similarity at all grid points, the mutual distances between the probe atom and the atoms of the molecules in the aligned dataset are also taken into account. To describe this distance-dependence and calculate the molecular properties, Gaussian-type functions are employed. Since these underlying Gaussian-type functional forms are 'smooth' with no singularities, their slopes are not as steep as the Coulombic and Lennard-Jones potentials in CoMFA; therefore, no
arbitrary cut-off limits are required to be defined. These functions tend to produce values within a reasonable range, even in the case of overlapping atoms. Despite the fact that CoMSIA also suffer from most of the limitations of CoMFA, it offers following distinctive advantages:

- Use of the Gaussian distribution of similarity indices, which avoids the abrupt changes in grid-based probe–atom interactions
- The choice of similarity probe, is not limited to either steric or electrostatic potential fields but also include hydrophobic and hydrogen-bonding (hydrogen bond acceptors and donors) fields
- Effect of the solvent entropic terms can also be included by using a hydrophobic probe
- The standard CoMFA contours highlights those regions in space where the aligned molecules would favourably or unfavourably interact with a possible receptor environment. On the other hand, the CoMSIA contours indicate those areas within the region occupied by the ligands that “favour” or “dislike” the presence of a group with a particular physicochemical property. This relationship between the required properties and a probable ligand shape is a more direct guide to substantiate whether all features imperative for activity are present in the structures being considered.

1.4.06. GERM

Genetically Evolved Receptor Models (GERM) is a technique for 3D-QSAR and for constructing useful three-dimensional models of macromolecular binding sites in the absence of a crystallographically-determined or homology-modeled structure of the target receptor\(^4\). The primary requirement for GERM is a structure–activity series for which a sensible alignment of realistic conformers has been determined. The methodology consists of enclosing the superimposed set of molecules in a shell of atoms (analogous to the first layer of atoms in the active site) and allocating these atoms with explicit atom types (aliphatic H, polar H, etc., to match the types of atoms usually found in the proteins). Aliphatic carbon atoms are disseminated uniformly over a sphere surrounding the training set of aligned ligands, and their positions are adjusted to obtain maximum van der Waals interaction between the model carbon atoms and the ligand molecules. Once the positions of the carbons have been recognized, they can be occupied by any of the atom types, including no atom at all. One practical problem arises when the number of shell atoms and their atom types increases, since the number of possible combinations rises to a huge value thereby rendering
it impossible to systematically find a best possible model. The method therefore makes use of the genetic algorithms (GA) to solve this highly multi-dimensional search problem. The ligands in the training set are then docked into a GA generated receptor active site model, one at a time, and the intermolecular non-bonded interaction energies (van der Waals and electrostatic terms) are computed using a CHARMm molecular mechanics force field. Finally these calculated interaction energies (dimensions) are correlated with the biological activities of the molecules. The affirmative feature of this method is that the model is presented as a 3D-display of the receptor properties in space. The limitation of GERM methodology is that it considers only a single conformation of each ligand in the training set, as well as its single orientation in the binding site. Since this method is based on the computation of interaction energies with the hypothetical receptor, it is subjected to all the limitations of such methods including the alignment problem. However, if all the molecules of the set do bind in a manner that doesn't alter the binding site too much; GERM could be a good approach.

1.4.07. COMPASS

COMPASS is a 3D-QSAR approach which differs from the other methods in three ways:\(^\text{42}\):

- Physical properties are calculated only near the surface of the molecules
- Conformations and alignments of the molecules are automatically selected by the algorithm
- A non-linear chemometric method (neural networks) is employed

The Compass methodology begins by performing a routine conformational analysis to identify the low energy and probable bioactive conformation of each ligand. An approximate initial alignment of the molecules is carried out by means of a user-defined pharmacophore or a substructure common to all molecules in the data set. From each pose (conformation in a specific alignment), a set of real value features are computed which serve as a measure of the surface shape or the polar functionality of the pose in the vicinity of a particular point in space. A non-linear statistical model using neural networks is developed, which quantitatively and predictively elucidates the association between these surface characteristics of the molecules and their biological activity. Based on the resulting model, the molecules are realigned by allowing them to rotate, translate, and alter their conformations to achieve the best fit to a binding site. This process continues until all the molecules find poses which are further closer to the bioactive ones than those formerly considered. Using these improved and realigned molecular poses, a refined model is built.
which can reveal the structure-activity relationship more precisely. These alternating steps of model building and reposing iterates continuously till convergence is reached, resulting in the final predictions of activity and bioactive pose for each molecule. The results are displayed graphically for easy visual interpretation. The activities and bioactive poses of the new molecules can be predicted by subjecting them to a conformational search, positioning them in initial alignments using the same common substructure or pharmacophore as was used in the best model, realigning them according to the neural network model, selecting their potential bioactive poses and eventually predicting their activities using the final model.

Compass represents each pose of a molecule by a different set of feature values. The method presently allows the use of only three types of molecular features (dimensions): steric, hydrogen bond donor and acceptor features. Steric features are determined in terms of the distance from a sampling point to the van der Waals surface (adjacent atom) of a molecule in a specific pose. Similarly hydrogen bond donor and acceptor features are computed as the distance from the sampling points scattered in the vicinity of the surface of the molecules to the nearest hydrogen bond donor and acceptor groups, respectively. Compass currently have some limitations like absence of entropic contribution, requires user to provide a common substructure or pharmacophore as a qualitative hypothesis of binding, no approximation of the reliability of predictions, ligand flexibility issues etc.

1.4.08. RSM / CoRSA

Receptor Surface Model (RSM) is a non-atomistic approach that uses explicit surfaces to characterize the shape of the active site. It is based on the principle that the most active molecules are more likely to exploit the best spatial and electronic interactions with the receptor, while the least active do not and tend to have unfavourable steric or electronic interactions\textsuperscript{43, 44}. All the molecules in the data set are geometrically optimized and superimposed (using a template or some pharmacophore hypothesis), preferably in conformations that reveal the active, “bound” poses. A set of the most active analogs is chosen to cover the variety of steric and electrostatic variations likely to be encountered in the test data. This is followed by the generation of a surface, enclosing a volume common to all the aligned molecules, to represent their aggregate molecular shape, which is assumed to be complementary to the shape of the receptor site itself. For generating this surface, a volumetric field exemplifying the molecular shape is created for every aligned molecule. These fields are referred to as the shape fields, and are computed around a set of field sources
like a set of atoms, at different points in space (usually on a 3D-grid). These field sources results in a local field defined by a distance-dependent function. The shape fields from each individual molecule are then combined and a final volumetric shape field is generated from which an isosurface of the field can be computed to create an explicit object with well-defined shape. After the surface is created, information corresponding to the putative chemical properties of the receptor is linked with every surface point. A scalar value for each of these properties (including partial charge, electrostatic potential, hydrogen-bonding tendency and hydrophobicity) is calculated and stored with every surface point in the model. This information is used to map properties onto the surface during graphical display to visually convey the important active site characteristics in an intuitive fashion, as well as in computing ligand-(virtual) receptor interactions. The location-specific interaction energies (van der Waals, electrostatic interactions etc.) computed between a molecule and the virtual receptor surface model (i.e., between each surface point from the virtual receptor and the atoms in a molecule), can be used as descriptors or dimensions to be correlated with the biological activity of the molecules using PLS technique. The resulting formalism is termed as the Comparative Receptor Surface Analysis (CoRSA)\textsuperscript{45}. The RSM (and CoRSA) approach has the following distinctive characteristics:

- As compared to a pharmacophore model, the receptor surface model approach is conservative in nature, since it includes information on the steric extent of the training molecules, and so can penalize or eliminate molecules that cannot also assume the appropriate steric shape. This property can be useful in focusing de novo construction or database search for the most likely molecules
- The receptor surface model naturally represents the essential features of a receptor site itself, rather than the common features of the molecules that bind to it, and thus is visually intuitive and can be graphically modifiable
- In order to obtain conformations which are consistent with the model, molecules can be energy minimized within the receptor surface model
- A receptor surface model can profitably be utilized in database searching, to extract molecules similar in shape and reliable in electrostatics to a given receptor surface model query
1.4.09. COMBINE

Comparative Binding Energy Analysis (COMBINE) method was developed to take advantage of the structural data from ligand-macromolecule complexes, in a 3D-QSAR paradigm. The technique is based upon the hypothesis that the free energy of binding can be correlated with a subset of energy components calculated from the structures of receptors and ligands in bound and unbound forms. The ligands are divided into fragments and the same number of fragments is allocated to all the compounds, adding "dummy" fragments to the ligands lacking a particular fragment. The non-bonded (van der Waals and electrostatic) interaction energies are computed between each residue of the receptor and every fragment of the ligand, using a molecular mechanics force field. The energies are also calculated between all pairs of residues/fragments for the complexes and for the free ligands and receptor. The electrostatic interactions are computed using a distance-dependent dielectric constant, and no cut-off limits are employed for the non-bonded interactions. The insignificant descriptors are then eliminated from the data using the variable selection utility in GOLPE program, and finally the biological activities of the molecules are correlated with the interaction energy values (dimensions) by employing PLS technique. Like all other interaction energy based 3D-QSAR approaches, COMBINE also suffers from the inherent errors involved in the computation of these energies. Also, the predictive ability of the method can be enhanced by making improvements in various aspects like the description of the electrostatic term, the inclusion of suitable descriptors for solvation and entropic effects, and the optimization of particular facets of the methodology, such as the choice of ligand fragment definitions and the details of the variable selection protocol.

1.4.10. CoMMA

Comparative Molecular Moment Analysis (CoMMA) is one of the unique alignment-independent 3D-QSAR methods, which involves the computation of molecular similarity descriptors based on the spatial moments of molecular mass (shape) and charge distributions up to and including second order as well as related quantities. With respect to each molecular structure, two Cartesian reference frames are defined. One frame is the principal inertial axes calculated with respect to the center-of-mass. For neutral molecular species, the other reference frame is the principal quadrupolar axes calculated with respect to the molecular “center-of-dipole”. Dipolar, quadrupolar, and displacement descriptors are then calculated with reference to the principal inertial axes translated such that its origin is
superposed on the center-of-dipole. It is noteworthy that these descriptors are obtained after translation to the center of mass as well as the center of dipole for each molecule, to keep the system alignment-independent. Finally these molecular moment descriptors are correlated with the biological activities of molecules using the PLS technique. Literature reports suggest that CoMMA descriptors are sensitive to molecular conformations, but less sensitive than CoMFA field parameters. The authors propose that the CoMMA descriptors have a potential role in addressing the issues like large scale screening and molecular diversity. A web version of the CoMMA program is provided by the IBM informatics group. A slight variant of this approach, termed as CoMMA2, has also been developed by the author.

1.4.11. VFA

The Voronoi Field Analysis (VFA) approach employs a non-regularly spaced grid based on steric and electrostatic potential indices calculated at lattice points; however the indices are manipulated to be assigned to the superimposed molecular regions defined by the Voronoi polyhedral division\(^{48}\), but not to each of the lattice points as in CoMFA. The methodology begins by subjecting the molecules in the data set to conformational analysis and energy minimization routines, followed by superimposition based on their common skeleton and accounting for their conformational flexibilities. The total union volume occupied by the superimposed set of molecules is divided into subspaces referred to as Voronoi polyhedra, each including a reference point (an atom) with certain coordinates. Each polyhedron is a set of every point closer to its reference point than to any other. In reality, each polyhedra is a region surrounded by a set of planes, each of which is perpendicular to, as well as, bisects the line connecting the reference point with each of the adjoining reference points.

The compound with the simplest structure in the data set is chosen as a template molecule and the positions of all of its atoms including hydrogens are defined as the initial reference points (atoms). The molecule with the largest number of atoms in the data set is next selected, and its skeletal structure (common in the series) is superimposed over that of the template molecule. The positions of atoms of the second molecule are compared with the reference points of the template. If no reference point of the template is within 1 Å region from each atom in this molecule, the ‘isolated’ atomic position is designated as a new reference point. Otherwise, the atoms are ignored and no longer referred any more. Other molecules in the data set are then selected and superimposed successively in decreasing order of their size, and their atomic positions are compared with the reference points defined in the previous steps. In
each step, ‘isolated’ atoms are defined as new reference points. These steps are repeated continuously until all compounds of the data set are superimposed to define every Voronoi reference point. A cuboid is defined with six tangential planes surrounding the union volume of the superposed set of molecules. The cuboid region is divided into a 3D-lattice with a spacing of 0.3 Å. The potential energy indices are then calculated at each lattice point using ‘hard-sphere potential’ model. As per this model, the steric potential indices at lattice points located on and inside the van der Waals surface of the molecule are set as unity but otherwise zero. The electrostatic potential indices are calculated in a manner similar to CoMFA, using a +1 charge as the probe and the standard Coulombic potential function. The electrostatic potentials at lattice points inside the van der Waals surface of the union volume of the superimposed set of molecules are neither calculated nor considered in each polyhedron. The potential indices calculated at each of the lattice points are then transformed into those belonging to each of the Voronoi polyhedra. The indices for the \(k\)-th polyhedron, \(V^{\text{st}}_k\) and \(V^{\text{el}}_k\), are defined as the summation of the steric and the average of the electrostatic potential indices estimated at lattice points accommodated inside the \(k\)-th polyhedron, respectively.

PLS technique is then used to extract new components and descriptors/dimensions from the polyhedral variables. In order to avoid problems in auto-scaling of the steric and electrostatic potential indices of different sizes, the PLS method is applied to each of the indices individually to generate independent \(i\)-th steric and \(j\)-th electrostatic latent variables, \(Z^{\text{st}}_i\) and \(Z^{\text{el}}_{ij}\). MLR analysis is finally performed to correlate the activity index with the latent variables and, if necessary, external parameters such as logP. The optimum steric and the electrostatic variables are then represented in the form of 3D-contour maps drawn as continuously connected cubic regions of 0.3 Å size, depicting those Voronoi polyhedral regions around which the variations in the polyhedral parameters is responsible for the potency variations within the set. Some of the distinguishing features of VFA approach include:

- The number of molecular field variables is significantly less than that in CoMFA, leading to the stabilization of the PLS solution
- The ‘hard-sphere potential’ model for the calculation of steric potentials is relatively simple, and gives results comparable to the Lennard-Jones potential function in CoMFA
- The PLS technique is applied separately to the steric and electrostatic field variables thereby avoiding problems related to auto-scaling of the two types of variables
The molecular field parameters are calculated at lattice points with the spacing of 0.3 Å, thus yielding robust models and rendering the methodology independent of the location of the superimposed molecular set relative to the lattice points.

1.4.12. PARM

The Pseudo Atomic Receptor Model (PARM) is a 3D-QSAR approach which generates an atomic-level pseudo-receptor model based on a set of known structure-activity relationships\textsuperscript{49}. The formalism defines fifteen types of pseudo-receptor atoms that can possibly be found in a protein. They are divided into three groups: positive group with positive partial atomic charge, negative group with negative partial atomic charge, and neutral group with neutral atom types. The training set molecules are superimposed on a specific pharmacophore and a set of grid points is generated around the common surface of the aligned ligands. Receptor-based models are built by positioning pre-defined atoms at these grid points in 3D-space to mimic a receptor active site and its interactions with the ligand molecules. Since any type of atom can be arbitrarily selected and placed at each point, the number of achievable models can be huge. Therefore the methodology makes use of a genetic algorithm to screen this enormous range of possible models, and select only those that have a good correlation between the calculated binding energy and bioactivity. The PARM approach is quite similar to the GERM algorithm, but differs only in the underlying genetic process. The genetic operation of PARM first generates a large number of receptor models (i.e., possible solutions or individuals). Taking into consideration the complementarities between the electrostatic field of the ligand and that of the receptor, unlike GERM, each preliminary model is developed not entirely on the basis of a random mechanism but in a charge-dependent, random manner. In other words, the PARM methodology presumes that the charge on the receptor surface point is complementary to the partial atomic charge of the adjoining ligand atom that can interact with it. Therefore, it is necessary to allocate one atom type to each bit (grid point position), while generating a possible model. This selection of atom type in PARM is not absolutely random but is based on the charge of the atom closest to the grid point. Thus, a formal charge is defined on each grid point. If the receptor model is generated over a single molecule, each grid point is given a formal charge which is equal to but opposite in sign to the charge of a ligand atom that is next to the grid point. However, if the model is developed over a set of ligands, every grid point is assigned a formal charge that is equal but opposite in sign to the average partial atomic charge of the closest ligand atoms in the whole set of molecules.
After the allocation of formal charge and specific atom type to each grid point based on the electrostatic complementarity principle, the algorithm generates an initial population of individuals or receptor models. The van der Waals and electrostatic interaction energies (dimensions) are computed between each ligand and the receptor model (i.e., individual), which are then correlated to their bioactivities using a linear regression technique. These models are finally used to predict the activities of the test set molecules. The activity of any new molecule can be predicted by computing its binding energy with the receptor model and substituting it in the QSAR equation of that model.

1.4.13. SOMFA

Self-Organizing Molecular Field Analysis (SOMFA) technique is somewhat similar to the molecular similarity analysis, CoMFA, Free-Wilson analysis and the hypothetical active site lattice (HASL) methods, but is theoretically simple and comparatively comprehensive in nature. A model is developed from a set of superimposed molecules with recognized activities, constituting the training set. The mean activity of the training set is subtracted from the activity of each molecule; to obtain their mean centered activity values. This procedure scales the data set in such a manner that the most active molecules have positive mean centered activity values while least active ones have negative values. 3D-grids are generated around the molecules with values at the grid points signifying the shape or electrostatic potential. The steric potentials (Shape values) are given a value of unity inside the van der Waals envelope, and zero outside. The values of electrostatic potential at grid points are computed in the usual manner from the partial charges distributed across the atom centers. The shape or electrostatic potential values at every grid point for a molecule are multiplied by the mean centered activity for that molecule. This is a type of variable filtering step, which imparts weights to the grid points such that the most active and least active molecules have higher values than the less interesting molecules close to the mean activity. The grid values for each molecule in the training set are then summed to give master grids for each property (steric or electrostatic potential). From these master grid values, the so-called SOMFA\textsubscript{property.1} descriptors/dimensions are calculated, which are finally correlated with the log-transformed activity values of the training set molecules. At each point, the master grid values can be displayed graphically to emphasize features favourable or unfavourable to activity. SOMFA approach does not require the computation of probe-based interaction energies and thus has the advantage of being considerably fast and simple, but it does suffer from the alignment problem.
1.4.14. **CoMSA**

Comparative Molecular Surface Analysis (CoMSA) is a non-grid 3D-QSAR approach that makes use of the molecular surface for defining those regions of the compounds which are required to be compared using the mean electrostatic potentials\(^1\). The methodology proceeds by subjecting the molecules in the data set to geometry optimization and assigning them with partial atomic charges. The Kohonen's self-organizing maps (SOM, a type of neural network) are then employed to transform the three-dimensional surface of the molecules into two-dimensional topographical maps, by extracting the signals from the Cartesian coordinates of the points sampled randomly at the van der Waals surface of the molecules. The partial atomic charges of the atomic molecular representations are also projected onto the 2D-topographical maps. The molecular electrostatic potentials (MEPs) are calculated at the surface points and a mean value of the potential analogous to the respective points found in each grid cell (of CoMFA like methods) is utilized to explain this cell. The calculated mean electrostatic potential values are converted into vectors (dimensions) and the vectors expressing all the molecules in the series are superimposed onto a matrix, by comparing the respective topographical maps of the molecules. The ensuing comparative matrix of the mean electrostatic potentials (transformed into vectors) is finally used to develop a 3D-QSAR model using the PLS technique. The distinctive feature of CoMSA is that, in contrast to CoMFA and related approaches, it compares the molecular properties explaining not a discrete set of points but the average property values (MEPs) calculated for a certain area of the molecular surface.

1.4.15. **AFMoC**

Adaptation of Fields for Molecular Comparison (AFMoC) is a 3D-QSAR method involving fields derived from the protein environments (and not from the superimposed ligands as in CoMFA), therefore it is also known as a 'reverse' CoMFA (=AFMoC) approach\(^2\). The methodology begins by placing a regularly-spaced grid into the receptor binding site, followed by mapping of the knowledge based pair-potentials between protein atoms and ligand atom probes onto the grid intersections resulting in the potential fields. Based on these potential fields, interaction fields are generated by multiplying distance-dependent atom-type properties of actual ligands docked into the active site with the neighbouring grid values. These atom-type specific interaction fields are then used as dimensions or descriptors to be correlated with the binding affinities of the molecules using PLS technique, which assigns
individual weighting factors to each field value. Finally the results are displayed graphically by using contribution maps, and binding affinities of novel ligands are predicted by applying the derived 3D-QSAR equation. The distinctive features of this approach include:

- A tailor-made scoring function is combined with a protein-based CoMFA approach, thereby overcoming the prerequisite to involve complete ligand training sets
- The gradual shift from generally valid knowledge-based potentials to protein-specific pair-potentials, reflects the amount and the degree of structural diversity existing in the ligand training data
- Atom-type specific interaction fields are used which are mutually orthogonal in nature and thus eases the interpretation of PLS results
- In addition to the enthalpic contribution, the methodology is also expected to include the entropic effects resulting from (de-)solvation, since structural knowledge from experimentally-determined complexes is converted into statistical pair-potentials

1.4.16. FLUFF-BALL

FLUFF-BALL is a 3D-QSAR approach empowered by a semiautomatic superimposition algorithm called FLUFF and an associated QSAR technique termed as BALL\textsuperscript{53}. The superimposition algorithm is based on a novel field-fitting procedure called Flexible Ligand Unified Force Field (FLUFF), executed by exploiting a customized MMFF94 force field. This technique superimposes a set of molecules over a given template molecule, in a manner so as to maximize the similarity of the steric and electrostatic field volumes of the ligand and the template. In the process, the molecules can either be made completely flexible to permit maximal adaptation and seek the best common conformation, or can be constrained by freezing selected atoms. This geometrically optimized superimposed set of molecules is then used to predict the biological activities by a QSAR technique based on the Boundless Adaptive Localized Ligand (BALL) approach. This method ties the internal coordinate system with the template molecule, by placing the anchor points (vertices) of the local grid at the atomic centers of the template. The algorithm then evaluates the similarities between the template and the ligand molecules and computes various van der Waals and electrostatic terms/dimensions, which are finally correlated with the biological activities using the PLS technique. The FLUFF-BALL approach offers the following characteristic advantages:
• The semiautomatic superimposition procedure is computationally simple, fast and requires only a minimum amount of human intervention in the form of a rough initial alignment. The flexibility option allows some changes in the template’s conformation without adverse effects on the predictive power of the model.

• Standard CoMFA-like methods are based on the uniform global positioning system, necessitating all molecules to be aligned to the same spatial coordinates. The template-based local coordinate modification approach of the FLUFF-BALL formalism renders it grid-independent or immune to the rotations and translations of the global coordinate system.

1.4.17. CoMASA

Comparative Molecular Active Site Analysis (CoMASA) is a rapid evaluation method for obtaining 3D-QSAR based on the extraction of molecular representational coordinates instead of the lattice points around the molecules as in CoMFA. To avoid lattice-based arbitrariness regularly encountered in CoMFA and related approaches, the method digs out (by using cluster analysis) a minimum collection of points which are spatially occupied by the molecules, and are essentially adequate to elicit a biological response. The methodology proceeds by superimposing the compounds in the data set based on a template molecule. All atoms are extracted from the aligned set of molecules and their interatomic distances are calculated. The closest atoms with a distance of less than 0.75 Å are removed and replaced by a new pseudo atom/point created from their weighted average. Addition of ring centroids for the generation of atomic represented points by cluster analysis is also tried. The process is continued until the distances between all the atoms or pseudo atoms/points are greater than the threshold value of 0.75 Å, thereby resulting in a bare minimum set of representational coordinates. Interaction fields of steric, electrostatic and hydrophobic properties (dimensions) are calculated for each molecule at these molecular represented points, by means of four different evaluation functions namely rapid similarity index function, SEAL function, Good’s Gaussian similarity index function, and simple indicator variables (a variant of the topological indices). Finally the PLS technique is used to derive a QSAR equation from the high-dimensional data, and the results are visualized graphically in the form of isocontours or isopleths explaining the relationships between the differences in the molecular active site and variations in the dependent variable. Though CoMASA approach is limited by the requirement for a superimposed set of molecules, it offers following advantages:
The resulting QSAR coefficient maps of CoMASA are easily interpretable and elucidate the relationships between the differences in the molecular active site and variations in the dependent variable. These maps can be easily transformed to pharmacophore and/or queries required for 3D-database searches.

- Orientation and translation of molecules against a lattice grid are not necessary
- Because of its rapid scoring functions and reduced interaction points, standard computers can be used to run the method

1.4.18. CoMPIA

Comparative Molecule/Pseudo receptor Interaction Analysis (CoMPIA) is a 3D-QSAR approach developed by incorporating the concepts of postulated pseudo-receptor and genetically optimized probes into the CoMFA method. The methodology involves geometry optimizations of the structures of all the molecules followed by their superimposition based on a common template molecule. The resulting space encompassed by the set of superimposed molecules is partitioned into grids with sufficient number of lattice points to accommodate all the probe atoms. Different types of probe atoms are positioned randomly at each lattice point to calculate steric, electrostatic and hydrophobic interactions between the probes and every molecule in the set. These interaction terms are presented as chromosomes to a genetic algorithm which then attempts to find the pseudo-receptor mode by optimizing the probe distributions and representing the combination mode of probe atoms at lattice points. Finally the resulting interaction energies are correlated to the molecular activities using PLS technique. The distinguishing features of CoMPIA approach are as follows:

- Nine different types of hybrid atoms are employed as lattice point probes
- The distributions of different probes are auto-determined by the algorithm itself to avoid illogical and biased selection of lattice point probes, thereby improving interpretabilities of the models.
- To avoid discrepancies in comparing molecular properties, CoMPIA directly compares ligand-receptor interaction patterns
- To generate an optimal amalgamation of improved flexibility and adaptability, genetic algorithms are employed to select different probe-atom types at different lattice points
- The method allows the use of some lattice points with no probe as ‘empty’, thereby favouring the organic combination of variable selection and latent regression (as GA-PLS) modelling to significantly reduce the noise
1.5. Statistical methods for QSAR analysis

Statistical or chemometric techniques form the mathematical foundation for QSAR analysis. Some of these methods employed in two major aspects of QSAR modelling (model development and model validation) are briefly described below:

1.5.1. Model development

Various chemometric methods can be used for developing a QSAR model. Some of these techniques are listed in Table 1.03, and are briefly described below:

Table 1.03. Statistical techniques for building QSAR models

<table>
<thead>
<tr>
<th>Linear regression analysis (RA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple linear regression</td>
</tr>
<tr>
<td>Multiple linear regression (MLR)</td>
</tr>
<tr>
<td>Stepwise multiple linear regression</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multivariate data analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal component analysis/regression (PCA/PCR)</td>
</tr>
<tr>
<td>Partial least square analysis (PLS)</td>
</tr>
<tr>
<td>Genetic function approximation (GFA and G/PLS)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pattern recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster analysis</td>
</tr>
<tr>
<td>Artificial neural networks (ANNs)</td>
</tr>
<tr>
<td>k-nearest neighbour (kNN)</td>
</tr>
</tbody>
</table>

1.5.1.1 Linear regression analysis

Among the increasing pool of various statistical methods available in the literature, linear regression analysis techniques are considered as easily interpretable methods indicated for QSAR analysis\(^{58}\). These regression techniques construct a statistical model to represent the correlation of one or more independent variables \(x\) with a dependent explicative variable \(y\). The model can be utilized to predict \(y\) from the knowledge of \(x\) variables, which can be either quantitative or qualitative. Simple linear regression, multiple linear regression, and stepwise multiple linear regression are some of its variants.

**Simple linear regression**: This method performs a standard linear regression calculation to generate a set of QSAR equations that include a single independent descriptor \(x\) and a dependent variable \(y\). Thus, a one-term linear equation is produced separately for each independent variable from the descriptor set. This technique is suitable for generating simple
relationships between structure and activity exploring some of the most important descriptors governing the activity. However, the interaction of multiple descriptors is ignored. The simple linear regression can be expressed by the equation:

\[ y = a + bx \]

where the dependent variable \( y \) is expressed in terms of the independent variable \( x \) by means of two parameters: the constant \( a \), also referred to as the intercept and the regression coefficient \( b \).

**Multiple linear regression (MLR):** This technique also referred to as the linear free-energy relationship (LFER) method, is an extension of the simple regression analysis to more than one dimension\(^{59}\). MLR generates QSAR equations by performing standard multivariable regression calculations to identify the dependence of a drug property on any or all of the descriptors under investigation. The possibility of chance correlation is checked through the values of correlation coefficient \( (r) \), Student’s \( t \)-value, Fisher’s \( F \) ratio, standard deviation \( (s) \), and through independent tests like the leave-one-out (LOO) method. The significance of correlation can be judged through cross-validated correlation coefficient \( (r^2_{cv} \) or \( q^2 \)) values and also by the \( y \)-scrambling technique described in the later sections. MLR assumes that all variable are independent, and not correlated. However, in the multivariate case, \( i.e., \) MLR analysis involving more than one independent variable, the relationship is expressed with the following single multiple-term linear equation:

\[ y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m + e \]

The MLR analysis estimates the regression coefficients \( (b_i) \), by minimizing the residual error \( (e) \), which quantifies the deviation of a particular point from the regression line, as in the case of simple linear regression.

**Stepwise multiple linear regression:** This method is a commonly used variant of MLR which also creates a multiple-term linear equation, but not all the independent variables are used\(^{60}\). In contrast to MLR, each independent variable is sequentially added to the equation and new regression is performed every time. The new term is preserved only if the model passes a test for significance. This regression technique is especially useful when the number of descriptors is large and the key descriptors are unknown.
1.5.1.2. **Multivariate data analysis**

The regression analysis methods described above have now been replaced by multivariate chemometric techniques which try to explain an extended set of variables by means of a reduced number of new latent variables possessing the maximum amount of information relevant to the problem. These techniques project multivariate data into a space of lower dimensions thereby providing insight to visualize, classify, and model large datasets. These latent variables are orthogonal and hence can be used in multiple linear regressions. Some of the multivariate data analysis methods are described below:

**Principal components analysis (PCA):** This is a data reduction technique that does not generate a QSAR model but seeks for relationships among independent variables\(^{61}\). It then creates a new set of orthogonal descriptors - referred to as principal components (PCs) which describe most of the information contained in the independent variables in order of decreasing variance. Consequently, PCA reduces dimensionality of a multivariate dataset of descriptors to the actual amount of data available. When principal components are employed as the independent variables to perform a linear regression, the method is termed as the principal components regression (PCR). In other words, PCR applies the scores from PCA decomposition as regressors in the QSAR model, to generate a multiple-term linear equation\(^{62}\).

**Partial least squares (PLS):** This is an iterative regression procedure that produces its solutions based on linear transformation of a large number of original descriptors to a small number of new orthogonal terms called latent variables\(^{63}\). PLS gives a statistically robust solution even when the independent variables are highly interrelated among themselves, or when the independent variables exceed the number of observations. Thus, PLS is able to analyze complex structure-activity data in a more realistic way, and effectively interpret the influence of molecular structure on biological activity. It is one of the standard statistical methods used for the development of predictive 3D-QSAR models.

**Genetic function approximation (GFA):** This method serves as a good alternative to standard regression analysis for building QSAR equations\(^{64}\). It employs the natural principles of evolution of species which leads to improvements by recombination (mutation and crossover) of independent variables. This technique results in multiple models generated by evolving random initial models using genetic algorithm. The method is suitable for obtaining QSAR equations when dealing with a larger number of independent variables. It can build
linear as well as higher-order non-linear equations, perform automatic outlier removal and
classification by utilizing spline-based terms. Genetic partial least squares (G/PLS or GA-
PLS) is a valuable analytical tool that has evolved by combining the best features of GFA and
PLS\textsuperscript{65}, and has been widely preferred by the researchers\textsuperscript{66-69}.

1.5.1.3. Pattern recognition

In recent years, other methods to perform qualitative or classification studies have been
spurred in the field of QSAR. The so-called pattern recognition methods based on the
principle of analogy are used for the detection of the distance or closeness within the large
amount of multivariate data\textsuperscript{70}. It searches for structural features such as the presence (or
absence) of certain groups, number of a certain type of atom, or mass spectral-fragmentation
so that new compounds can be classified as similar or dissimilar to the members of the
existing classes. Some of the pattern recognition methods are briefly explained below:

Cluster analysis: It is a statistical pattern recognition method used to investigate the
relationship between observations associated with several properties and to partition the
dataset into categories consisting of similar elements\textsuperscript{71}. It allows for the consideration of the
inactive compounds in the analysis and can be used to study a large set of substituents to
identify which of the subsets share similar physical properties.

Artificial neural networks (ANNs): This technique has its origin from the real neurons
present in an animal brain. ANNs are parallel computational systems consisting of groups of
highly interconnected processing elements called neurons, which are arranged in a series of
layers\textsuperscript{72}. The first layer is termed the input layer, and each of its neurons receives data from
outside/user, corresponding to one of the independent variables used as inputs in QSAR.
Subsequent to the input layer, there are one or many layers of processing neurons,
collectively termed as the hidden layers. The last layer is the output layer, and its neurons
handle the output from the network. Each layer may make its independent computations and
may pass the results to another layer. The working of ANNs is given below:

- Each input descriptor value is multiplied by the connection weight, as per its significance
- The weighted inputs are summed up and supplied to the hidden layers, where a nonlinear
  transfer function does all the required processing
- The results of the transfer function are communicated to the neurons in the output layer,
  where the results are interpreted and finally presented to the user.
**k-Nearest Neighbour (kNN):** This method is one of the simplest machine learning algorithms, most commonly used for classifying a new pattern (e.g., a molecule). The technique is based on a simple distance learning approach whereby an unknown/new molecule is classified according to the majority of its k-nearest neighbours in the training set. The nearness is determined by a Euclidean distance metric (e.g., a similarity measure computed using the structural descriptors of the molecules). Typically, the kNN approach is executed as follows:

Euclidean distances between an unknown object (u) and all the objects in the training set are computed.

Based on the calculated distances, k objects from the training set most similar to object u are selected.

Object u is assigned to the group to which the majority of the k objects belong.

An optimal k value is selected by optimization through the categorization of a test set of samples or by leave-one-out cross-validation.

### 1.5.2. Model validation

Validation is a crucial element of any QSAR analysis. The reliability of a 3D-QSAR model depends on how well the model can predict the activity of compounds outside the training set rather than how well the model reproduces the biological activity of compounds included in the model. Based on this criterion, various validation approaches can be broadly classified into two categories, internal and external validation techniques, and are described below:

#### 1.5.2.1. Internal validation techniques

These methods are used for validating the reliability of the QSAR model in reproducing the biological activities of those molecules which have been included in its training set or which have been used for building/developing that model. These techniques include:

**Correlation coefficient (r):** This is a measure of the degree of linearity of the relationship. It signifies the quality of fit of the model and quantifies the variance in the data. In an ideal situation, the correlation coefficient must be equal to or approach 1, but in reality due to the complexity of biological data, any value above 0.9 is appreciable. Correlation coefficients for the variables in a dataset are compiled in a correlation matrix, which shows the relationship of one descriptor with another. The correlation matrix ensures that variables of significance...
are orthogonal to each other. The addition of every new variable to the model always increases the $r$, unless the new variable is a constant of a linear combination of other variables, which would not produce any effect. The increase in $r$ caused by adding new variable signifies over-fitting of the data.

**Coefficient of multiple determinations or Pearson's correlation coefficient** ($r^2$): This is the squared correlation coefficient which informs about how well the model reproduces the experimental data\textsuperscript{24}. It is a quantitative measure of the precision of adjustment for the fitted values to the observed ones. The closer it approaches to the unity, the more similar are the adjusted values to the experimental ones, suggesting that the model fits the data unerringly. However, an $r^2$ close to 1 does not mean that the model is perfect; the addition of any new descriptor to the model induces an ever-increasing of $r^2$, even if the newly added descriptor does not contribute to the model. Thus, other measures are required to determine the predictive capacity of the model.

**Cross-validation (CV):** It is one of the most extensively employed methods for the internal validation of a statistical model\textsuperscript{25}. In cross-validation, the predictive ability of a model is estimated using a reduced set of structural data. Usually, one element of the set is extracted each time, and a new model is derived based on the reduced dataset, which is then employed to predict the activity of the excluded molecule. The procedure is repeated $n$ number of times until all compounds have been excluded and predicted once. This is the so-called leave-one-out (LOO) cross-validation method\textsuperscript{26}. Analogously, leaving out more than one molecule of the dataset at a time is termed as leave-n-out or leave-many-out cross-validation method\textsuperscript{26}. The outcome of LOO procedure is a cross-validated correlation coefficient $r^2_{cv}$ (or $q^2$) which is a criterion of both robustness and predictive ability of the model:

$$q^2 \text{ or } r^2_{cv} = \frac{\text{PRESS}_0 - \text{PRESS}}{\text{PRESS}_0}$$

where PRESS\textsubscript{0} is the mean of the observed biological activity while PRESS (predicted residual sum of squares) is the sum of the squares of the differences between the predicted and the observed activity values\textsuperscript{27}. Many researchers consider high $q^2$ as the ultimate proof of high predictive power of the QSAR model which is incorrect. It has been established that, in cases where test sets with known values of biological activities were available for prediction, there existed no correlation between the $q^2$ and $r^2$. Therefore, $q^2$ should be regarded as a measure of internal consistency of the derived model rather than as a true indicator of the predictability. It should be noted that, since it is easier to fit the experimental data than to
predict them from the QSAR model, $r^2$ of the model is always higher than $q^2$. Cross-validation is not foolproof. In highly redundant datasets with fewer degrees of freedom, it can give an over-optimistic result. It may also improperly indicate a lack of correlation if all the compounds in the dataset are unique. Therefore, it can be concluded that despite its wide acceptance, a high value of $q^2$ alone is an insufficient criterion for a QSAR model to be highly predictive.

Bootstrapping: This is another technique that can be used along with cross-validation to evaluate the robustness and the statistical confidence of the QSAR model. It involves simulating a large number of datasets which are of the same size as original and are produced by randomly selecting samples from the original dataset. In each PLS run some objects may be excluded while some others might be sampled more than once. The statistical calculation is run on each of these bootstrap samplings. The difference between the parameters calculated from the original dataset and the mean of the parameters calculated from many bootstrap samplings is a measure of the biasness of the original calculations. Since it demands heavy computation with relatively smaller gains compared to cross-validation, the technique is not very attractive.

Randomization or y-scrambling: This method is a rigorous alternative to cross-validation and bootstrapping in which the biological activity values are re-assigned arbitrarily to different molecules in the same dataset, and a new regression is performed. Only if the results from a PLS model, using the original sequence of the biological data, is significantly better than the results from the ‘scrambled’ models, can one be sure that significant correlation indeed exists between the biological data and the independent variables, and it has not been resulted from a chance correlation. The randomization test analyzes the ability of the statistical model to derive real structure-activity relationships.

Fischer statistic or F-value: The F-value is one of the several variance-related parameters that can be used as a measure of the level of statistical significance of the regression model. A higher F value implies that a more significant correlation has been reached. It is used as a criterion to determine whether a more complex model is significantly better than a less complex one.
1.5.2.2. **External validation techniques**

The extrapolative or external predictive ability of the model can be evaluated by forecasting the activity of an external test set of molecules using the QSAR models derived from the training set. Following methods are used for validating the reliability of the model in predicting the biological activities of those molecules which have not been included in its training set or which have not been used for building/developing that model:

**Predictive correlation coefficient** ($r^2_{\text{pred}}$): The predictive correlation coefficient is analogous to the cross-validated $r^2$ (or $q^2$), but is a measure of the predictive ability of the derived QSAR model. It is calculated by the following formula:

$$r^2_{\text{pred}} = \frac{(\text{SD} - \text{PRESS})}{\text{SD}}$$

where SD is the sum of squared deviations between the biological activities of the test set molecules and the mean activity of the training set molecules, while PRESS is the sum of squared deviations between the observed and the predicted activities of the test set molecules. Higher the $r^2_{\text{pred}}$ value, better is the external predictive ability of the QSAR model.

**Predictive correlation coefficient** ($p^2$): Vedani *et al.* have defined $p^2$ as a true measure of the external predictive ability of a QSAR model. It is calculated exactly in the same manner as the $r^2_{\text{pred}}$ value using the formula:

$$p^2 = \frac{(\text{SD} - \text{PRESS})}{\text{SD}}$$

The only difference is that here SD is the sum of squared deviations between the biological activities and the mean activity of the test set molecules. However, it should be noted that the $r^2_{\text{pred}}$ and $p^2$ values tend to be equal to each other, if the mean activities of the training and test set of molecules are same.
1.6. Problem statement and the scope of the thesis

Multidimensional QSARs represent an un-restrained extension of 3D-QSAR methods attempting to overcome their shortcomings. These higher-dimension QSARs are more automatable, which increases efficiency and allows the processing of larger data sets. Though these modern methods are noble attempts at handling the conformational flexibility of the molecules, they are often more complex to use and still in the infancy of their development. It is not uncommon to find the various QSAR methods to be of greater or lesser utility in different circumstances. In other words, it is very hard to find a single method that will always give a useable, extrapolative model. Owing to the different QSAR methodologies, deciding which QSAR method to use depends on the composition of the system of interest and the desired results. However, there are presently no clear guidelines to facilitate the selection of one method over the other, because there have been only few wide-ranging comparisons of various approaches for modelling structure-activity relationships.

In light of the facts mentioned above, the objective of the current work is to compare the different QSAR methodologies in order to investigate the effect of adding extra dimensions on them. For this, a variety of large and diverse protein-ligand databases were assembled from the literature. Different thermodynamic events of ligand–receptor binding leading to a biological response like interaction energies (steric, electrostatic and hydrophobic), solvation, entropy changes etc., were taken into consideration as additional dimensions. For all the datasets, QSAR tables were generated by incorporating different combinations of these dimensions in an incremental order. With a chemometric method like G/PLS (Genetic-partial least-square analysis), rigorous QSAR models were built correlating the structural dimensions/descriptors with the molecular activities for the respective datasets. The models were validated both internally using various cross-validation techniques, as well as, by predicting the activity of an external test set of molecules. Finally the influence of each added dimension on the predictive power as well as interpretability of the various QSAR models was examined.

Due to the widespread application of QSAR methods in concert with synthesis and screening of a series of analog, it is anticipated that the outcomes of this work will facilitate a better application of the QSAR methodologies by the medicinal/synthetic chemists. An improved correspondence between expected and actual prediction accuracy will enhance the usefulness of QSAR models for all involved in the lead optimization process.


