Chapter 1  

Computer Aided Drug Design: An Overview

1.1 Abstract
In this chapter a brief introduction to the computer-aided drug design (CADD) methodologies is given. The theoretical basis of CADD involves quantum mechanics and molecular modeling studies like structure based drug design; ligand-based drug design; database searching and binding affinity predictions. Finally, a brief description of the present work is given.

1.2 Introduction
Drug discovery and developing a new medicine is a long, complex, costly and highly risky process that has few peers in the commercial world. This is why computer-aided drug design (CADD) approaches are being widely used in the pharmaceutical industry to accelerate the process. The cost benefit of using computational tools in the lead optimization phase of drug development is substantial. On an average, it takes 10-15 years and US $500-800 million to introduce a drug into the market, with synthesis and testing of lead analogs being a large contributor to that sum. Therefore, it is beneficial to apply computational tools in hit-to-lead optimization to cover a wider chemical space while reducing the number of compounds that must be synthesized and tested in vitro. The computational optimization of a hit compound involves a structure-based analysis of docking poses and energy profiles for hit analogs, ligand-based screening for compounds with similar chemical structure or improved predicted biological activity, or prediction of favorable affinity or optimize drug metabolism and pharmacokinetics (DMPK) or absorption, distribution, metabolism, excretion, and the potential for toxicity (ADMET) properties. The comparably low cost of CADD compared with chemical synthesis and biological characterization of compounds make these methods attractive to focus, reduce, and diversify the chemical space that is explored.

CADD is capable of increasing the hit rate of novel drug compounds because it uses a much more targeted search than traditional high throughput screening (HTS) and combinatorial chemistry. It not only aims to explain the molecular basis of therapeutic activity but also to predict possible derivatives that would improve activity. In a drug discovery campaign, CADD is usually used for three major purposes: (1) filter large compound libraries into smaller sets of compounds that can be tested experimentally; (2)
guide the optimization of lead compounds, to increase its DMPK properties including ADMET; (3) design novel compounds, either by "growing" starting molecules one functional group at a time or by piecing together fragments into novel chemotypes.³

CADD can be classified into two general categories: structure-based and ligand-based. Structure-based CADD relies on the knowledge of the target protein structure to calculate interaction energies for all the compounds to be tested, whereas ligand-based CADD exploits the knowledge of known active and inactive molecules through chemical similarity searches or construction of predictive, quantitative structure-activity relationship (QSAR) models.⁴ Structure based CADD is generally preferred where high-resolution structural data of the target protein are available, i.e., for soluble proteins that can readily be crystallized. Ligand based CADD is generally preferred when no or little structural information is available, often for membrane protein targets. The central goal of structure based CADD is to design compounds that bind tightly to the target, i.e., with larger reduction in free energy, improved DMPK/ADMET properties, and are target specific, i.e., have reduced off-target effects.⁵ A successful application of these methods will result in a compound that has been validated in vitro and in vivo and its binding location has been confirmed, ideally through a co-crystal structure. One of the most common uses in CADD is the screening of virtual compound libraries, also known as virtual high-throughput screening (vHTS). Fig. 1.1 illustrates the stages in the drug discovery process⁶ and Fig. 1.2 explains virtual drug discovery process⁷.

Figure 1.1 Stages in the drug discovery process.⁶
1.2.1 Receptor theory

The concept that therapeutic agents produce their selective action in modifying disease symptoms by acting as "magic bullets" at discrete molecular targets within the body, is generally attributed to Paul Ehrlich at the turn of the 19th century as part of the now seminal "lock and key" hypothesis. This hypothesis has described drugs as receptor’s ligands or enzyme substrates that selectively modulate the function of unknown molecular targets to produce beneficial effects. The receptor theory involves, to a very major extent, the classical enzyme kinetic model based on the law of mass action, which was derived by Michaelis and Menten in 1913. The interaction between receptor and a ligand can be looked upon as:

\[
[\text{Receptor} + \text{Ligand}] \rightarrow [\text{RL}] \rightarrow R + \text{Cellular Effect} \quad (1)
\]

The ligand L binds to the receptor R and alters the nature of the receptor interaction with its associated membrane components to effect a change in cellular and ultimately, tissue function. Ligands interacting with the receptors have two intrinsic properties: Affinity and Efficacy. Affinity is the ability to recognize and bind to the receptor while the ability of the ligand to effect a change in cellular processes via activation of transmembrane transduction mechanisms involving G-protein complexes or ion channels is a measure of efficacy. In addition to the affinity of a receptor for its ligand, the response to the ligand is also dependent on the number of receptors on a given tissue. An additional ligand property is that of selectivity, the degree to which the ligand interacts with the target of choice as compared to related structural targets. The degree of selectivity typically determines the side effect profile of a new compound, given that the targeted mechanism itself does not produce untoward effects when stimulated beyond the therapeutic range. Ligands may be either agonists or antagonists. Agonists have intrinsic
efficacy and their binding to the receptor leads to activation of the intracellular components involved in the physiological or pharmacological responsiveness of cell or tissue. This efficacy may be manifested by changes in the activity of an enzyme like adenylate cyclase or by an alteration in the contractile response of an isolated, intact tissue preparation. Antagonists bind to the receptor and block the interaction of the agonist while producing no effect on the tissue on their own. Antagonism can be of several types: competitive, non-competitive and inverse. \(^1\) Competitive antagonism is usually associated with ligands that interact directly with the agonist binding site i.e. the recognition element of the receptor. The non-competitive or uncompetitive antagonists interact at sites distinct from the agonist recognition site and can modulate agonist binding. A third class of ligand is that of the inverse agonist. Ligands of this class interact with a defined recognition site on a receptor and are not only able to block the effects of an agonist at the receptor but, to varying degrees, are also able to produce effects opposite to that of the agonist. It is clear that a biological response is produced by the interaction of a drug with the biological receptor. This selective binding and its extent is governed by the molecular recognition phenomenon. In molecular modeling, this process of molecular recognition is simulated to understand the drug- receptor interaction. Most of the molecular recognition phenomena of ligand and receptor involve the following type of interaction:

\[
\begin{align*}
\text{Ligand} + \text{Receptor} & \quad \xrightarrow{k_1} \quad \text{L-R Complex} \quad \xrightarrow{k_2} \quad \text{Response} \\
& \quad \text{.................................................. (2)}
\end{align*}
\]

The rate constant for association of the complex is \(k_1\); the rate constant for dissociation of the complex is \(k_2\) and the affinity or association constant

\[
k_{\text{ass}} = \frac{k_1}{k_2}.
\]

The thermodynamic parameters of interest for the above reactions are standard free energy (\(G^0\)), enthalpy (\(H^0\)) and entropy (\(AS^0\)) of association. These parameters are related by the Gibbs free energy equation,

\[
\begin{align*}
\Delta G^0 &= -RT \ln k_{\text{ass}} \quad \text{.................................................. (3)} \\
\Delta G^0 &= \Delta H^0 - T \Delta S^0 \quad \text{.................................................. (4)}
\end{align*}
\]

The most fundamental forces involved in the interaction of ligand and receptor is covalent, reinforced ionic, ionic, ion-dipole, dipole-dipole, van der Waals and hydrophobic forces. In molecular modeling every effort is made to measure the free energy of
association ($\Delta G$). Various computational chemistry methods and assumptions are adopted to arrive at a measure of association.

## 1.3 Molecular Modeling and Computational Chemistry

The definition currently accepted of what molecular modeling is, can be stated as this: “molecular modeling is anything that requires the use of a computer to paint, describe or evaluate any aspect of the properties of the structure of a molecule”.\(^{13}\) Methods used in the molecular modeling arena regard automatic structure generation, analysis of three-dimensional (3D) databases, construction of protein models by techniques based on sequence homology, diversity analysis, docking of ligands or continuum methods. Thus, today molecular modeling is regarded as a field concerned with the use of all sorts of different strategies to model and deduce information of a system at the atomic level. On the other hand, this discipline includes all methodologies used in computational chemistry, like computation of the energy of a molecular system, energy minimization, monte Carlo methods or molecular dynamics. In other words, it is possible to conclude that computational chemistry is the nucleus of molecular modeling. Identification of bimolecular moieties involved in the interaction with a specific receptor permits to understand the molecular mechanism responsible of its specific biological activity. In turn, this knowledge is aimed at designing new active molecules that can be successfully used as drugs. Due to the fact that simulation accuracy is limited to the precision of the constructed models, when it is possible, computational simulations have to be compared with experimental results to confirm model accuracy and to modify them if necessary, in order to obtain better representations of the system.\(^{14}\)

## 1.4 Quantum Mechanics and Molecular Mechanics

There are two different approaches to compute the energy of a molecule. First, quantum mechanics, a procedure based on first principles. In this approach, nuclei are arranged in the space and the corresponding electrons are spread all over the system in a continuous electronic density and computed by solving the Schrödinger equation. When chemical reactions do not need to be simulated, classical mechanics can describe the behavior of a bimolecular system. This mathematical model is known as molecular mechanics, and can be used to compute the energy of systems containing a large number of atoms, such as molecules or complex systems of biochemical and biomedical interest. In contrast to quantum mechanics, molecular mechanics ignore electrons and compute the energy of a system only as a function of the nuclear positions. Then, it is possible to take into account
in an implicit way the electronic component of the system by adequate parameterization of the potential energy function. The set of equations and parameters which define the potential surface of a molecule is called force field.\textsuperscript{15}

1.5 Force Fields

In molecular mechanics the electrons and nuclei of the atoms are not explicitly included in the calculations. Molecular mechanics considers a molecule to be a collection of masses interacting with each other through harmonic forces. Thus, the atoms in molecules are treated as ball of different sizes and flavors joined together by springs of variable strength and equilibrium distances (bonds). This simplification allows using molecular mechanics as a fast computational model that can be applied to molecules of any size.

In the course of a calculation the total energy is minimized with respect to the atomic coordinates, and it consists of a sum of different contributions that compute the deviations from equilibrium of bond lengths, angles, torsions and non-bonded interactions:

\[
E_{\text{tot}} = E_{\text{str}} + E_{\text{bend}} + E_{\text{tors}} + E_{\text{vdw}} + E_{\text{elec}} + ... \quad (5)
\]

where \(E_{\text{tot}}\) is the total energy of the molecule, \(E_{\text{str}}\) is the bond-stretching energy term, \(E_{\text{bend}}\) is the angle-bending energy term, \(E_{\text{tors}}\) is the torsional energy term, \(E_{\text{vdw}}\) is the van der Waals energy term, and \(E_{\text{elec}}\) is the electrostatic energy term. The equilibrium values of bond lengths and bond angles are the corresponding force constants used in the potential energy function in the force field and it defines a set known as force field parameters. Each deviation from these equilibrium values will result in increasing total energy of the molecule. So, the total energy is a measure of intramolecular strain relative to a hypothetical molecule with an ideal geometry of equilibrium. By itself the total energy has no strict physical meaning, but differences in total energy between two different conformations of the same molecule can be compared.\textsuperscript{16-19}

1.6 Energy-Minimizing Procedures

Energy minimization methods can be divided into different classes depending on the order of the derivative used for locating a minimum on the energy surface. Zero order methods are those that only use the energy function to identify the regions of low energy through a grid search procedure. The most well-known method of this kind is the SIMPLEX method. Within first-derivative techniques, there are several procedures like the steepest descent method or the conjugate gradient method that make use of the gradient of the function. Second-derivative methods, like the Newton-Raphson algorithm make use of the hessian to locate minima.\textsuperscript{20,21}
1.6.1 Steepest Descent Method

In the steepest descent method, the minimizer computes numerically the first derivative of the energy function to find a minimum. The energy is calculated for the initial geometry and then again after one of the atoms has been moved in a small increment in one of the directions of the coordinate system. This process is repeated for all atoms which finally are moved to a new position downhill on the energy surface. The procedure stops when a predetermined threshold condition is fulfilled. The optimization process is slow near the minimum, and consequently, the steepest descent method is often used for structures far from the minimum as a first, rough and introductory run followed by a subsequent minimization employing a more advanced algorithm like the conjugate gradient.

1.6.2 Conjugate Gradient Method

The conjugate gradient algorithm accumulates the information about the function from one iteration to the next. With this proceeding the reverse of the progress made in an earlier iteration can be avoided. For each minimization step the gradient is calculated and used as additional information for computing the new direction vector of the minimization procedure. Thus, each successive step refines the direction towards the minimum. The computational effort and the storage requirements are greater than for steepest descent, but conjugate gradients is the method of choice for larger systems. The greater total computational expense and the longer time per iteration is more than compensated by more efficient convergence to the minimum achieved by conjugate gradients.\textsuperscript{22,23}

As a summary, the choice of the minimization method depends on two factors: the size of the system and the current state of the optimization. For structures far from minimum, as a general rule, the steepest descent method is often the best minimizer to use for 100-1000 iterations. The minimization can be completed to convergence with conjugate gradients.

There are several ways in molecular minimization to define convergence criteria. In non-gradient minimizers only the increments in the energy and the coordinates can be taken to judge the quality of the actual geometry of the molecular system. In all gradient minimizers, however, atomic gradients are used for this purpose. The best procedure in this respect is to calculate the root mean square gradients of the forces on each atom of a molecule. The value chosen as a maximum derivative will depend on the objective of the minimization. If a simple relaxation of a strained molecule is desired, rough convergence criterions like a maximum derivative of 0.1 kcal mol\textsuperscript{-1} Å\textsuperscript{-1} is sufficient while for other cases
convergence to a maximum derivative less than 0.001 kcal mol\(^{-1}\) Å\(^{-1}\) is required to find a final minimum.

### 1.7 Computer-aided Drug Design

Computer-aided drug design, often called structure based drug design involves using the biochemical information of ligand-receptor interaction in order to postulate ligand refinements. For example, if we know the binding site the steric complementarity of the ligand could be improved to increase the affinity for its receptor. Indeed, using the crystal structure of the complex we can target regions of the ligand that fit poorly within the active site and postulate chemical modifications that lower the energetic potential by making more negative van der Waals terms, thus improving complementarity with the receptor. In a similar fashion, functional groups on the ligand can be changed in order to augment electrostatic complementarity with the receptor.

When a target is selected for the design of new lead compounds three different situations can be faced regarding the amount of information of the system that is available: 1) the structure of the receptor is well known and the bioactive conformation of the ligand is not known, 2) only the bioactive conformation of the ligand is known and 3) the target structure and the bioactive conformation of the ligand are unknown (Fig. 1.3).

The best possible starting point is an X-ray crystal structure of the target site. If the molecular model of the binding site is precise enough, one can apply docking algorithms that simulate the binding of drugs to the respective receptor site, like Autodock.\(^{24}\) In the first step the program creates a negative image of the target site through the use of several atom probes that determine affinity potentials for each atom type in the substrate molecule at different points in a grid, place the putative ligands into the site and finally they evaluate the quality of the fit. The program will try a set of different conformers of the ligand in order to obtain the best disposition of the atoms of the molecule for maximizing the scoring function that quantifies ligand receptor interaction.

A different strategy for obtaining new lead compounds through rational drug design is the *de novo* design of ligands with the use of a builder program, like Ligbuilder.\(^{25}\) This program also determines the shape and the electrostatic properties of the binding site cavity through the use of several atom probes and then it combines from a library of chemical fragments those that better fill the cavity based on steric and electrostatic complementarity.
1.7.1 Structure-Based Computer-Aided Drug Design

Structure-based computer-aided drug design (SBDD) relies on the ability to determine and analyse 3D structures of biological molecules. The core hypothesis of this approach is that a molecule’s ability to interact with a specific protein and exert a desired biological effect depends on its ability to favourably interact with a particular binding site on that protein. Molecules that share those favorable interactions will exert similar biological effects. Therefore, novel compounds can be elucidated through the careful analysis of a protein’s binding site. Extensive use of biophysical techniques such as X-ray crystallography and NMR spectroscopy has led to the elucidation of a number of 3D structures of human and pathogenic proteins. For example, the PDB has over 81,000 protein structures, whereas data bases such as PDBBIND\(^{26}\) and protein ligand database house has 5,671 and 129 (as of 2003) ligand-protein co-crystal structures, respectively. Drug discovery campaigns leveraging target structure information have sped up the discovery process and have led to the development of several clinical drugs.

Table 1.1 Example of Marketed Drugs Involving use of Structure based Drug Design.\(^{27-31}\)

<table>
<thead>
<tr>
<th>Year</th>
<th>Generic Name</th>
<th>Brand Name</th>
<th>Manufacturer</th>
<th>Against / Inhibits</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989</td>
<td>Zanamivir</td>
<td>Relenza</td>
<td>GlaxoSmithKline</td>
<td>Neuraminidase</td>
</tr>
<tr>
<td>1997</td>
<td>Nelfinavir</td>
<td>Viracept</td>
<td>Hoffman-La Roche</td>
<td>HIV protease</td>
</tr>
<tr>
<td>1998</td>
<td>Raltitrexed</td>
<td>Tomudex</td>
<td>AstraZeneca</td>
<td>Thymidylate</td>
</tr>
<tr>
<td>1999</td>
<td>Amprenavir</td>
<td>Agenerase</td>
<td>GlaxoSmithKline</td>
<td>HIV protease</td>
</tr>
<tr>
<td>2007</td>
<td>Raltegravir</td>
<td>Isentress</td>
<td>Merck</td>
<td>HIV integrase</td>
</tr>
</tbody>
</table>
A prerequisite for the drug discovery process is the ability to rapidly determine potential binders to the target of biological interest. Computational methods in drug discovery allow rapid screening of a large compound library and determination of potential binders through modeling/simulation and visualization techniques.

### 1.7.1.1 Preparation of a Target Structure

Success of virtual screening depends upon the amount and quality of structural information known about both the target and the small molecules being docked. The first step is to evaluate the target for the presence of an appropriate binding pocket.\(^{32-33}\) This is usually done through the analysis of known target-ligand co-crystal structures or using in-silico methods to identify novel binding sites.\(^{34}\)

A target structure experimentally determined through X-ray crystallography or NMR techniques and deposited in the PDB is the ideal starting point for docking. Structural genomics has accelerated the rate at which target structures are being determined. In the absence of experimentally determined structures, several successful virtual screening campaigns have been reported based on comparative models of target proteins.\(^{35-37}\)

### 1.7.1.2 Homology Modeling

In the absence of experimental structures, computational methods are used to predict the 3D structure of target proteins. Comparative modeling is used to predict target structure based on a template with a similar sequence, leveraging that protein structure is better conserved than sequence, i.e., proteins with similar sequences have similar structures. Homology modeling is a specific type of comparative modeling in which the template and target proteins share the same evolutionary origin. Comparative modeling involves the following steps: (1) identification of related proteins to serve as template structures, (2) sequence alignment of the target and template proteins, (3) copying coordinates for confidently aligned regions, (4) constructing missing atom coordinates of target structure, and (5) model refinement and evaluation. Fig. 1.4 illustrates the steps involved in homology modeling. Several computer programs and web servers exist that automate the homology modeling process e.g., PSIPRED\(^ {38}\) and MODELER.\(^ {39}\)

### 1.7.1.3 Molecular dynamics-based detection

The dynamic nature of biomolecules sometimes makes it insufficient to use a single static structure to predict putative binding sites. Multiple conformations of target are often used to account for structural dynamics of target. Classic molecular dynamic (MD) simulations can be used for obtaining an ensemble of target conformations beginning with a single
structure. The MD method uses principles of Newtonian mechanics to calculate a trajectory of conformations of a protein as a function of time. Classic MD methods tend to get trapped in local energy minima. To overcome this, several advanced MD algorithms such as targeted-MD\textsuperscript{40}, conformational folding simulations\textsuperscript{41}, temperature accelerated MD simulations\textsuperscript{42}, and replica exchange MD\textsuperscript{43} have been implemented for traversing multiple minima energy surface of proteins.

**Figure. 1.4** Steps involved in homology model building process\textsuperscript{38-39}

1.7.1.4 **Small Molecules and Target Protein representation for Docking Simulations**

There are three basic methods to represent target and ligand structures \textit{in silico}: atomic, surface, and grid representations\textsuperscript{44,45}. Atomic representation of the surface of the target is usually used when scoring and ranking is based on potential energy functions. An example is DARWIN, which uses CHARMM force-field to calculate energy\textsuperscript{46}. Surface methods represent the topography of molecules using geometric features. The surface is represented as a network of smooth convex, concave, and saddle shape surfaces. These features are generated by mapping part of van der Waals surface of atoms that is accessible to probe a sphere (Connolly, 1983)\textsuperscript{47}. Docking is then guided by a complementary alignment of ligand and binding site surfaces. Earliest implementation of DOCK\textsuperscript{48} used a set of nonoverlapping spheres to represent invaginations of target surface and the surface of the
ligand. For the grid representation, the target is encoded as physicochemical features of its surface. A grid method described by Katchalski-Katzir et al.\textsuperscript{49} digitizes molecules using a 3D discrete function that distinguishes the surface from the interior of the target molecule. Molecules are scanned in relative orientation in three dimensions, and the extent of overlap between molecules is determined using a correlation function calculated from a Fourier transform.

1.7.1.5 Sampling Algorithms for Protein-Ligand Docking

Docking methods can be classified as rigid-body docking and flexible docking applications depending on the degree to which they consider ligand and protein flexibility during the docking process.\textsuperscript{50,51} Rigid body docking methods consider only static geometric/physiochemical complementarities between ligand and target and ignore flexibility and induced-fit\textsuperscript{51} binding models. More advanced algorithms consider several possible conformations of ligand or receptor or both at the same time according to the conformational selection paradigm.\textsuperscript{52} Rigid docking simulations are generally preferred when time is critical, i.e., when a large number of compounds are to be docked during an initial vHTS. However, flexible docking methods are still needed for refinement and optimization of poses obtained from an initial rigid docking procedure.

1.7.1.5.1 Systematic Methods

Systematic algorithms incorporate ligand flexibility through a comprehensive exploration of a molecule’s degrees of freedom. In systematic algorithms, the current state of the system determines the next state. Systematic methods can be categorized into (1) exhaustive search algorithms and (2) fragmentation algorithms. Exhaustive searches elucidate ligand conformations by systematically rotating all possible rotatable bonds at a given interval. Large conformational space often prohibits an exhaustive systematic search. Algorithms such as GLIDE\textsuperscript{53} use heuristics to focus on regions of conformational space that are likely to contain good scoring ligand poses. Fragmentation methods sample ligand conformation by incremental construction of ligand conformations from fragments obtained by dividing the ligand of interest. Ligand conformations are obtained by docking fragments in the binding site one at a time and incrementally growing them or by docking all fragments into the binding site and linking them covalently. FLEXX\textsuperscript{54} uses the “anchor and grow method” for ligand conformational sampling.

1.7.1.5.2 Molecular Dynamics Simulations

Molecular dynamics (MD) simulation calculates the trajectory of a system by the application of Newtonian mechanics. However, standard MD methods depend heavily on
the starting conformation and are not readily appropriate for simulation of ligand-target interactions. Because of its nature, MD is not able to cross high-energy barriers within the simulation’s lifetime and is not efficient for traversing the rugged hyper surface of protein-ligand interactions. Strategies like simulated annealing have been applied for more efficient use of MD in docking.\textsuperscript{55}

\subsection*{1.7.1.5.3 Monte Carlo Search with Metropolis Criterion (MCM) Simulations}
MCM samples conformational space faster than molecular dynamics in that it requires only energy function evaluation and not the derivative of the energy functions. Although traditional MD drives a system toward a local energy minimum, the randomness introduced with Monte Carlo allows hopping over the energy barriers, preventing the system from getting stuck in local energy minima. MCM simulations have been adopted for flexible docking applications such as in MCDOCKER.\textsuperscript{56}

\subsection*{1.7.1.5.4 Genetic Algorithms}
Genetic algorithms introduce molecular flexibility through recombination of parent conformations to child conformations. In this simulated evolutionary process, the “fittest” or best scoring conformations are kept for another round of recombination. In this way, the best possible set of solutions evolves by retaining favorable features from one generation to the next. In docking, a set of values that describe the ligand pose in the protein are state variable. State variables may include set of values describing translation, orientation, conformation, number of hydrogen bonds, etc. The state corresponds to the genotype; the resulting structural model of the ligand in the protein corresponds to the phenotype, and binding energy corresponds to the fitness of the individual. Genetic operators may swap large regions of parent’s genes or randomly change (mutate) the value of certain ligand states to give rise to new individuals. Genetic Optimization for Ligand Docking (GOLD)\textsuperscript{57} explores full ligand flexibility with partial target flexibility using a genetic algorithm.

\subsection*{1.7.1.6 Scoring Functions for Evaluation of Protein Ligand Complexes}
Docking applications need to rapidly and accurately assess protein-ligand complexes, i.e., approximate the energy of the interaction. A ligand docking experiment may generate hundreds of thousands of target-ligand complex conformations, and an efficient scoring function is necessary to rank these complexes and differentiate valid binding mode predictions from invalid predictions.

\subsection*{1.7.1.6.1 Force-Field or Molecular Mechanics-Based Scoring Functions}
Force-field scoring functions use classic molecular mechanics for energy calculations.\textsuperscript{58}
These functions use parameters derived from experimental data and \textit{ab initio} quantum
mechanical calculations. The binding free energy of protein-ligand complexes are estimated by the sum of van der Waals and electrostatic interactions. DOCK uses the AMBER force fields in which van der Waals energy terms are represented by the Lennard-Jones potential function while electrostatic terms are accounted for by coulombic interaction with a distance-dependent dielectric function.

1.7.1.6.2  **Empirical Scoring Functions**

Empirical scoring functions fit parameters to experimental data. An example is binding energy, which is expressed as a weighted sum of explicit hydrogen bond interactions, hydrophobic contact terms, desolvation effects, and entropy. Empirical function terms are simple to evaluate and are based on approximations. The weights for different parameters are obtained from regression analysis using experimental data obtained from molecular data. Empirical functions have been used in several commercially available docking suits like LUDI, FLEXX and SURFLEX.

1.7.1.6.3  **Knowledge-Based Scoring Function**

Knowledge based scoring functions use the information contained in experimentally determined complex structures. They are formulated under the assumption that interatomic distances occurring more often than average distances represent favorable contacts. On the other hand, interactions that are found to occur with lower frequencies are likely to decrease affinity. Several knowledge based potentials have been developed to predict binding affinity like potential of mean force, DRUGSCORE, SMOG and BLEEP.

1.7.1.6.4  **Consensus-Scoring Functions**

Consensus approaches rescore predicted poses several times using different scoring functions. These results can then be combined in different ways to rank solutions. Some strategies for combining scores include (1) weighted combinations of scoring functions, (2) a voting strategy in which cut-offs established for each scoring method is followed by decision based on number of poses a molecule has, (3) a rank by number strategy ranks each compound by its average normalized score values, and (4) a rank by rank method sorts compounds based on average rank determined by individual scoring functions.

1.7.1.7  **Structure-Based Virtual High-Throughput Screening**

Structure-based virtual high-throughput screening (SB-vHTS), the in silico method for identifying putative hits out of hundreds of thousands of compounds to the targets of known structure, relies on a comparison of the 3D structure of the small molecule with the putative binding pocket. SB-vHTS selects for ligands predicted to bind a particular binding site as opposed to traditional HTS that experimentally asserts general ability of a ligand to
bind, inhibit, or allosterically alter the protein’s function. To make screening of large compound libraries within finite time feasible. SB-vHTS often uses limited conformational sampling of protein and ligand and a simplified approximation of binding energy that can be rapidly computed. The key steps in SB-vHTS are: (1) preparation of the target protein and compound library for docking, (2) determining a favorable binding pose for each compound, and (3) ranking the docked structures.68

1.7.2 Ligand-Based Computer-Aided Drug Design

The ligand-based computer-aided drug discovery (LBDD) approach involves the analysis of ligands known to interact with a target of interest. These methods use a set of reference structures collected from compounds known to interact with the target of interest and analyse their 2D or 3D structures. The overall goal is to represent these compounds in such a way that the physicochemical properties most important for their desired interactions are retained, whereas extraneous information not relevant to the interactions is discarded. It is considered as an indirect approach to the drug discovery in that it does not necessitate knowledge of the structure of the target of interest. The two fundamental approaches of LBDD are (1) selection of compounds based on chemical similarity to known actives using some similarity measure or (2) the construction of a quantitative structure activity relationship (QSAR) model that predicts biological activity from chemical structure. The methods are applied for in silico screening for novel compounds possessing the biological activity of interest, hit-to-lead and lead-to drug optimization, and also for the optimization of DMPK/ADMET properties. LBDD is based on the similar property principle which states that molecules that are structurally similar are likely to have similar properties.69 LBDD approaches in contrast to SBDD approaches can also be applied when the structure of the biological target is unknown. Additionally, active compounds identified by ligand-based virtual high-throughput screening (LB-vHTS) methods are often more potent than those identified in SB-vHTS.70

1.7.2.1 Molecular Descriptors

Molecular descriptors can include properties such as molecular weight, geometry, volume, surface areas, ring content, rotatable bonds, interatomic distances, bond distances, atom types, planar and nonplanar systems, molecular walk counts, electronegativities, polarizabilities, symmetry, atom distribution, topological charge indices, functional group composition, aromaticity indices, solvation properties, and many others.71 These descriptors are generated through knowledge-based, graph-theoretical methods, molecular mechanical, or quantum-mechanical tools72,73 and are classified according to the
“dimensionality” of the chemical representation from which they are computed\textsuperscript{74}: 1-dimensional (1D), scalar physicochemical properties such as molecular weight; 2D, molecular constitution-derived descriptors; 2.5D, molecular configuration-derived descriptors; 3D, molecular conformation-derived descriptors. These different levels of complexity, however, are overlapping with the more complex descriptors, often incorporating information from the simpler ones.

1.7.2.2 Molecular Fingerprint and Similarity Searches

Molecular fingerprint-based techniques attempt to represent molecules in such a way as to allow rapid structural comparison in an effort to identify structurally similar molecules or to cluster collections based on structural similarity. These methods are fewer hypotheses driven and less computationally expensive than pharmacophore mapping or QSAR models. They rely entirely on chemical structure and omit compound with known biological activity, making the approach more qualitative in nature than other LBDD approaches.\textsuperscript{75} Additionally, fingerprint-based methods consider all parts of the molecule equally and avoid focusing only on parts of a molecule that are thought to be most important for activity. This is less error prone to overfitting and requires smaller datasets to begin with. Fingerprint methods may be used to search databases for compounds similar in structure to a lead query, providing an extended collection of compounds that can be tested for improved activity over the lead. In many situations, 2D similarity searches of databases are performed using chemotype information from first generation hits, leading to modifications that can be evaluated computationally or ordered for \textit{in vitro} testing.

1.7.2.3 Quantitative Structure-Activity Relationship Models

Quantitative structure-activity relationship (QSAR) models describe the mathematical relation between structural attributes and target response of a set of chemicals.\textsuperscript{76} Classic QSAR is known as the Hansch-Fujita approach and involves the correlation of various electronic, hydrophobic, and steric features with biological activity. In the 1960s, Hansch and others began to establish QSAR models using various molecular descriptors to physical, chemical, and biological properties focused on providing computational estimates for the bioactivity of molecules.\textsuperscript{77} In 1964, Free and Wilson\textsuperscript{78} developed a mathematical model relating the presence of various chemical substituents to biological activity (each type of chemical group was assigned an activity contribution), and the two methods were later combined to create the Hansch/ Free-Wilson method.

The general workflow of a QSAR-based drug discovery project is to first collect a group of active and inactive ligands and then create a set of mathematical descriptors that
describe the physicochemical and structural properties of those compounds. A model is then generated to identify the relationship between those descriptors and their experimental activity, maximizing the predictive power. Finally, the model is applied to predict activity for a library of test compounds that were encoded with the same descriptors. Success of QSAR, therefore, depends not only on the quality of the initial set of active/inactive compounds but also on the choice of descriptors and the ability to generate the appropriate mathematical relationship. One of the most important considerations regarding this method is the fact that all models generated will be dependent on the sampling space of the initial set of compounds with known activity, the chemical diversity. In other words, divergent scaffolds or functional groups not represented within this “training” set of compounds will not be represented in the final model, and any potential hits within the library to be screened that contain these groups will likely be missed. Therefore, it is advantageous to cover a wide chemical space within the training set.

1.7.2.3.1 3D-QSAR
Comparative field molecular analysis (CoMFA)\(^7\) is a 3D-QSAR technique that aligns molecules and extracts aligned features that can be related to biological activity. This method focuses on the alignment of molecular interaction fields rather than the features of each individual atom. CoMFA was established over 20 years ago as a standard technique for constructing 3D models in the absence of direct structural data of the target. In this method, molecules are aligned based on their 3D structures on a grid and the values of steric (van der Waals interactions) and electrostatic potential energies (Coulombic interactions) are calculated at each grid point. A comparative molecular similarity index (CoMSIA) is an important extension to CoMFA. In CoMSIA, the molecular field includes hydrophobic and hydrogen-bonding terms in addition to the steric and coulombic contributions. Similarity indices are calculated instead of interaction energies by comparing each ligand with a common probe and Gaussian-type functions are used to avoid extreme values.\(^7\) The overview of workflow of 3D-QSAR is represented in Fig. 1.5.

1.7.2.3.2 Multidimensional QSAR: 4D and 5D Descriptors
Multidimensional QSAR (mQSAR) seeks to quantify all energy contributions of ligand binding including removal of solvent molecules, loss of conformational entropy, and binding pocket adaptation.
Figure 1.5. Overview of 3D-QSAR and virtual screening workflow.

4D-QSAR is an extension of 3D-QSAR that treats each molecule as an ensemble of different conformations, orientations, tautomers, stereoisomers, and protonation states. The fourth dimension in 4D-QSAR refers to the ensemble sampling of spatial features of each molecule. A receptor-independent (RI) 4D-QSAR method was proposed by Hopfinger in 1997. This method begins by placing all molecules into a grid and assigning interaction pharmacophore elements to each atom in the molecule (polar, nonpolar, hydrogen bond donor, etc.). Molecular dynamics simulations are used to generate a Boltzmann weighted conformational ensemble of each molecule within the grid. Trial alignments are performed within the grid across the different molecules, and descriptors are defined based on occupancy frequencies within each of these alignments. These descriptors are called grid cell occupancy descriptors. A conformational ensemble of each compound is used to generate the grid cell occupancy descriptors rather than a single conformation.

5D-QSAR has been developed to account for local changes in the binding site that contribute to an induced fit model of ligand binding. In a method developed by Vedani and Dobler, induced fit is simulated by mapping a “mean envelope” for all ligands in a training set on to an “inner envelope” for each individual molecule. Their method involves several protocols for evaluating induced-fit models including a linear scale based on the
adaptation of topology, adaptations based on property fields, energy minimization, and lipophilicity potential. By using this information, the energetic cost for adaptation of the ligand to the binding site geometry is calculated.

Vedani from the Biographics Laboratory developed a receptor modeling concept, Quasar, based on 6D-QSAR that explicitly allows for the simulation of induced fit. Quasar concept, previously 3,4,5D extended to six dimensions allows for the simultaneous consideration of different solvation models which can be achieved explicitly by mapping parts of the surface area with solvent properties (position and size are optimized by the genetic algorithm).

### 1.7.2.4 Pharmacophore Mapping

In 1998, the International Union of Pure and Applied Chemistry (IUPAC) formally defined a pharmacophore as “the ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target structure and to trigger (or to block) its biological response”. In terms of drug activity, it is the spatial arrangement of functional groups that a compound or drug must contain to evoke a desired biological response. Therefore, an effective pharmacophore will contain information about functional groups that interact with the target, as well as information regarding the type of noncovalent interactions and interatomic distances between these functional groups/interactions.

A pharmacophore model of the target binding site summarizes steric and electronic features needed for optimal interaction of a ligand with a target. Most common properties that are used to define pharmacophores are hydrogen bond acceptors, hydrogen bond donors, basic groups, acidic groups, partial charge, aliphatic hydrophobic moieties, and aromatic hydrophobic moieties. Pharmacophore features have been used extensively in drug discovery for virtual screening, \textit{de novo} design, and lead optimization. A pharmacophore model of the target binding site can be used to virtually screen a compound library for putative hits. Apart from querying database for active compounds, pharmacophore models can also be used by \textit{de novo} design algorithms to guide the design of new compounds.

Structure-based pharmacophore methods are developed based on an analysis of the target binding site or based on a target-ligand complex structure. Ligand Scout uses protein-ligand complex data to map interactions between ligand and target. A knowledge based rule set obtained from the PDB is used to automatically detect and classify interactions into hydrogen bonds, charge transfers, and lipophilic regions. The algorithm
creates regularly spaced grids around the ligand and the surrounding residues. Probe atoms that represent a hydrogen bond donor, a hydrogen bond acceptor, and a hydrophobic group are used to scan the grids. An empirical scoring function, SCORE, is used to describe the binding constant between probe atoms and the target. SCORE includes terms to account for van der Waals interactions, metal-ligand bonding, hydrogen bonding, and desolvation effects upon binding. A pharmacophore model is developed by rescoring the grids followed by clustering and sorting to extract features essential for protein-ligand interaction.

The most common software packages used for ligand based pharmacophore generation include Phase, MOE, Catalyst, DISCO, and GASP.

1.8 Database Searching

The pharmacophores obtained from similarity analysis and 3D-QSAR analysis can be used to search the compounds from database having same features as defined in the pharmacophores. Whereas QSAR focuses on a set of descriptors like electrostatic and thermodynamic properties, pharmacophore mapping is a geometric approach. There are various programs like UNITY, CATALYST, MENTHOR, MACCS-3D, CAVEAT that convert these pharmacophores into search queries. Various databases available commercially are Comprehensive Medicinal Chemistry-3D (CMC-3D), Fine Chemical Directory-3D (FCD-3D), National Cancer Institute (NCI), Maybridge, Derwent World Drug Index, Biobyte etc. These search queries can be combined with ORACLE program to perform rational database search to arrive at a potential molecule with drug-like properties.

1.9 de novo Drug Design

With increased understanding of the drug-receptor theory along with thermodynamics of binding it is now possible to design new molecules from scratch. This methodology will allow designing molecules belonging to newer classes. This method, coupled with docking algorithms, gives a powerful tool for new molecule discovery. There are various methods available for de novo design but the basic principle involved in these methods is quite similar. The work by Bohm has led to program LUDI. Other methods in de novo design are Group Build, SMOG, MCSS & LeapFrog.
1.10 Aim of the Present Work

The present work involves the application of computer-aided drug design, involving Quantitative Structure-Activity Relationship (QSAR), Pharmacophore generation, Molecular modeling methods to design and develop New Chemical Entities (NCEs) as anti-inflammatory agents. The work also involves synthesis of NCEs and determination of their activity by *in vivo* pharmacological model such as cargenanan induced rat paw edema model. In this study, applications of these methodologies have been used to develop new potential anti-inflammatory agents.

Thus, the research work presented in the thesis comprises both experimental and computational medicinal chemistry aimed to find new chemical entities of therapeutic interest. The details of research work carried out are presented in this thesis.

During this entire work, we have extensively utilized computational chemistry tools. The molecular modeling software’s SYBYL-X 2.0, Schrödinger’s Phase 3.4 and Glide 5.8 licensed to Prof. V. M. Kulkarni, were used to design potential therapeutic agents.
1.11 References


91. SYBYL/ UNITY4.0, Tripos Inc., 1699, S. Hanley Road, St. Louis, MO 63144-2913, USA.

92. CATALYST, Accelrys, San Diago, USA.


103. SYBYL, Ligand-Based Design Manual, Tripos Inc., St. Louis, MO.