Chapter 2. Computational Approaches for the Study of Enzymes and their Reactions

A series of computational approaches can be used to overcome the limitations of experimental techniques in studying enzymes and their mechanisms. Identifying the most critical energetic components of the reaction is of crucial importance for understanding enzymatic catalysis. Several algorithms of varying accuracy and computational cost have been developed to study enzymatic catalysis. From ligand binding and activation to the catalytic process itself, including formation of the transition state and final generation of products, different modelling and simulation techniques can be applied.

Selection of the appropriate computational tool is not straightforward; it is one of the first challenges a computational chemist must face. Complete understanding of enzymatic reactions requires study of processes that take place over a variety of scales of time and distance. Therefore, no single simulation technique will be adequate to address all events occurring during catalysis, and it will be necessary to use a range of theoretical approaches [8,39,132-135]. In this chapter a brief summary of the main computational techniques used is presented, describing the underlying principles of the approaches, their applications and limitations.

2.1 Docking

Docking techniques, designed to find the correct conformation of a ligand and its receptor, have now been used for decades (for recent reviews and comparisons see [136-141]. The accurate prediction of the binding mode of a ligand within a protein receptor is of great importance in modern structure-base drug design [140], where docking is often used in virtual screening methods [139,142-146] to reduce large virtual libraries of compounds to a meaningful subset, which includes molecules with high binding affinities for the target receptor (Figure 2-1).

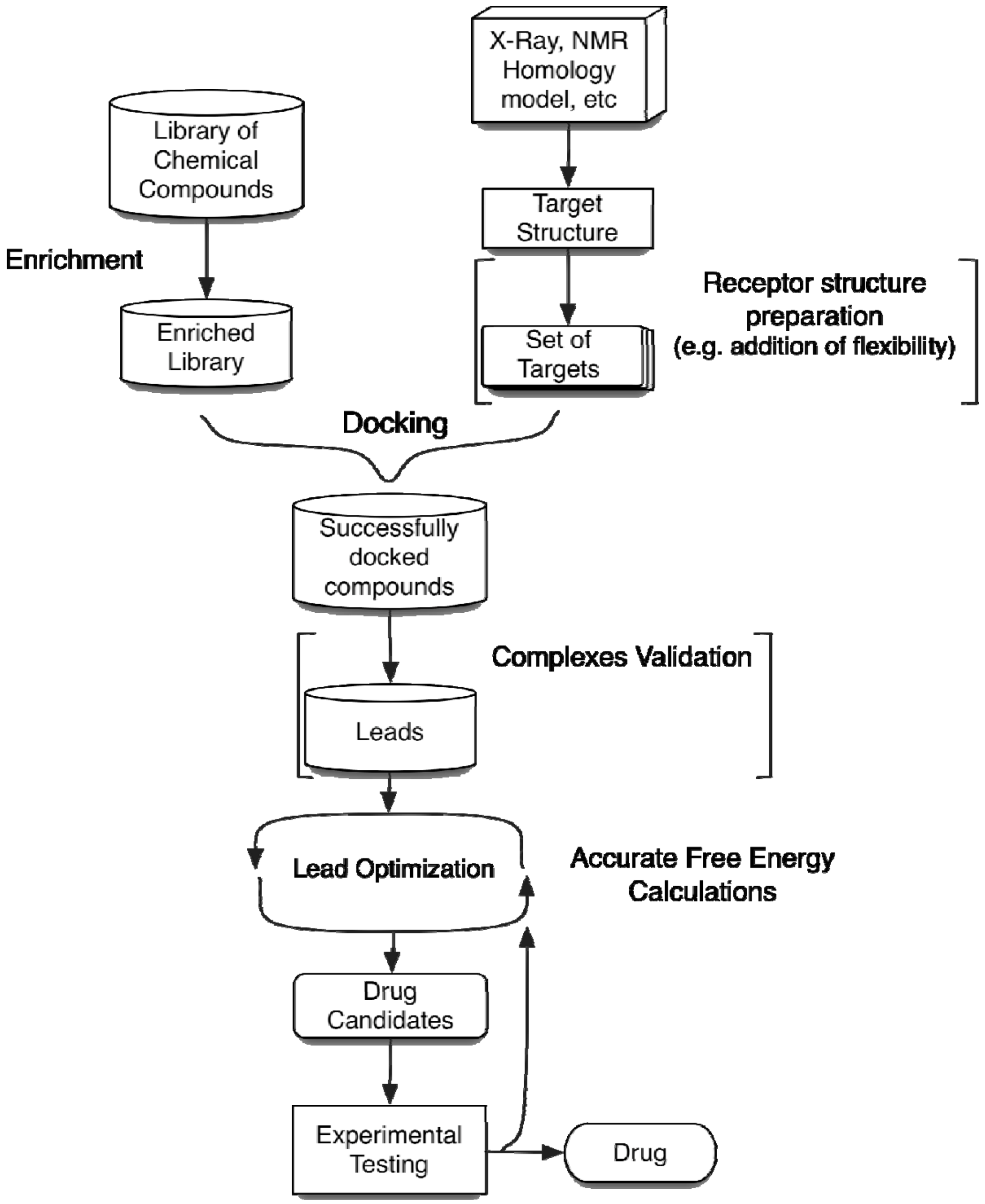

Figure 2-1 Drug design process.

Schematic representation of the protocol commonly followed during a drug-design process, when the structure of the protein target is known or can be modelled. Steps within square brackets are not always performed.

The process of binding a small molecule to its protein target is not simple; several entropic and enthalpic factors influence the interactions between them. The mobility of both ligand and receptor, the effect of the protein environment on the charge distribution over the ligand [147], and their interactions with the surrounding water molecules, further complicate the quantitative description of the process.

The three dimensional (3-D) structure of both ligand and protein are usually necessary for the application of docking techniques. While the manifold of conformational structures of small molecules may be relatively easy to predict, the lowest energy conformation obtained may not correspond to that of the bound ligand. The structures of proteins, on the other hand, present a bigger challenge. Although experimental techniques involving X-ray and NMR analysis are now routinely used, inherent difficulties in the preparation of samples and data collection and interpretation mean we are still far from a complete automated and high-throughout process [148]. Many proteins targeted for drug design do not have an experimentally determined structure and, therefore, docking studies cannot be performed directly. In some cases, computational techniques can be used to predict the 3-D structure of a protein provided the structure of a closely related protein homologue is known. Homology modelling [149,150] or sequence threading techniques may be used to generate models of protein structures [151] which, although not as good as experimentally determined structures, can be used as docking targets [152-157]. There are docking algorithms specifically designed for modelled structures [156].

2.1.1 Programs and Algorithms

For molecular docking programs to be useful tools, accuracy and speed are key factors. The docking problem involves both efficient searching algorithms to comprehensively explore all possible binding modes between the ligand and receptor, and accurate scoring functions which can efficiently and effectively discriminate between the experimental and the non-native solutions (for recent reviews and comparisons see [137,158-161]).

Ideally, the search procedure would explore all six degrees of translational and rotational freedom of the ligand together with the internal conformational degrees of freedom of both the ligand and the protein [141]. However, due to the size and complexity of the search space, and the time and computer limitations, such an approach is impractical. Therefore, in order to sample such a large conformational space the problem is simplified by applying constraints and approximations to reduce the dimension of the problem and locate the global minimum as efficiently as possible.

The early docking algorithms were limited to treatment of both the ligand and the target as rigid bodies, exploring only the six degrees of translational and rotational freedom and based mainly on geometric complementarity of the ligand and the active site [162-164]. The first example of such a program was DOCK, created by Kuntz and coworkers [163].

Until recently, although more flexible algorithms allowed significant mobility of the ligand, the protein receptor was usually held rigid, or with only a limited extent of flexibility at the active site. But in the last few years new algorithms and advances in computer power have allowed for a more flexible treatment of the complete system. The most popular docking programs nowadays include DOCK [163,165], AutoDock [166], FlexX [167], GOLD [168], and GLIDE [169,170] among others.

2.1.2 Ligand Flexibility

Although much smaller than proteins, ligands are often very flexible molecules with the number of possible conformations increasing in proportion to the power of the number of rotatable bonds. Ligand flexibility algorithms can be divided into three types of searches: systematic, stochastic, and deterministic. In systematic algorithms, a combinatorial approach is used to explore all the degrees of freedom in a molecule; an example is the anchor-and-grow or incremental construction algorithm. Stochastic search algorithms make random changes, usually changing one degree of freedom of the system at a time; examples include Monte Carlo (MC) methods and evolutionary algorithms. In deterministic searches, the initial state determines the move that can be made to generate the next state, which generally has to be equal or lower in energy; examples are energy minimizations and molecular dynamics (MD) simulations. For comparative studies of different algorithms for flexible molecular docking see [171,172].

2.1.2.1 Fragment-Based Methods or Incremental Construction Algorithms

The ligand is generally divided into an “anchor” rigid region and flexible parts. Usually, the anchor is docked first with the flexible parts added sequentially, with a systematic scanning of the torsional angles (Figure 2-2).

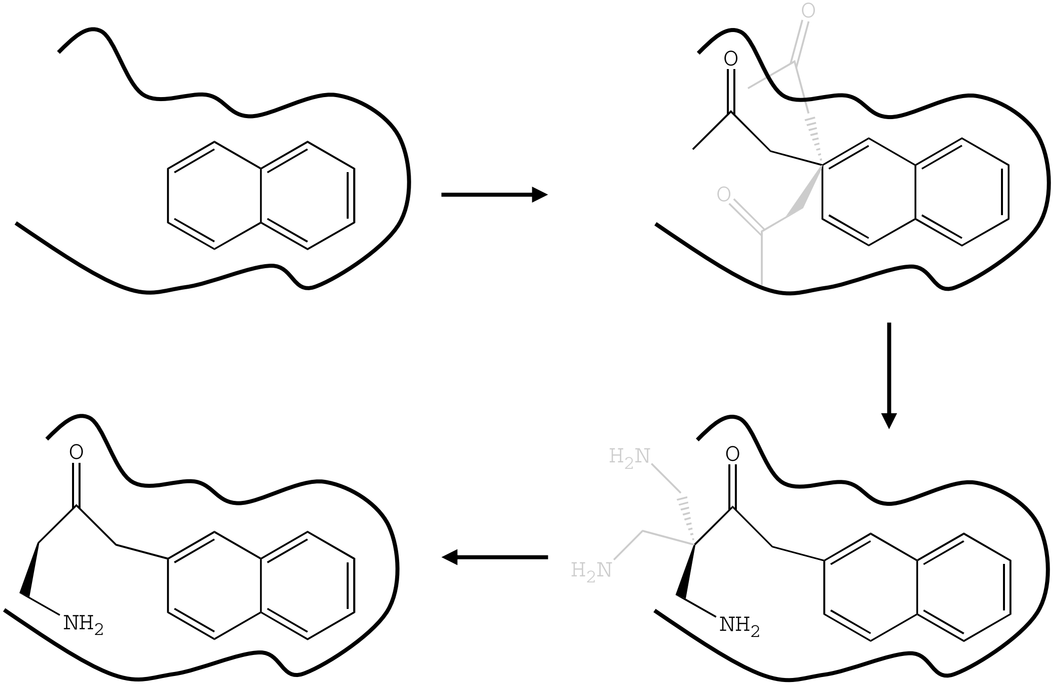

Figure 2-2 Docking with an incremental construction algorithm.

A selected rigid portion of the ligand or “anchor” region is initially docked; followed by the systematic addition of flexible parts. Only the most stable conformations of the partially grown ligand are kept, and used in subsequent rounds of growth and selection.

The first incremental construction algorithm [162] was implemented within the DOCK program [165]. The anchor fragment is docked based on steric complementarity, followed by the incremental addition of flexible regions. For each new addition, dihedral angles are explored and energy minimizations are carried out on the partial structure of the growing ligand [173].

The most popular program that performs fragment docking is FlexX [167,174], which docks the anchor base by modelling chemical interactions instead of steric complementarity. The choice of the anchor fragment is crucial as it is the first portion of the ligand that is docked, and it should contain all key interactions with the receptor. Next, flexible fragments are added incrementally, in all possible positions and conformations. The placements are ranked, clustered, and the highest ranked solutions are used in the next iteration. The process is repeated until the complete ligand is built, and the final structures are scored using an empirical scoring function. In a recent extension of the program [175], the possibility of exploring the placement of potential water molecules during the docking process has been incorporated.

Other docking programs that use a fragment approach for molecular docking include ADAM [176], Hammerhead [177], and SLIDE [178].

2.1.2.2 Monte Carlo Methods

Monte Carlo (MC) methods involve the application of random translational and rotational changes to the ligand, as well as torsional movements. After each move, the structure is usually minimized and the change accepted or rejected according to a Boltzmann probability. Early implementations of AutoDock [179,180] used Metropolis MC simulated annealing with a grid-based evaluation of the energy to dock flexible ligands into the binding pocket of a rigid receptor. Each simulation cycle was done at a lower temperature than the previous one, producing an annealing effect [179].

PRODOCK [181] and a method by Caflisch and coworkers [182] use an MC minimization technique to dock flexible ligands into flexible binding sites. PRO_LEADS [183] implements a “tabu search”. Here, the conformational space of the ligand is explored though random moves, as in MC, and a record is kept of the conformations already sampled. When a new solution is not lower in energy than the previous ones, it is only kept if it is not similar to anything in the “tabu” list [184].

2.1.2.3 Evolutionary Algorithms

The basic principle behind genetic algorithms (GA) is the evolution of a population of possible solutions via genetic operators, including mutations and crossovers, to a final population. Only the best solutions are carried on to the next generation, as determined by a predefined fitness function. The degrees of freedom (e.g. translation, rotation, internal motions) are encoded into “genes”, often represented as binary sub-strings, and the collection of “genes” is assigned a fitness value. The mutation operator randomly changes the value of a “gene”, and the crossover exchanges a set of “genes” between two candidate solutions. The population size, mutation rates, crossover rates, and number of evolutionary rounds are all parameters that can influence the final results.

GOLD [168] uses the GA search strategy, and multiple subpopulations of the ligand, rather than a single large one, are evolved at the same time. Migration of members from one subpopulation to another is allowed, and has been shown to increase the efficiency of the docking process.

AutoDock 3.0 [166] uses a Lamarckian genetic algorithm (LGA). This approach switches between “genotypic space” and “phenotypic space”. Mutations and crossover occur in genotypic space, while phenotypic space is determined by the energy function to be optimised. Energy minimization is performed after genotypic changes have been made to explore the local space. The phenotypic changes resulting from the energy minimization are then mapped back onto the genes. The LGA approach has been shown to outperform other simpler applications of the GA [166].

2.1.2.4 Pregenerated Conformational Libraries

In this approach, the cost of generating multiple conformers per ligand molecule is incurred only once, saving time in future studies. Moreover, the internal energy of each pre-generated conformer can be pre-calculated before it is docked rigidly into the receptor. This approach has been implemented in DOCK [185], FLOG [186], and EUDOC [187].

2.1.2.5 Implicit Ligand Flexibility

The flexibility of the ligand is taken into account indirectly, by allowing for overlap between the protein and the docked molecule through a “soft” scoring function [188]. Rigid docking programs based on surface complementarity include FTDOCK [189], LIGIN [190] and SANDOCK [191].

2.1.3 Protein Receptor Representation and Flexibility

Although all atoms in the ligand molecule are usually treated explicitly, there are three different basic representation forms for the protein receptor: atomic, surface, and grid [138]. Due to its computational cost, an atomic representation of the receptor is often restricted only to the final ranking procedures, when a potential energy function is used to characterize the complex. Surface representations are commonly used in protein-protein docking [192,193], based on the original work of Connolly on molecular surface representations [194,195]. However, the most common way of representing the protein receptor during small-ligand docking is using grids. The use of potential energy grids was pioneered by Goodford [196]; in this method the receptor's energy contributions are stored on grid points that need only to be read during ligand scoring.

Several studies have highlighted the importance of the conformation of a protein receptor for docking analysis [155,183,197]. The level of enrichment attained during a docking screening process decreases from the *holo* to the *apo* to the modelled structure of a protein, as the conformation of the receptor is less prepared to accommodate the ligand [155]. Sometimes, the conformation of the *holo* protein is such that only molecules structurally similar to that present in the original crystal-structure determination are recognized as potential ligands [183]. In the case of *apo* structures, their conformations may be inadequate to accommodate a ligand, due to wrongly positioned residues or the presence of loops blocking access to the binding site. Modelled molecules, even those modelled with a template of high sequence identity, can have badly placed side chains or missing loops or residues, hindering the docking process. In the latter case, it has been reported that the use of multiple homology models constructed from different crystal structures could provide a better representation of the protein receptor and improve the docking [156].

When receptor flexibility is included during the docking process the risks associated with ill-prepared conformation of the protein target are reduced [198-200]. Although originally restricted to the docking of rigid ligands into rigid receptors, recent advances in docking algorithms have allowed incorporation of ligand flexibility and, to less extent, protein mobility [168,181,182,201], during the docking procedure (Figure 2-6). The size and complexity of proteins makes it difficult to fully account for their mobility during a docking process and, therefore, its treatment is usually restricted to selected residues.

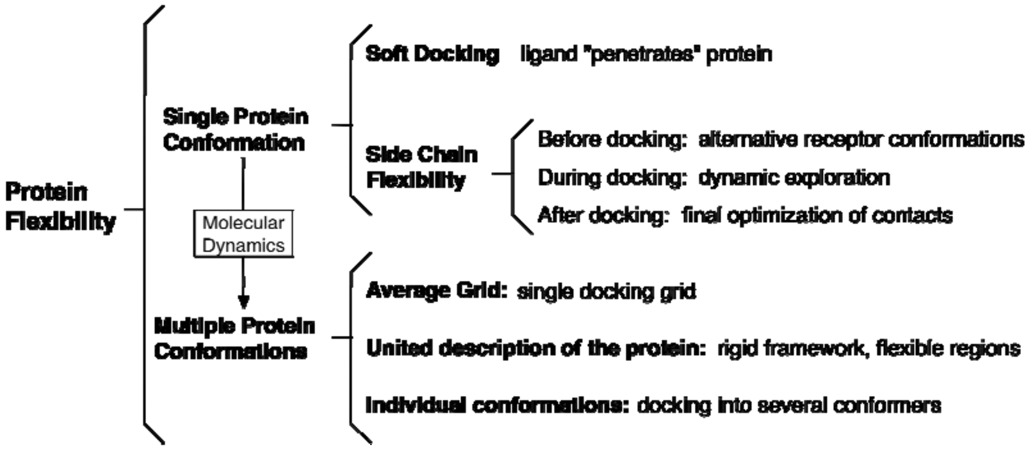
Protein Flexibility
Single Protein Conformation
Molecular Dynamics
Multiple Protein Conformations
Soft Docking ligand "penetrates" protein
Side Chain Flexibility
Before docking: alternative receptor conformations
During docking: dynamic exploration
After docking: final optimization of contacts
Average Grid: single docking grid
United description of the protein: rigid framework, flexible regions
Individual conformations: docking into several conformers

Figure 2-3 Protein flexibility.

Approaches that can be used during docking studies to incorporate flexibility to the protein receptor, at least partially.

2.1.3.1 Soft Docking

The simpler approaches deal with protein flexibility in an indirect way. Despite treating the receptor as a rigid object, the repulsive terms of the Lennard-Jones potential can be attenuated by generating a “soft” interaction. Thus, the ligand is allowed to “penetrate” the protein surface to some extent and to account for small and localized changes that would take place in a flexible environment [202-204] (Figure 2-4).

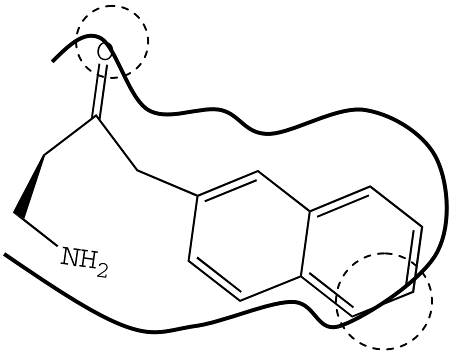

Figure 2-4 Soft docking.

During a soft docking approach, the ligand is allowed to “penetrate” the protein surface (regions within dashed circle). The implementation of a less repulsive Lennard-Jones potential allows, indirectly, accounting for small conformational changes of the enzyme upon ligand binding.

Although this approach does not increase computational costs, the changes in protein conformation that can be accounted for are minimal. While a “soft” scoring function usually performs better than a “hard” one when a single configuration of the receptor is used, use of the “hard” function is recommended when multiple conformations of the receptor are considered [205].

2.1.3.2 Sidechain Flexibility

In a different and more comprehensive approach, the mobility of some residues, particularly those within an enzyme active site, can be treated explicitly either during the docking process or after the ligand has been approximately placed [178,206]. As an example, the docking algorithm SLIDE [178] incorporates side-chain mobility after initial positioning of the ligand, allowing rotations of single bonds of non-anchor regions of the ligand and protein side chains.

Side-chain flexibility within the binding site can also be accounted for by pre-generating an ensemble of protein conformers using a rotamer library. Leach [207] was one of the first to introduce such an approach, with one of the major original limitations being the use of discrete pre-determined conformations of both the ligand and side chains. Increasing computer power allowed for improvements, and rotamer libraries have been used recently for the successful docking of the synthetic inhibitor RS-104966 of human collagenase-1 [208] and a series of inhibitors of tyrosine phosphatase B1 [209].

Although consideration of side-chain flexibility increases the computational cost of the docking process, it allows localized protein movement and results in improved fit of the ligand. As only the side chains of selected residues are allowed to move, important changes in the protein backbone, such as those involved in loop movements, are not taken into account.

2.1.3.3 Combined Protein Grid

Several alternative structures of a protein receptor can be combined into a single representation of the ensemble to account for bigger conformational changes that may be critical for the binding process. The averaging can be done over atom coordinates, to generate a final average structure, or over the grid representation of all receptor conformations, to produce an average docking grid.

These grids [196], or pre-calculated two-body potentials, are usually focused around the binding site and are used during the docking process to determine the interaction energy of different conformations of the ligand and the active site, in a fast and computationally inexpensive way. Different grids produced from several conformations of the receptor can be combined into a single global grid using a simple average-weighted scheme, or a differential weighting scheme which favours the contribution of some conformations over others, and which usually provides better results [210-212]. These ensemble-based grids minimize the effect of steric clashes between particular conformations of receptor and ligand, allow the establishment of more favourable interactions, and offer a relatively inexpensive approach for considering receptor conformational variability during the docking process.

2.1.3.4 United Description of the Receptor

The docking program FlexE [201] implements a different solution to the protein flexibility problem. Instead of combining different conformations of the protein receptor into a single docking grid, a united protein description of the target is created (Figure 2-5). Alternative protein conformations are superimposed and a rigid average structure is constructed from the most conserved structural features. For the variable regions, different conformations are explicitly considered and retained as an ensemble, which is combinatorially explored during the docking process to generate novel protein structures. Although this approached proved to give very good results [201], it was found that the scoring function implemented in FlexX needs to be revised to accurately identify the best complexes.

In a similar approach, Wei *et al.* [213] used a modified version of the program DOCK [163] to incorporate receptor flexibility during the docking process. In this case, the interactions between a given configuration of the ligand and different flexible parts of the receptor were calculated independently. Finally, those flexible regions that presented the best interaction energies with the ligand were recombined into a final representation of the protein receptor. One of the most significant observations of this study was the importance of considering the receptor conformational energy during the final ranking. When this energy was ignored, many known ligands ranked poorly due to the presence of "decoys" that could complement better some high-energy conformations of the receptor.

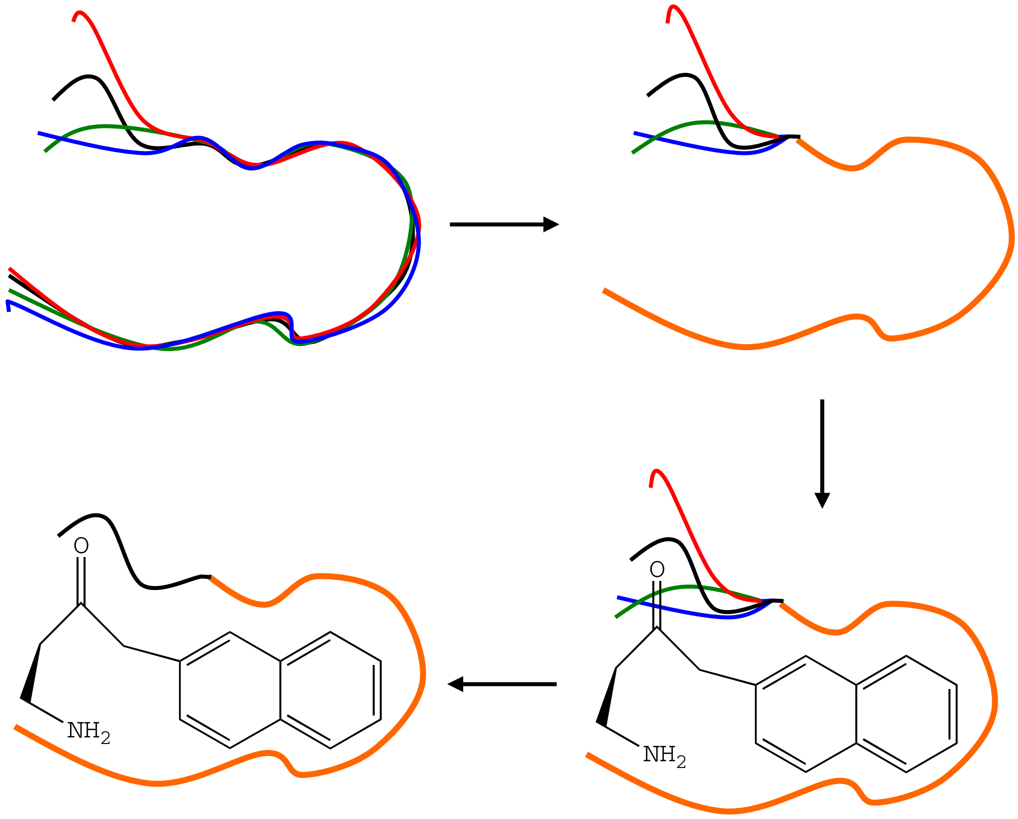

Figure 2-5 United description of the protein receptor.

Several conformations of the receptor are superimposed, and a consensus structure is built from the conserved areas. Those regions of greater variability are included as an ensemble of structures, which are systematically explored during the docking process. Finally, those protein conformations that present the best interactions with the docked ligand are combined into a final representation of the receptor.

2.1.3.5 Docking into Several Individual Protein Conformations

Docking the ligand against each protein structure in the ensemble constitutes the most comprehensive, although expensive, approach. While this strategy is not a realistic option for the virtual screening of a large library, it is a valid approach for difficult docking problems where even minor conformational changes of the receptor are expected to have a major influence on the binding process, and takes into account the possibility that a ligand may bind to only a few conformations of the receptor.

Carlson *et al.* [214] developed "dynamic" pharmacophore models of HIV-1 integrase using several snapshots from an MD simulation. They showed that the composite model was able to accommodate known inhibitors while the single crystal structure failed to do so [215]. In the relaxed-complex scheme of McCammon and coworkers [216], a long MD simulation of the *apo* receptor is first conducted to sample its conformational space, followed by the rapid docking of mini-libraries of candidate inhibitors against a large ensemble of snapshots. Schames *et al.* [217] discovered a novel binding trench in HIV-1 integrase by docking the 5CITEP inhibitor to snapshots of a 2 ns trajectory. The discovery of this new binding trench would not have been possible without the initial MD simulations of the receptor.

Although docking against several structures of the protein increases the chances of finding a receptor in the right conformational state to accommodate a particular ligand, it also reduces the selectivity of the docking process. A more relaxed representation of the protein will be able to accommodate a wider variety of ligands, many of which might be false positives. It is important, therefore, to use accurate scoring functions during the final screening process to maximise selection of the most active ligands.

2.1.4 Scoring Functions

A search algorithm may produce an immense number of solutions, and the purpose of the scoring functions is to distinguish the experimental binding modes from all the other modes explored during the search. Estimating binding free energies accurately is a time-consuming process, and although free energy perturbation (FEP) and thermodynamic integration (TI) calculations have been shown to provide very accurate results, they are computationally far too expensive to be used in the normal screening of large ligand libraries.

Scoring functions implemented within docking programs are not sufficiently accurate to identify, in every case, the most stable conformation of a given ligand or drug with the highest binding affinity among a set of compounds (for recent reviews and comparisons see [218-220]. Although some algorithms are able to rank the correct solution within the top 100 or even within the top 10 poses for some predictive docking cases [221-225], for most complexes the highest ranked structures are false positives [226]. It has also been shown that the performance of docking algorithms strongly depends on the characteristics of the target protein structure and the properties of the active site [159,161].

Scoring functions can be divided into four main classes according to the type of algorithm they use to estimate binding energies: first principles methods, semiempirical scoring functions, empirical methods, and knowledge-based potentials [227,228].

2.1.4.1 First Principles Methods.

These scoring methods use energy potentials similar to those found in molecular mechanics force fields. DOCK [229] and AutoDock [179] implement this kind of scoring function. To account for the screening effect of the solvent on electrostatic interactions, a distance-dependent dielectric constant is used. Internal ligand energies and entropic terms are completely ignored. Force field-based scoring functions ignore most solvent effects as well as solute entropies, and the calculated scores are just energies or enthalpies rather then free energies. The effects of solvent on protein-ligand interactions were later on incorporated into DOCK using implicit solvent models [185].

2.1.4.2 Semiempirical Methods.

One example of these methods is the linear interaction energy (LIE) approach of Åqvist and coworkers [230]. In this method, only the free ligand in solution and the ligand bound to the receptor are simulated using common molecular mechanics force fields. The final binding energy is obtained from the difference in electrostatic and van der Waals energy of the ligand in water and in the protein environment (see section 2.5.1.2 for further details). The GOLD scoring function is another example [168], where the energy is calculated based on a hydrogen bonding term based on empirical values, and van der Waals and internal energy terms calculated from molecular mechanics approaches. The docking program AutoDock also implemented a semiempirical function [166], where the force field terms are scaled by empirically determined constants, and special terms account for hydrogen bonding and desolvation effects.

2.1.4.3 Empirical Methods.

The idea behind empirical scoring functions is that binding energies can be approximated by a sum of individual uncorrelated terms, as first proposed by Böhm [231]. Each term is scaled by a coefficient obtained from regression analysis of empirical binding energies and X-ray structure information. As these scoring functions are based on a limited number of receptor-ligand structures, accurate scores are only expected for complexes that present similar interactions to those found within the training set. Moreover, it is difficult to know what each term and the error associated with it, exactly accounts for. Although the experimental conditions under which binding-constant affinities are measured (e.g. pH, salt concentration, temperature, etc) significantly affect the final values, these effects are generally ignored during the calculations, further increasing the possibility of errors. The first empirical scoring function was implemented in LUDI [232], and a similar function was included in the docking program FlexX [167].

2.1.4.4 Knowledge-Based Potentials.

Also known as potentials of mean force (PMF), knowledge based potentials are derived from known structures of protein-ligand complexes. They are designed to reproduce experimental structure rather than binding energies. A number of atom-type interactions are defined depending on their molecular environment, and their frequencies are converted into free energies using Boltzmann distributions. Their main difference from empirical potentials is that no binding data are needed. In theory, PMF include all forces that play a role in complex formation, although some effects, such as solvent effects, may be underestimated [136]. FlexX implements DrugScore [233], a smooth-grained PMF where the potentials were derived over short distances only, to ensure specific interactions dominate.

Given the inaccuracies of current scoring functions, a recent trend is to use consensus-scoring schemes [234]. In this approach, the information from different scoring functions is combined into a consensus one, balancing individual errors and improving the probability of identifying ‘true’ ligands. Charifson *et al.* [234], showed that consensus scores not only perform better in discriminating between active and inactive enzyme inhibitors, but they also result in a smaller number of false positives. In a recent work of Wang *et al.* [235], the ability of eleven functions to reproduce experimentally determined structures and binding affinities of 100 protein-ligand complexes were compared. It was found that the best results were obtained after combining two or three scoring functions into a consensus-scoring scheme. A similar conclusion was drawn in several other studies [145,236].

Lack of a reliable method for quickly locating correct solutions is the major obstacle in using predictive docking for practical applications [138]. Although library-screening processes require fast and inexpensive scoring functions, more accurate and expensive calculations can be employed in the last stages of a docking process, when only a few possible candidates are left, or during lead optimisation [237]. MD-based methods are among the most accurate current techniques available for the calculation of free energies (see section 2.5).

2.1.5 Refinement of Docked Complexes

A two-step protocol seems to be the most practical and convenient approach to obtain an adequate representation of the substrate-enzyme complex. Fast and less accurate docking algorithms are first used to scan large databases of molecules and reduce their size to a reasonable number of hits. Then, more accurate but time-consuming MD methods can be applied to refine the conformation of the complexes and produce accurate free energies [237,238].

MD simulations present an attractive approach for structural refinement of the final docked complexes. They incorporate flexibility of both ligand and receptor, improving interactions and enhancing complementarity between them, and thus accounting for induced fit [239]. Moreover, the time-dependent evolution of the system during the simulation provides a dynamic picture of the complex and helps to discriminate the correctly docked conformations from the unstable ones. Incorrectly docked structures are likely to produce unstable trajectories, leading to the disruption of the complex, while realistic complexes will show stable behaviour. In addition, the ability to incorporate explicit solvent molecules and their interactions is very important for understanding the role of water and its effect on the stability of the ligand-protein complexes.

Published studies where MD simulations have been applied after docking calculations to refine and analyse ligand-protein interactions include the work of Park H *et al.* [240] on the differential inhibition of two cyclin-dependent kinases, the study of propidium within human acetylcholinesterase (HuAChE) [241], and the binding of D-glucose onto the surface of insulin [242], and my own work on DfrB dihydrofolate reductase [243] (see Chapter 3), among others.

2.2 MD simulations

MD simulations are one of the most versatile and widely applied computational techniques for the study of biological macromolecules [81,244-247]. They are very valuable for understanding the dynamic behaviour of proteins at different timescales, from fast internal motions to slow conformational changes or even protein folding processes [248]. It is also possible to study the effect of explicit solvent molecules on protein structure and stability, to obtain time-averaged properties of the biomolecular system, such as density, conductivity, and dipolar moment, as well as different thermodynamic parameters, including interactions energies and entropies. MD is useful not only for rationalizing experimentally measured properties at the molecular level, but it is well known that most structures determined by X-ray or NMR methods have been refined using MD methods [249]. Therefore, the interplay between computational and experimental techniques in the area of MD simulations is longstanding, with the theoretical methods assisting in understanding and analysing experimental data. These, in turn, are vital for the validation and improvement of computational techniques and protocols.

Although the first protein MD simulation (bovine pancreatic trypsin inhibitor; 58 residues and ~450 atoms) was done *in vacuo* and for only 9.2 ps [250], enormous increases in computer power nowadays permit simulations of systems comprising 10^4-10^6 atoms [251] and simulation times in the order of ns to μs [252]. Simulations of more realistic systems, including explicit water molecules, counterions, and even a complete membrane-like environment [253,254] are possible, and new properties can now be studied as they evolve in real time. This progress in system representation has been accompanied by methodological improvements: better force fields [255,256], improved treatment of long-range electrostatic interactions [257], and better algorithms used to control temperature and pressure. However, despite all these advances, the set up of an MD simulation can be far from trivial.

Parameters used to describe proteins and their interactions are normally found within modern force fields, but adequate descriptors for non-standard molecules, such as ligands, might be missing. In such cases, the determination and fitting of new parameters is usually straightforward, but may be a time-consuming process if it needs to be done for many ligands, limiting the general applicability of the method. Commonly used programs for MD simulations of biomolecules include Amber [258], CHARMM [259], GROMOS [260], GROMACS [261], and NAMD [262], among others.

2.2.1 Programs and algorithms

The core of all MD simulations is the MM force field [255,263]. Force fields are classical descriptions of interactions energies that have been parameterised to reproduce a set of experimental data. The most popular force fields for protein simulation include AMBER by Cornell *et al.* [264,265], the CHARMM force fields [266], the force field developed at MSI called CVFF [267], the GROMOS force field [268,269], and the OPLS-AA force field [270]. Some of these force fields have also been developed to incorporate simulations of carbohydrates [271-273], nucleic acids [274-276], and membranes [277-279].

Most commonly used force fields are two-body additive, where the potential energy function is a function of pairs of atoms. The most general and simple form of such a potential is shown below,

$$E_{MM} = \sum_{bonds} k_r\left(r - r_{eq}\right)^2 + \sum_{angles} k_\varphi\left(\varphi - \varphi_{eq}\right)^2 + \sum_{dihedrals} \frac{V_n}{2}\left[1 + \cos\left(n\phi - \gamma\right)\right] + \sum_{impropers} k_{imp}\left(\theta - \theta_{eq}\right)^2 + \sum_{i<j}\left[\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}\right]$$

The internal energy of the molecule is determined by the first four terms, taking contributions from bonds, angles, dihedrals, and improper dihedrals; and the enthalpic contributions are estimated through the Lennard-Jones potential, which estimates the van der Waals interactions for the system, and the Coulomb equation, which determines the electrostatic interactions. The current electrostatic models treat electronic polarizability implicitly by choosing partial atomic charges that overestimate the molecular dipoles; however, new methods with explicit inclusion of electronic polarization are being developed [280]. Beyond these core terms, some force fields incorporate additional or alternate terms for both internal and external aspects of the force field. The largest differences among different force fields are in the philosophies used for the optimisation of the nonbonded parameters and treatment of hydrogen bonds.

In OPLS-AA [281] partial charges and Lennard-Jones parameters for liquids are assigned empirically from iterative pure liquid and dilute solution simulations aimed at reproducing bulk properties (e.g. density, heat of vaporization, free energy of hydration, etc), while charges for protein residues are adjusted from *ab initio* calculations according to CHELPG (Charges from Electrostatic Potentials using a Grid base method [282]).The AMBER force field [264] uses partial charges derived from restrained electrostatic potential (RESP) fits to *ab initio* charge distributions obtained at the HF/6-31G* level, while the CHARMM force field [266] was largely fit to scaled *ab initio* interactions energies at the HF/6-31G* level. A recent comparative study [263] supports the quality of all these three force fields, OPLS, AMBER and CHARMM.

2.2.1.1 Treatment of solvation

Biological molecules are surrounded by water and, therefore, an accurate treatment of the condensed aqueous environment is an essential aspect of a force field. This can be done using explicit solvent, which accounts for specific interactions, or an implicit model. Several potential functions for representing water have been developed.

The most popular explicit water models used in biomolecular simulations include the TIP3P [283], TIP4P [283,284], SPC, and SPC/E [285]. The parameters in all these models have been empirically adjusted to reproduce the enthalpy of vaporization and the density of water. However, none of them can describe correctly the temperature dependence of the density of water, which has been addressed by a recent development of the Jorgensen group, TIP5P [286].

When selecting a particular water model, it is important to confirm its compatibility with the biomolecular force field being used, as these have usually been developed in conjunction with a specific water model (e.g. AMBER and CHARMM with TIP3P, OPLS with TIP3P and TIP4P, and GROMOS with SPC).

There have been important advances in implicit solvation models [287,288]. These models represent the solvent as a dielectric continuum that mimics the solvent-solute interactions, including their apolar (van der Waals contacts and the entropic effects of creating a cavity in the solvent) and polar (mostly the screening of the electrostatic interactions) components. They offer significant computational savings while yielding accurate estimations of free energies of solvation [289]. Implicit solvent models are usually employed when extensive sampling of conformational space is required, but a caveat is that they can fail when highly specific water-solute interactions are important. In the Poisson-Boltzmann (PB) model, contributions from solvent polarization are taken into account [290-293]. Although these can yield accurate free energies of solvation, they are relatively computationally expensive. An alternative is to use approximate continuum solvation models, among which the generalized-Born (GB) formalism is one of the most popular [288,294,295]. In addition, both the PB or GB methods can be combined with free energy solvent accessibility (SA) terms that account for the hydrophobic effect [296,297].

2.2.1.2 Boundary conditions

In order to perform simulations under realistic conditions, biological molecules are usually solvated in water, and periodic boundary conditions are applied [298]. These produce an artificial replication of the system in all three spatial directions that ensures that all atoms are surrounded by neighbours and that the system does not end abruptly in a vacuum. However, the procedure introduces an artificial periodicity into the system. The alternative is to enclose the system in a solvent sphere, with restraints and stochastic forces acting at the boundary to mimic an extended system [299,300]. In this case, it might be necessary to simulate a considerably large system in order for large-scale dynamics to be possible. Stochastic boundary conditions are particularly useful when investigating only a particular region, such as the binding site in an enzyme [133]. This allows considerable saving of computational time by minimizing the extent of the system that is not of interest.

2.2.1.3 Long-Range Interactions

To a large degree electrostatic interactions determine the conformations of a biological macromolecule, and they are also very important in many association processes (for a recent review see [257]). Electrostatic forces are long-range and strongly dependent on the solvent and ions surrounding the biomolecule. Most electrostatic models used in structural biology comprise a full-atom representation of the molecule, and either treat the solvent explicitly, relying on Coulomb's law to compute the electrostatic energy, or treat the solvent implicitly, removing the necessity of calculating the expensive solvent-solvent interactions.

The most expensive part of an MD simulation is generally the calculation of the electrostatic interactions; for a finite system with *N* charges, ~N^2 interactions have to be computed. Traditionally, this has been addressed by imposing a cut-off, a truncation to the range within which electrostatic interactions are calculated [259]. There are different implementations of the cut-off, including: abrupt truncation, where the interaction energy between two atoms is set to zero when they are further apart than a given distance; switching, where the interaction energy is driven smoothly to zero between two cutoffs [259]; and force shifting, where the energy is modified such that both the energy and its derivative are zero at and above the cutoff [301]. To further reduce the computational costs, a list of neighbouring atoms or residues is kept and updated only every certain number of steps. In a twin-range cut-off method [298], the short-range interactions are calculated at every step of the MD simulation, and the long-range ones are only computed when the non-bonded list is updated.

An alternative approach is the Ewald summation method [302,303], which was originally derived for computing the electrostatic energy of infinite periodic systems, such as crystals. The possible artefacts introduced by this artificial infinite treatment of the system have been shown to be small [304-307]. A derived method, particle mesh Ewald (PME) [302,308] has been widely employed for MD simulations of proteins and nucleic acids [274,309]. When using Ewald methods it is important to use a periodic system of adequate size to avoid possible artefacts due to the molecule interacting with the electrostatic potential of its own images [310,311].

The fast multipole method [302,312] can be used to decrease the computational demand of electrostatic calculations from $O(N^2)$ to $O(N)$, with N being the number of charges in the system. In this approach, the charges are divided into sub-groups, so that a collection of distant charges can be considered as a single charge. The electrostatic interactions for particles within the same or neighbouring groups are treated exactly, while the potential for more distant particles is determined through multipolar expansions.

Instead of trying to include explicitly all interactions in an infinite system, or truncating the sum at some cut-off, a third alternative is to include the effects of the surroundings using a reaction field, which takes account of the response of the dielectric medium beyond the cut off [313-315].

2.2.1.4 Temperature and Pressure Control

Experiments are usually performed at constant temperature and volume (i.e. canonical ensemble) or constant pressure and temperature (i.e. isobaric-isothermal ensemble), so it is often desirable to perform MD simulations under the same conditions. During a simulation at constant energy, the temperature will fluctuate due to the interconversion of the kinetic and potential energies. The atomic velocities can be rescaled or modified to keep the temperature constant during the course of a simulation. Different approaches include the simple Berendsen scaling of velocities [316] or the Nosé-Hoover approach [317,318].

To maintain a constant pressure during a simulation, the volume needs to be allowed to fluctuate by adjusting the dimensions of the periodic box and rescaling the atomic positions accordingly. Several methods have been developed to run MD simulations at constant pressure, including the extended system algorithm [319,320], the constraint algorithm [321,322], the weak coupling method [316], the hybrid method [323], and the Langevin piston method [324].

2.2.2 Limitations and sampling improvement

To improve the exploration of the free energy landscape and reproduce changes taking place within the micro/milli second time scale within feasible computation times, it is necessary to increase the sampling power of conventional MD simulations [247,325]. Several accelerated MD variants have been proposed in the literature [326,327], which can be broadly grouped into three classes. The first class alters sampling of conformational space by modifying the potential surface, the second uses non-Boltzmann sampling to increase the probability of high-energy states, and the final one includes those methods that enhance the sampling of certain degrees of freedom at the expense of others.

2.2.2.1 Modified Potentials

The basic goal of modified potential methods is to reduce the amount of time that the system spends in a local energy minimum and speed the transitions from one minimum to another by modifying the potential energy surface. The methods include the deflation method [328], conformational flooding [329], umbrella sampling [330], local elevation [331], potential smoothing [326], puddle-skimming [332], hyperdynamics [333] and accelerated MD [334]. All these approaches use different strategies to modify the potential energy surface. For example, the local elevation method promotes sampling by adding a penalty potential to any conformation previously visited, while the conformational flooding approach increases the energy of previously visited conformations by adding a Gaussian potential to the system, thus forcing it to explore the rest of the conformational space. In umbrella sampling [330], the system is forced to sample specific regions of the conformational space by applying a biasing potential. In general, a series of calculations are performed at selected points, dragging the system through the configurational space surrounding the reaction coordinate.

2.2.2.2 Modified Sampling

Modified sampling methods include high-temperature MD [335], locally enhanced sampling [336], replica exchange [200,337,338], parallel tempering [339], self-guided MD [340], targeted MD [341], milestoning [342], repeated annealing [343], among others.

In the local enhanced sampling method [336], the simulation contains N non-interacting copies of a fragment of the system (e.g. side chain or ligand). Each copy interacts with the rest of the system with 1/N of its original magnitude. Therefore, the use of multiple copies and the reduction in the interaction potential significantly enhances the conformational sampling of the fragment.

In high-temperature MD [335] the system is simulated at elevated temperatures, physiologically irrelevant but useful for speeding the conformational search [344]. This can also be applied selectively to certain components of the system, and enhance the sampling specifically (e.g. ligand).

There are several approaches that use simultaneous or parallel MD simulations of the same system [345-347]. One special case is replica exchange MD (REMD) [200]. The replicas are simulated over a range of temperatures, and at particular intervals the temperatures of the simulations may exchange temperatures according to a Monte Carlo-like transition probability. When simulated at high temperature, the system can jump from basin to another, while the lower temperature replicas explore a single valley. This technique has been implemented in most MD programs, reflecting its wide application in protein studies.

2.2.2.3 Modified Dynamics

In these methods the dynamics along the “slow” degrees of freedom are accentuated relative to the “fast” degrees of freedom. A clear example is the SHAKE algorithm [348], used in almost all simulations. The length of the time step used in MD simulations is restricted by the requirement that Δt must be small compared with the period of the highest frequency motions being simulated. In the case of biomolecules, the highest frequency motions are the bond-stretching vibrations of hydrogen atoms, which are usually of minimal interest. Therefore, improvements in efficiency can be obtained by constraining the bonds of hydrogen atoms to fixed lengths using algorithms such as SHAKE [348,349], RATTLE [350] and LINCS [351].

2.2.3 MD Simulations and Enzyme Flexibility

As mentioned before, an ensemble of protein receptor conformations rather than a single structure is expected to provide a more realistic representation of an enzyme. Multiple structures could be obtained from experimental studies, such as NMR and X-ray analysis, or generated using computational tools.

Philippopoulos and Lim [352] suggested that the best source of protein conformations is NMR studies. A set of 15 NMR structures of *E. coli* ribonuclease HI was shown to explore a larger conformational space than that of a conventional 1.7 ns MD simulation of the system. Although both NMR and MD sampled similar conformations, NMR conformers covered a larger space with increased side-chain and protein-backbone mobility. As a caveat on these conclusions, it should be noted that this study employed conventional MD single-trajectory simulations, which are known to be inadequate for exploring large conformational spaces. Modified MD simulations designed to improve the conformational sampling of the system (see section 2.2.2), would likely produce better results.

MD simulations provide an easy practical alternative to explore the conformational space of the protein receptor in the many cases where multiple experimental conformations are not available. Several different studies have shown that MD simulations are generally in good agreement with experimental results in reproducing the general protein structure and dynamic processes occurring on the ps-timescale [198,353,354].

When preparing MD simulations for the exploration of the conformational space of a protein receptor, it should be remembered that the dynamic behaviour of the free and ligand-bound forms of the enzyme might be very different. Better binding energies are expected when docking against snapshots from simulations of the substrate-enzyme complex, as increased stability is expected to result from the induced fit produced during the simulation [355]. However, receptor conformations obtained from the simulation of a ligand-bound protein may be biased to accommodate particular types of ligands with particular binding modes (Figure 2-6).

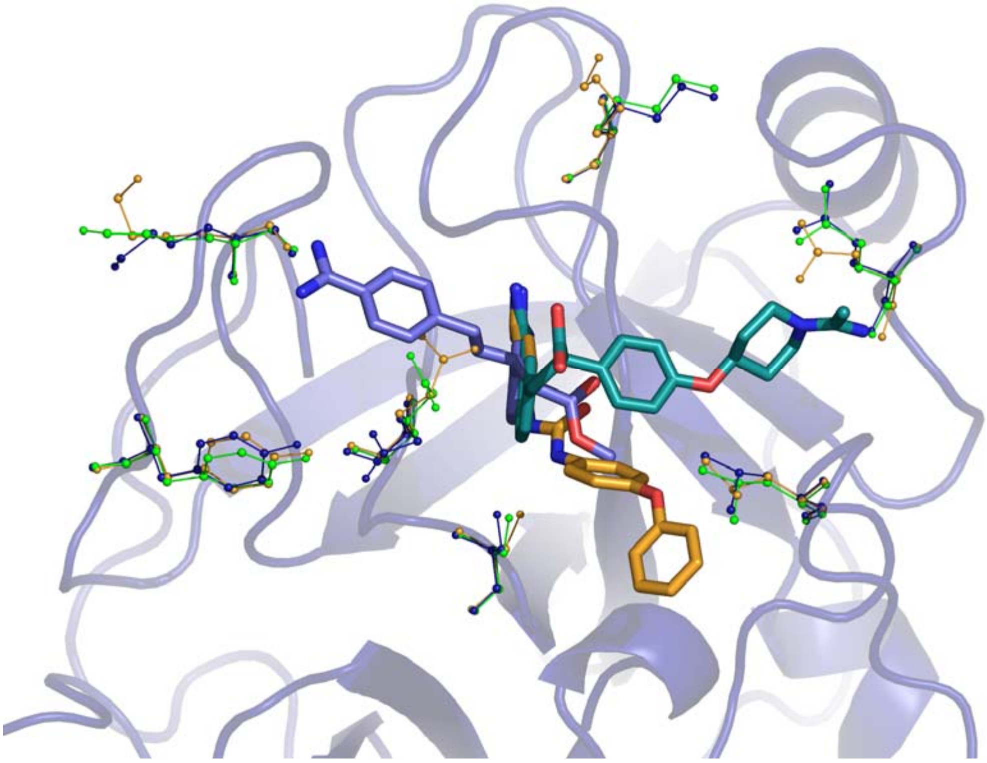

Figure 2-6 Differential ligand binding.

Several distinct binding modes for different ligands to a single protein receptor are possible. This superposition of three different complexes of the enzyme trypsin with the bis-phenylamidine inhibitor (1AZ8, blue), BX5633 (1MTV, green) and 1-(2-amidinophenyl)-3-(phenoxyphenyl)urea (1BJV, orange) shows that the ligands can adopt quite different orientations, with the side chains of the protein presenting different conformations depending on which ligand is bound.

2.3 Quantum Mechanical Methods

As nearly all enzymatic reactions involve bond-breaking and bond-forming steps, the computational methods used to study them must be capable of describing these processes. In principle, *ab initio* quantum mechanical (QM) methods could be used to reliably describe the system and generate the necessary structural and energetic data. However, it is practically impossible to solve the Scrhödinger equation for systems as big as proteins, and usually unnecessary. The reaction takes place within the active site, where only a small number of atoms directly participate in the bond-formation and bond-breaking process, while the surrounding environment usually influences the properties and reactivity of the reactants. Therefore, it is possible to use accurate and expensive QM calculations to simulate a small sub-system that is directly involved in the reaction, while using a simpler and less costly approach to describe the rest of the system, e.g. molecular mechanics.

2.3.1 *Ab Initio* Calculations

Given the position of the atomic nuclei, and the total number of electrons, the electronic energy, density and other properties of the system can be calculated by solving the Schrödinger equation [134]. There are two main approaches to solve the electronic Schrödinger equation: wavefunction- and density functional-based approaches.

The wavefunction-based approaches expand the electronic wavefunction as a sum of Slater determinants, the orbitals and coefficients of which are optimised. Hartree-Fock theory is the simplest method of this type, and involves the optimisation of a single determinant. However, its usefulness is limited as the electron correlations are completely neglected. Approaches which treat the correlation effect include the second-order Moller-Plesset perturbation theory (MP2) [356], and methods based on couple-cluster, like CCSD(T) [357], where excited states are included in the representation of the system.

MP2 methods [356] formally scale with N^5, with N being the number of electrons in the system. Advances in numerical methods during recent years have led to significant improvements, enabling systems containing a few hundred atoms to be treated routinely [134]. The performance of MP2 methods for properties involving making or breaking electron pairs is inferior to that of CCSD(T), but it gives very good results for non-bonded interaction and internal conformational energies.

Coupled cluster methods are the most computationally expensive, but also the most accurate of the approaches. Highly excited determinants are incorporated into the wavefunction without the combinatorial explosion of effort implied by configuration-interaction approaches. At present, the CCSD(T) approach (coupled cluster single, double, and triple perturbative excitations) provides the best balance between accuracy and efficiency [357]. As the formal scaling with the number of electrons (N) is N^7, calculations are limited to small- to medium-size molecules.

2.3.2 Semi-Empirical Calculations

Although *ab initio* QM calculations are expected to provide the best results, they are costly to apply. In those cases where a rapid but less accurate calculation will suffice, semi-empirical methods are commonly used [358,359]. Semiempirical methods are based on explicit treatment of valence electrons, pseudominimal basis sets, neglect of three- and four-centre integrals, and use of parameterised expressions for two-centre integrals. There are basically four types of semiempirical approximations: the extended Hüekel theory [360], complete neglect of differential overlap (CNDO) [361], intermediate neglect of differential overlap (INDO) [362] and neglect of diatomic differential overlap (NDDO) [363]. The most successful methods are based on the NDDO approximation [364], and include the modified neglect of diatomic overlap (MNDO) [365] and its successors AM1 [366] and PM3 [364,367] methods. In the NDDO approximation, all integrals involving diatomic differential overlap are neglected. The remaining integrals are evaluated using empirical formulas derived by Dewar and Thiel, except for the overlap integrals, which are computed analytically and are used to determine the two-centre resonance integrals. The parameters used in the MNDO, AM1, and PM3 procedures were optimised against experimental molecular geometries, heats of formation, dipole moments, and ionisation potentials. Therefore, their performance is satisfactory for ground-state properties, but these are not reliable for calculating transition state structures and energies. Proton-transfer activation energies are usually overestimated by as much as 20 kcal/mol by PM3 [368]. Basicities are always overestimated and nucleophilicities underestimated, resulting in inaccurate results for ion-molecule structures [369]. Nevertheless, continuous improvements are being made, which combined with their high computational efficiency, still makes them an attractive alternative to *ab initio* or DFT methods in many cases [358].

Warshel and coworkers have used the empirical valence bond (EVB) theory extensively [370]. In this approach, it is assumed that a reaction can be described by some VB resonance structures. The analytical form of these VB functions can be approximated by appropriate molecular mechanics potentials, and the parameters of these MM potentials are calibrated to reproduce experimental or *ab initio* molecular-orbital (MO) data in the gas phase as well as in the condensed phase. Although the EVB approach has been shown to provide very useful results when employed properly, its use requires considerable experience and parameterisation; this has limited its routine application to different systems.

2.3.3 Density Functional Theory Methods

Density Functional Theory (DFT) techniques show considerable advantages in generality and accuracy over semiempirical methods. Although the computational speed is comparable with *ab initio* HF calculations, DFT has the advantage of including electronic correlation effects. Within the local density formulation of the Kohn-Sham theory [371] the ground state energy is given as a functional of the electron density in the presence of an external potential.

In contrast to methods that attempt to solve the Schrödinger equation and for which the wave function is the basic variable, DFT methods express the energy of the system as a function of electron density. These are more cost effective for the treatment of large molecular systems [372], as they are a function of only three space variables [373], while in traditional QM methods the wave function is a function of 3N space variables, with N the number of electrons in the system. However, while the results of wave function-based methods can be improved systematically, the accuracy of DFT methods depends on the exchange-correlation functional, whose analytical form is unknown. Therefore, wave function-based methods are still used, at least to provide a benchmark against which to compare the accuracy of the DFT results.

There are two principal classes of DFT functionals that have been extensively deployed and tested in large scale applications as well as small-molecule benchmarks [374]: gradient-corrected (e.g. BLYP), and hybrid (e.g. B3LYP) functionals.

The accuracy of DFT calculations, in therms of energy and structure, are often as good as more expensive correlated MO-based methods. However, there are limitations to their accuracy, especially for hydrogen-bonded and weakly bound systems, for the determination of activation energies for some reactions, and when unpaired electrons are involved [375-377].

2.3.4 Linear-scaling QM methods

The cost of *ab initio* and semiempirical methods usually scale with the cube or higher power of the size of the system (N^3). The most cost-demanding components are the evaluation of interelectronic interaction terms and the diagonalization of the Hamiltonian matrix to find the electronic orbitals. Therefore, a QM treatment of the entire protein-substrate-solvent system is impracticable using conventional methods. But in the past decade much effort has been put into the development of algorithms whose costs scale linearly with the size of the system, the so-called linear-scaling semiempirical and *ab initio* DFT and HF methods [378-382].

One of the methods is the so called divide-and-conquer approach [383,384], originally developed for *ab initio* DFT studies. A large system is divided into many subsystems, and the electron density of each region is determined separately. Finally, the contributions from all subsystems are combined to obtain the total electron density and determine the energy of the system. In a different approach, the so-called frozen density matrix approach, the entire protein solvent system is treated quantum mechanically while freezing the electron density of the groups in the outer region [385,386].

Unfortunately, the treatment of the entire protein by *ab initio* approaches is extremely expensive and cannot be used in free energy calculations of enzymatic reactions, but only for structural/static properties. On the other hand, although linear-scaling semiempirical methods have been applied to the study of enzymes and their reactions [387,388], they are not likely to provide correct energies given the inherent inaccuracies of semiempirical methods [389,390].

2.4 Hybrid Potential Methods

To account for bond breaking/forming steps it is necessary to use QM methods, which can accurately predict structures and energetics of reacting groups. However, traditional QM methods are computationally expensive and it is impossible to treat the entire system at this level, limiting their use to small and simple systems. For this reason, less expensive empirical molecular mechanics approaches are commonly used for the study of biomolecules in solution. However, because the electronic structures of the molecules are not explicitly represented, MM approaches are not appropriate for the simulation of chemical reactions that involve bond forming and breaking. One solution is to combine QM and MM methods so that the reaction system, or active site, is treated explicitly by a QM method, while the surrounding environment, which constitutes the largest part of the system, is approximated by a standard MM force field [39,391,392] (Figure 2-7).

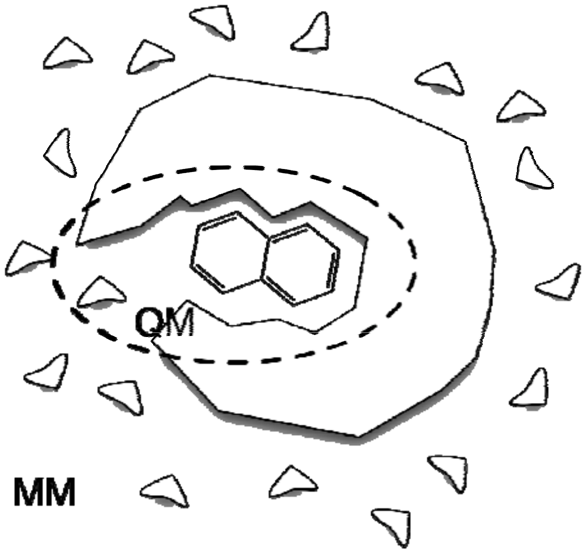

Figure 2-7 Partitioning of the protein system for the application of hybrid potential methods.

Schematic representation of a protein-ligand complex within an aqueous environment. The system has been divided in two regions, a smaller central area, including the ligand and the active site, which will be treated quantum chemically (QM), and the rest of the protein and solvent environment which will be modelled at the molecular mechanics (MM) level. This kind of separation reduces computational costs, while still providing an accurate description of the region of interest.

The so-called quantum mechanical/molecular mechanical (QM/MM) approaches [393-397] take advantage of the accuracy and generality of the QM treatment for chemical reactions, and of the computational efficiency of the MM calculations. Because the reactant electronic structure and the solute-solvent interactions are treated at the QM level, there is no need of parametrization for every new reaction (except for the EVB approach).

Using the Born-Oppenheimer approximation and assuming that there is no charge transfer between the regions, the effective Hamiltonian of the complete system can be separated into three terms:

$$H = H_{QM} + H_{QM/MM} + H_{MM}$$

where H_{QM} is the QM Hamiltonian of the region of interest (e.g. active site), H_{MM} represents the rest of the protein and $H_{QM/MM}$ is the Hamiltonian that couples the two regions. The total energy of the system is, therefore, composed of three different terms: one corresponding to the QM region, one for the MM part and the last one that accounts for the interactions between regions.

Different hybrid potential methods have been implemented [396], with most of the differences among them being in the number of regions into which the system can be divided, the potentials used to represent each region, and the way the interactions between regions are handled. Warshel and Levitt [397] were the first to introduce hybrid QM/MM potentials; these involved a combination of semiempirical methods to describe the QM region and standard protein force fields for the rest of the protein. Nowadays, the calculations can be carried out at different levels of QM electronic structure, including *ab initio* [392,398-400], semiempirical [397,401], DFT [402] or approximate DFT [403].

A critical aspect in these calculations is the correct choice for the partitions between the QM and the MM region. For some systems, there will be no covalent interactions between the enzyme and the substrate, and the separation into regions may be straightforward, provided the complete ligand can be modelled at the QM level. In this case the QM/MM interaction will include only

MM terms for the van der Waals interactions [404] and QM/MM electrostatic interactions. In most cases, however, the boundary between the QM and MM regions will need to separate covalently bonded atoms. When dealing with covalent bonds crossing the boundaries of the QM and MM regions, different approaches can be used for the so-called frontier bonds in the QM calculations [405]. The most common approach is the link-atom treatment [406,407], in which the unsaturated valences are satisfied by the addition of hydrogen atoms or halogen [401]. Another alternative is the local self-consistent field (LSCF) method, which describes bonds between the QM and the MM atoms as strictly localized bond orbitals [408,409]. The main disadvantage is that these localized orbitals have to be parameterised for each new system. The generalized hybrid orbital (GHO) method is transferable among systems [410], as the frozen orbitals are located at specially parameterised atoms that reproduce the bonding properties of full QM systems. The hybridisation of the hybrid orbital on the boundary atom is determined by the molecular geometry and is dynamically varied during the simulation [411].

The treatment of interactions between regions has been done so far using simple electrostatic and van der Waals approaches. Further developments are required to provide a better representation of the interactions at the interface [394,395,412], including hydrogen bonds [408] and polarization effects [413]. In the nonpolarization models, both the QM and MM portions are represented with static charge distributions. Most methods, however, allow polarization of the QM region by static charges of the MM region, while very few account for mutual polarization of the QM and MM regions.

The ONIOM method of Morokuma and coworkers [414,415] allows not only performing of QM/MM but also QM/QM calculations. Different regions of the system can be treated using several QM methods of varying precision, reducing the problem of dealing with interactions between regions, as the differences between methods on either side of the boundary are smaller.

2.5 Free Energy Calculations

Interactions between proteins and their ligands play central roles in many biological processes and, therefore, it is crucial to accurately predict their energies. As mentioned before, there are several different scoring functions for this purpose [218,219,235]. However, the type of scoring functions currently implemented in docking programs cannot be expected to distinguish energetically between close conformations of the same molecule, or even to rank properly a group of ligands of similar activity. If stringent rankings or accurate energies are needed, different MD-based calculations can be carried out on the final complexes to estimate the binding free energy [207,230,416-423].

Despite providing very accurate free energies, thermodynamic integration (TI) and free energy perturbation methods (FEP) are not widely applied as they are computationally expensive [421,424,425]. The main limitation of these approaches is the exhaustive conformational sampling required to obtained a proper averaged ensemble, and their slow convergence. Recently developed approaches that provide relatively good energy estimates at a moderate costs include MD-based methods such as the linear interaction energy (LIE) method [230,426-429], and the so-called MM-PBSA method (Molecular Mechanics/Poisson-Boltzmann Surface Area) [430].

2.5.1.1 Free Energy Perturbation and Thermodynamic Integration

Free energy perturbation (FEP) [431-434] and thermodynamic integration (TI) [435-438] are the most rigorous methods currently available to calculate the relative binding strength of different complexes. The difference in binding free energy between two given ligands, *L* and *L'*, and the receptor *R* is calculated using the following thermodynamic cycle:

$$\begin{tikzcd}
L + R \arrow[r, "\Delta G_{bind}(L)"] \arrow[d, "\Delta G^{w}_{mut}"'] & LR \arrow[d, "\Delta G^{p}_{mut}"] \\
L' + R \arrow[r, "\Delta G_{bind}(L')"'] & L'R
\end{tikzcd}$$

Instead of calculating the individual binding energies (ΔG_{bind}) to determine the relative binding free energy ($\Delta\Delta G_{bind}$), the energies of the non-physical transformations *L→L'* in solution (ΔG^{w}_{mut}), and *LR→L'R* within the protein environment (ΔG^{p}_{mut}) are determined. This type of alchemic transformation can be used to determine relative free energies, as the free energy is a state function that can be calculated by any reversible path between the initial and final states:

$$\Delta\Delta G_{bind} = \Delta G_{bind}(L') - \Delta G_{bind}(L) = \Delta G^{p}_{mut} - \Delta G^{w}_{mut}$$

In order to determine the energy of these non-physical mutations (ΔG_{mut}), the states L and L' are linearly combined using a coupling parameter λ, and an MD simulation is used to slowly transform one ligand (*L,* λ=0) into the other (*L',* λ=1) in both the free and receptor-bound forms. In the case of FEP calculations, the free energy associated with the change is calculated as:

$$\Delta G_{mut} = -RT\sum_{i=1}^{N-1} \ln\left\langle \exp\left(-\frac{H(\lambda_i + 1) - H(\lambda_i)}{RT} \right) \right\rangle_{\lambda_i}$$

where $H(\lambda_i)$ represents the Hamiltonian of the system at λ_i and $\langle\ \rangle_{\lambda i}$ indicates an ensemble average. In the TI method, the average of derivatives of the Hamiltonian are calculated for each λ, and then numerical integration over λ provides the total energy difference between the two states,

$$\Delta G_{mut} = \int_0^1 \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle d\lambda$$

These approaches have been used successfully to predict binding energies in several systems [439-443]. One of the most important limitations in free energy calculations is the sampling of the conformational space [438,444]. Exploration of the appropriate conformations is not guaranteed simply by longer simulations. In order to avoid convergence problems and inadequate sampling during the simulations, only transformations between similar molecules are feasible, constraining the type of ligands that can be compared. This, together with the computational cost of such approaches, has prevented the wide application of FEP for determining binding free energies, despite its accuracy.

2.5.1.2 Linear Interaction Energy Method

Åqvist *et al.* [230] introduced the linear interaction energy (LIE) semiempirical MD approach for the estimation of binding free energies [422,445]. This method assumes that the binding free energy can be extracted from simulations of the free and bound state of the ligand.

The energy is divided into electrostatic and van der Waals components, and the final binding energy is calculated as

$$\Delta G_{bind} = \alpha \left\langle V_{bound}^{elec} - V_{free}^{elec} \right\rangle + \beta \left\langle V_{bound}^{vdw} - V_{free}^{vdw} \right\rangle + \gamma$$

where $\left\langle V_{bound}^{elec} - V_{free}^{elec} \right\rangle$ represents the averaged change in electrostatic energy and $\left\langle V_{bound}^{vdw} - V_{free}^{vdw} \right\rangle$ the averaged change in van der Waals energy in going from an aqueous solution to a protein environment. α, β and γ are empirically determined constants. Two different MD simulations, one for the ligand bound to the protein and another for the free ligand in water, are used to calculate the energies. During the early applications of the LIE approach, only two coefficients, α and β, were considered. Although α, the electrostatic coefficient, appeared to have a constant value of 0.5 for several protein systems, as predicted by the linear response approximation [230], the van der Waals coefficient, β, seemed to adopt various values depending on the characteristics of the protein receptor [426,427,446,447]. Kollman and coworkers [429] suggested that the value of β depended on the hydrophobicity of the binding site, and that it could be predicted by calculating the weighted desolvation non-polar ratio (WDNR) of the system. Jorgensen's group extended the method to calculate both the hydration and binding free energy, adding a new term to account for the solvent accessible surface and scaling it by a new empirical coefficient [419,448,449]. It was later found, however, that the non-polar component γ, although considered zero in many cases [230,428], could adopt different values [450] and account for the variability earlier assigned to β. In a recent study, Åqvist and coworkers [451] showed that the coefficients of the LIE method are independent of the force field used and that only γ might need to be optimised to account for the hydrophobicity of the active site.

This approach has been successfully used to predict binding free energies in several cases [452-455]. The binding free energies obtained in all these cases were in very good agreement with experimental results and the LIE approach seems to be a good alternative to the more expensive FEP calculations. The two main shortcomings of the method are the need for two different MD simulations, one of the complex structure and another for the free ligand in water, and the use of empirically derived constants which may need to be modified for each particular system.

2.5.1.3 Molecular Mechanics/Poisson-Boltzmann Surface Area Method

The MM/PBSA (Molecular Mechanics/Poisson-Boltzmann Surface Area) method [417,456] was originally introduced by Srinivasan *et al.* [457]. It combines MM and continuum solvent approaches to estimate binding energies. An initial MD simulation in explicit solvent provides a thermally averaged ensemble of structures. Several snapshots are then processed, removing all water and counterion molecules, and used to calculate the total binding free energy of the system with the equation:

$$\Delta G_{bind} = \overline{G}_{complex} - \left[\overline{G}_{protein} + \overline{G}_{ligand}\right]$$

where the average free energy $\overline{G}$ of the complex, protein, and ligand, are calculated according to the following equations:

$$\begin{aligned}
\overline{G} &= \overline{E}_{MM} + \overline{G}_{solvation} - T\overline{S}; \\
\overline{E}_{MM} &= \overline{E}_{\text{int}} + \overline{E}_{elec} + \overline{E}_{vdw}; \\
\overline{G}_{solvation} &= \overline{G}_{polar} + \overline{G}_{nonpolar};
\end{aligned}$$

$\overline{E}_{MM}$ is the average MM energy in the gas phase, calculated for each desolvated snapshot with the same MM potential used during the simulation but with no cut-offs. $\overline{G}_{solvation}$, the solvation free energy, is calculated in two parts, the electrostatic component $\overline{G}_{polar}$ using a Poisson-Boltzmann approach, and a non-polar part using the solvent-accessible surface area (SASA) model [458]. The entropy ($T\overline{S}$) is the most difficult term to evaluate; it can be estimated by quasi-harmonic analysis [459-461] of the trajectory or using normal mode analysis [457,460,461]. The entropy change can be omitted if only the relative binding energies of a series of structurally similar compounds are required, but if the absolute energy is important, or if the compounds are notably different, then its contribution to the final free energy cannot be ignored.

Although only a single MD simulation of the complex is commonly used to determine the conformational free energy [430], as the structures for both the free ligand and ligand-free protein molecules are extracted from the simulation for the protein-ligand complex, this approach might not be the best. A recent study by Pearlman [462] showed that using a single simulation to generate all structures provides final results that are significantly worse than those from separate simulations, and that savings achieved in computing time are minimal and do not justify the simplification.

Application of the MM-PBSA approach has produced reasonable binding energies for several systems [416,418,420,463-467], but not for others [462]. It has been shown to produce accurate free energies at a moderate computational cost in most cases. Its main advantages are the lack of adjustable parameters and the option of using a single MD simulation for the complete system to determine all energy values. Nevertheless, this approach does have drawbacks, including the difficulties of predicting the entropic component of the free energy and the fact that the changes in internal energy of the ligand and receptor upon complex formation are neglected, which would produce significant errors in flexible systems where there is an important induced-fit effect.