Molecular dynamics-driven drug discovery:
Leaping forward with confidence

Aravindhan Ganesan¹, Michelle L. Coote² and Khaled Barakat¹*

¹Faculty of Pharmacy and Pharmaceutical Sciences, University of Alberta, Edmonton, Alberta, Canada.
²ARC Centre of Excellence for Electromaterials Science, Research School of Chemistry, Australian National University, Canberra ACT 2601, Australia.

*Corresponding author
Phone: 1-780-492-5783
Email: kbarakat@ualberta.ca

[Keywords: Molecular dynamics, drug discovery, computer-aided drug design, protein flexibility, binding free energy]

(21-September-2016)
Abstract

Given the high costs and time in developing a commercial drug, it remains important to constantly reform the drug discovery pipeline with novel technologies that can narrow down on the most promising lead compounds for clinical testing. The past decade has witnessed tremendous growth in computational capabilities that allow *in silico* approaches to expedite drug discovery processes. Molecular dynamics (MD) has become a particularly important tool in drug design and discovery. From classical MD methods to more sophisticated hybrid classical/quantum mechanical approaches, MD simulations are now able to offer extraordinary insights into ligand-receptor interactions. In this review, we discuss how the applications of MD approaches are significantly transforming the current drug discovery and development efforts.
Introduction

The quest for new drugs has always remained crucial throughout human history. From epidemic influenza of 1800s and 1900s[1] to the very recent Ebola virus outbreaks[2], the world’s population has constantly faced several dreadful epidemics, in addition to life-threatening diseases such as cancers. Thus drug discovery continues to be the most significant challenge for the scientific community. The overall drug discovery process, from the identification of potential lead compounds to the FDA approval of a drug, is not only extremely complex but also highly expensive and time consuming. A very recent report\(^1\) published by the Tufts Center for the Study of Drug Development (CSDD) estimates the overall costs for developing an approved drug at a staggering $2.6 billion, with an average of approximately 14 years to complete the entire development cycle of a single drug (from research labs to market)[3].

Drug design and development have matured over the last two decades by exploiting the advantages of new experimental techniques and complementary technologies. The early 1990s saw rapid advancements in combinatorial chemistry and high-throughput gene sequencing technology. These allowed the synthesis of huge compound libraries within a short span of time and their screening for various targets, thereby accelerating the discovery processes. This raised the hope of transforming the drug discovery field, making the natural products obsolete. But over time, the field of combinatorial chemistry started to face a lot of technical challenges. Particularly, the combinatorial libraries did not cover many structurally diverse compounds[4]. Further, the compounds in these libraries were also not stereochemically rich as the natural products. Therefore, the steep expansion in these compound libraries did not provide the expected fruitful outcomes; on the contrary, they only escalated the costs of testing[5] and resulted in reduced success rates. For example, until recently, only two compounds generated de novo have reacted the market as a drug[6]. One of them is sorafenib from Bayer, which was first approved by FDA in 2005 as a drug for cancer. The second drug that was possibly generated from de novo design is ataluren, which was approved in the European union in 2014 as a drug for the treatment of genetic disorders[6]. Nevertheless, there have been some significant efforts towards improving the

\(^1\) [http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study](http://csdd.tufts.edu/news/complete_story/pr_tufts_csdd_2014_cost_study)
combinatorial chemistry field. For instance, some of the schemes for addressing the lack of diversity were developed; and this includes diversity-oriented synthesis[7], which employs a “build/couple/pair strategy”[8]. In addition, strategies, such as ‘split and pool solid phase synthesis’, were developed as more powerful approach for synthesizing huge combinatorial chemistry. Despite many efforts, the field of combinatorial chemistry has still not reached its full capacity. Kodadek[9] discusses various recent advances in the combinatorial chemistry. This has led to a focus on computational methods as low-cost tools for driving the early search process for compounds with desired biological activity and pharmacological profiles, before initiating experiments.

Structure-based drug design (SBDD) is one of the vital computational approaches that has been found to be very effective in the identification of hits for in-vitro testing. As the name indicates, in principle, knowledge of the three-dimensional structures of proteins and the ligands are mandatory to perform SBDD. Recently, there has been dramatic accumulation of biological data, from gene sequences to three-dimensional structures of proteins and compound databases, which offers excellent support to SBDD research. As of June 2016, the Protein Data Bank (PDB) (www.pdb.org) contains more than one hundred thousand experimentally-determined (e.g., via X-ray, NMR and electron microscopy) protein structures, of which almost 26% correspond to human proteins. The UniProtKB/Swiss-Prot genome database (www.uniprot.org) contains ~540,000 amino acid sequences. These huge databases offer a gamut of potential targets for several human diseases. Moreover, when the experimentally-determined 3D structures of any proteins (or enzymes) are not available in the PDB, computational models of the unknown proteins for subsequent in-silico studies can be constructed using SBDD-based methods such as homology modelling, threading and de novo designing[10]. The success of virtual screening (see Glossary) and SBDD is also dependent on the availability of different compound libraries that comprise chemical compounds from diverse structural classes, so as to increase the probability of obtaining novel hits. There are a number of freely available compound databases, such as ZINC15[11,12] (~120 million compounds), Chemspider[13] (35 million compounds), ChEMBL[14] (~2 million compounds), DrugBank[15] (~14000 compounds), PubChem[16] (64 million compounds), among others.
When a specific target and compound libraries are selected, molecular docking-based virtual high-throughput screening is employed to identify only those compounds (from the libraries) with higher affinities to the protein’s active site[17]. The proteins are dynamic biological molecules and their flexibilities play vital roles in the process of ligand recognition, and thus in SBDD. In addition, ligand binding also tends to induce measurable levels of conformational changes in the proteins so as to adapt a biophysical state that is suitable to form a strongly-bound complex (known as induced-fit effects). Nevertheless, accounting for receptor flexibilities remains a major challenge and regular molecular docking methods are mostly unable to capture such conformational changes in proteins.

Molecular dynamics (MD) is a computational method that can take on this challenge and predict the time-dependent behaviour of a molecular system, thus becoming an invaluable tool in SBDD. It has been particularly valuable in exploring the energy landscapes of proteins and identifying their physiological conformations, which are, in many cases, not even accessible through high-resolution experimental techniques. MD is also useful in the structural refinements of post-docking complexes, such that the complementarity between the ligand and the receptor are enhanced in the complex-state, thereby allowing better re-scoring of complexes.

This review will discuss in detail the various applications of MD approaches in modern drug discovery efforts. Although, there have been a number of recent reviews[18-26] focussing on the usefulness of MD in drug design, they are mostly focussed on the theoretical background, applications of MD for accounting protein flexibility and selected few binding free energy methods. However, the current review aims to complement the existing ones in the literature, by addressing various aspects of SBDD, for which MD methods and QM/MM approaches can offer some valuable solutions. The review begins by briefly introducing molecular docking and virtual screening in SBDD. We discuss the recent developments in docking methods and how they struggle to account protein flexibility in SBDD. Subsequently, we discuss in detail how MD is helping to fill this gap and various applications of MD in SBDD, including post-docking structural refinements.
and accurate binding free energy estimations. Various binding free energy methods and
their recent developments are presented, along with a number of examples. In addition, we
also discuss an emerging trend of using solvent information more explicitly from MD
simulations, which provide significant information the effects of water molecules in drug
design. Further, we also caution about various limitations in MD methods and,
subsequently, we discuss about the applications of advanced hybrid QM/MM MD in drug
design. Finally, we present our perspectives by presenting a simple and practical workflow
for integrating various computational methods discussed in this review for SBDD.

Molecular docking and flexibility challenges
Molecular docking (see Glossary) protocols predict the optimal placement of each of the
compounds within a pre-defined active site of a protein target. They generate a
comprehensive set of conformations of the ligand-receptor complex (predominantly based
on the ligand poses). These poses are subsequently ranked based on their stability using
different scoring functions[27]. There are a number of programs for ligand-protein
docking, including, DOCK[28], AutoDock[29], Gold[30], and GLIDE[31]. These
docking-based methods have been of great use in modern drug discovery campaigns,
mainly because of their speed and simple set-ups.

Early docking methods assumed that the ligand-protein binding phenomenon could be
modelled as a simple ‘lock-and-key’ scheme. That is, the aim was to identify a ligand (i.e.,
a key) with the exact shape complementarity to fit within a stiff active site cavity (as a
keyhole) of the protein. In this way, most early docking algorithms treated the ligand and
the receptor as two rigid counterparts. This assumption holds well only for very rare cases,
such as the trypsin-BPTI complex, in which the interface of the bound and unbound states
is almost identical in their conformations[32]. However, it does not reflect the reality in
vast majority of cases, where both ligands and receptors undergo mutual changes to
accommodate each other in the complex state. Thus, ligand-protein binding mechanism is
now described as a ‘hand-and-glove’ scheme (Fig. 1), indicating that the best fit is still an
essential factor but under a flexible environment[33]. Most of the current docking software
programs have adopted ligand sampling as one of the basic elements in their docking
protocols. Several sampling algorithms, such as shape matching, systematic search and stochastic algorithms, are currently employed in docking to generate several ligand conformers (often referred as poses) around the given receptor environment[34]. For example, software programs such as GLIDE[31] and LUDI[35] implement systematic search methods in docking; while AutoDock incorporates stochastic methods for accounting ligand flexibility in docking. Thus, there have been significant advancements in the methods to allow exhaustive ligand flexibility in docking-based virtual screening[34].

On the other hand, protein flexibility has been almost ignored in docking calculations. Very few techniques, such as soft-docking and rotamer libraries[34,36], have been developed to tackle this problem. In soft-docking, the protein flexibility is implicitly included during the calculation, by softening the interatomic van der Waals terms in the scoring function such that it allows small levels of overlaps between the receptor and ligand[34,36]. Software programs such as GOLD and AutoDock implements soft-docking. Some programs attempt to implement protein conformational changes into docking calculations by treating the side chains as flexible, while retaining the rigidity of backbone atoms [34,36]. These methods employ rotamer libraries, which comprise a list of side-chain conformations determined by experiments and statistical analyses. GLIDE[31], for example, adapts an induced-fit docking method, where selected side chains are mutated into alanine residues to avert steric clashes during docking[31]. Later on, these side-chains’ conformations are adjusted to generate possible configurations that can adopt to the new environment, followed by an energy minimization of the binding site.

Nevertheless, such attempts only allow local movements of some selected residues in the active site, but are not able to capture the overall effects of ligand binding on the conformation of proteins. To overcome that, an ensemble of protein structures can be used to account for the full receptor flexibility during docking. This method has become one of the most widely accepted techniques in SBDD. In this approach, all protein structures are combined together to form a single representation [18,36], so that it can include conformational changes occurring during the ligand binding process. This is usually
achieved by averaging the grids of the different protein conformations (in the ensemble) into a single global receptor grid and employing this final grid in molecular docking. Knegtel et al. [37] made one of early attempts in employing an averaged grid that is generated from different experimentally-determined structures for ligand docking. The authors employed this approach for different test cases, including HIV protease, ras p21 protein, uteroglobins and retinol binding protein. They found that the averaged grids approach exhibited better accuracies when compared to those of a single structure. The issue of protein flexibility in docking was also addressed by using a united description scheme [38]. In this way, multiple experimentally-derived protein structures are superimposed, where the similar segments in the ensemble structures are aligned and fused together, while the variable regions are used as an ensemble. The ensemble of varied segments of proteins is combinatorially-explored to produce possible new conformations of proteins for docking calculations [18,34]. However, this approach relied heavily on the quality of the ligand conformational sampling. In addition, such approaches account only for the ligand-protein interaction energy where the internal energy of the protein is mostly neglected [18] (Fig. 2).

An alternate ensemble-based strategy, to model protein flexibility in molecular docking, is to explicitly consider multiple individual receptor conformations [39] and perform rigid docking of ligands against all those target structures. An ensemble of protein configurations is usually generated from an NMR structure of the chosen receptor or a set of X-ray crystal structures for the same receptor but with different ligands. Nevertheless, the main pitfall with using an ensemble of X-ray crystal structures is that the subsequent docking (or virtual screening) could be biased towards the structures available. This could be even more troublesome if all the available structures are co-crystalized with analogous ligands. On the other hand, in the absence of those experimental structures, modeling and MD simulations can be carried out to collect statistically significant protein conformations from the resulting (MD) trajectories. More discussions about this strategy are provided in the following sections. In fact, this combination-approach (i.e., mixing MD and molecular docking) is becoming more common [40-43], irrespective of the availability of experimental structures. For example, in a recent study, Campell et al. [44] presented an
approach that employs a biased-MD simulation on the known X-ray crystal structure(s) of ligand-protein complex(es), followed by rigid ligand docking to identify the best ranking pose for the complex(es). To demonstrate this scheme, the authors selected two test systems, cyclin-dependent kinase 2 (CDK2) and factor Xa (FXa) [44]. The authors collected the available crystal structures of these systems and performed MD simulations by introducing an external bias potential to retain the initial ligand conformation, thereby also maintaining the binding state that has been known to exist. Later, the authors collected a cluster of protein conformations from the MD trajectory and employed them for the ensemble-based docking of a new set of ligands in the known pocket. This work[44] demonstrates that despite the availability of crystal structures, MD simulations may be very useful to account protein flexibility in docking-based virtual screening. A recent study[45] showed that enrichment performances of virtual screening against three different targets, neuraminidase, HIV protease and peroxisome proliferator-activated alpha receptors, displayed excellent improvements when employing MD-based screening. Therefore, MD methods are now recognized as a valuable tool in SBDD.

Classical molecular dynamics
Molecular dynamics (MD) is the most widely employed computational technique to study the equilibration structures and dynamic interactions of biological systems [20,24,40-43]. They extend insights into time-dependent variations and configurational changes in the structures of the biological systems, which may be related to their functionalities[46]. Classical molecular dynamics regards atoms as solid spheres and the bonds connecting them as springs. This allows the atoms in the system to only oscillate within a specified distance. Classical MD is based on Newton’s equations of motion,

\[ m_i \frac{\delta^2 r_i}{\delta t^2} = F_i \]  

Here, \( F_i \) is the component of the net force acting on the i-th atom with a mass, \( m_i \). \( r_i \) denotes the position of the atom at time t. The force can then be computed as,
where, $U(r_1, r_2, \ldots, r_n)$ is the potential energy function of the specific conformation and can be described by using the concept of a force field with pre-defined parameters [47]. A force field is a mathematical expression comprising the functional form of the potential energy, which includes the possible bonded (bonds, angles and dihedrals) and non-bonded interaction (van der Waals potentials and Coulomb potentials) terms between the different atoms in the system. The bond stretching and angle terms are commonly modelled using a harmonic potential function, while the dihedrals are expressed as a cosine function. The non-bonded terms are modelled using Lennard-Jones potentials[48] and Coulomb’s law. Particle-mesh Ewald (PME) method[49] under periodic boundary conditions is normally employed in the classical MD simulations in order to treat long-range electrostatic interactions in the system. A number of force fields have been developed for MD simulations of biological systems, such as CHARMM[50], AMBER[51], GROMOS[52], etc. Most of these methods have different functional forms to treat MD simulations, which makes it difficult to transfer parameters from one force field to another.

It is generally difficult to compare the performance of different force fields, as the outputs will significantly depend on the type of system and properties studied[53]. However, there have been some efforts to compare different force fields and most of them find that the results concerning the structure and dynamics of systems could vary depending on the force field. For example, Todorova et al[54] compared five popular force fields, such as CHARMM27, OPLS, AMBER03, and the united-atom GROMOS 43A1 and GROMOS 53A6 force fields, for simulating insulin. The study addressed the effects of each force field on the conformational evolution and structural properties of the peptides and compared them against the established experimental data. The results found that different structural trends emerged depending on the force field used; however, the CHARMM27 and GROMOS 43A1 delivered the best representation of the experimental behavior[54]. Similar conclusions were drawn from a number of other studies as well; but some studies concluded that no major differences (in properties and performance) were detected when
comparing different force fields. Therefore, it is important to make a careful selection of a
force field before employing it in MD simulations. ‘Learning from experience’ is one of
the practical approaches for choosing a force field for MD simulations. Before choosing a
force field, the users need to be clear about the system they are working on and what is the
key question (or property) that they are trying to address through MD simulations.
Subsequently, the users need to do literature search to find out if MD simulations of similar
systems or properties have been reported earlier and if yes, what types of force fields were
applied to those simulations. If more than one force field have been applied, then which
one among them was able to provide more accurate results needs to be identified. It is also
important to benchmark the selected set of atomic force fields to test against some reliable
metrics. Sometimes the choice of force field may also depend on the type of water models
involved in the simulations, as force fields have been developed for certain water models
(such as TIP3P, TIP4 and SPC) [54,55]. For instance, it has been suggested that the
combinations of TIP3P-AMBER, TIP3P-CHARMM, TIP4P-OPLS and SPC-GROMOS
have been more relevant to the experiments[54,55]. Although there are some exceptions
shown in the literature, for instance see reference[55]. Becker et al[56] have listed a number
of considerations for choosing an appropriate force field in material science and
engineering and these suggestions also holds well for biomolecular simulations.

Solving Newton’s equations of motion analytically is unpractical for the thousands of
degrees of freedom typically involved in many MD problems. As a result, numerical
integration algorithms, such as Verlet integrator[57], velocity Verlet integrator[58] and
leapfrog integrator[59], are usually employed to solve these equations and predict the next
move for all atoms during MD simulations. Since the dynamics of the covalent bonds
involving hydrogen atoms are not very crucial in biological problems, they are usually
constrained using integration algorithms, such as SHAKE[60], RATTLE[61] and
LINCS[62]. Hence, a time step value in the range of 1.5fs to 2fs is possible and has shown
to be suitable for MD simulations of many biological systems[46].

The main advantage of MD approach is its abilities to mimic the experimental conditions
in which a typical biological question is addressed. For instance, experiments are carried
out by controlling different factors, such as temperature, pressure, number of atoms, ionic concentration and the type of solvent used to solvate the interacting molecules. All these factors can be readily adjusted and controlled in MD simulations within the context of statistical mechanics ensembles [63]. These ensembles include the microcanonical ensemble (constant total energy), canonical ensemble (constant temperature), and isothermal-isobaric ensemble (constant temperature and pressure). The microcanonical ensemble is the most basic approach and involves a constant number of particles (N), a constant volume (V) and constant energy (E). However, as the condition of maintaining a constant total energy is not realistic [64], the canonical ensemble (NVT)[65] and isothermal-isobaric ensemble (NPT)[66] are commonly used. A number of thermostats and barostats, such as Langevin[67], Berendsen[68] and Nose-Hoover[69,70], have been developed to fix the temperature and pressure in MD simulations. In fact, the isothermal-isobaric (NPT) ensemble is the most widely used ensemble in MD simulations, as it reflects the actual experimental conditions. There are a number of classical MD programs, including but not limited to, AMBER (www.ambermd.org), CHARMM (www.charmm.org), NAMD (www.ks.uiuc.edu/Research/namd/), GROMACS (www.gromacs.org), Desmond (www.deshawresearch.com) and Hyperchem (www.hyper.com). Some important quantities that are frequently used when analyzing MD trajectories are provided in BOX 1.

**MD simulations and protein flexibility**

The dynamic nature of proteins is a well-established phenomenon [71]. Proteins are very flexible biological molecules that can adopt multiple conformational states in solution [18]. Very few of these conformations are able to bind efficiently to the ligands and/or other systems in the environment. For example, certain configurations of proteins may adapt an open state that keeps the channels accessible for water molecules and ligands to bind/unbind freely [72,73]. On the other hand, in some other conformations of the same proteins, the highly malleable loops may be blocking the channel partially or completely thereby restricting ligand access. In addition, binding of the ligand may also lead to conformational changes in proteins, from local reorganization of side-chains to hinge dynamics of domains [40-42]. As a result, proteins often shift between different
conformational states separated by low- and high-energy barriers in the free-energy landscapes during chemical reactions. Histone deacetylase 8 (HDAC 8) is one of the best examples for dynamic mobility in proteins. These unusual dynamics of HDAC8 have been captured by at least 21 different experimentally-determined structures (PDB IDs: 1T64, 1T67, 1T69, 1VKG, 2W22, 2V5W, 2V5X, 3EW8, 3EWF, 3EZF, 3EZT, 3F06, 3F07, 3F0R, 3SFF, 3SFH, 3MZ3, 3MZ4, 3MZ6, 3MZ7, and 3RQD)[46,71,74]. By comparison of all the reported experimental structures, it was found that an 11 Å deep active-site pocket of the enzyme changes between a broadly open conformation to a partially open state and a fully closed structure[71,74]. Some experimental structures of HDAC8 also display an extra pocket that lies parallel to the main pocket (Fig. 3). All these structures are proposed to exist in equilibrium and are involved in ligand binding/unbinding, product release or water transfers[46,71]. Furthermore, some proteins could have additional druggable binding sites, which are cryptic in nature and have the potency to modulate the functionalities of the concerned receptors allosterically. Such cryptic or allosteric binding sites are usually not easily detectable in the ligand-free structures, as in TEM1 β-lactamase[75] and p38 MAP kinase[76] for instance, and require significant conformational changes in the receptors to become visible. Hence, these sites are usually not detectable from a single representative structure and requires large conformational sampling to reveal them. One of the well-known success stories of MD in such applications is with regards to the discovery of a novel-ligand binding trench in HIV-integrase enzyme. In 2004, Schames et al[77] performed MD simulations of HIV-integrase enzyme along with the docked ligand and discovered a novel ligand binding region, the trench. The existence of this cryptic site was later also confirmed by X-crystallography. Subsequently, scientists from Merck along with their collaborators performed intense experimental research[78] on this novel binding site, which eventually led to the development of novel anti-HIV inhibitors such as raltegravir[19].

It is therefore logical to use an ensemble of protein conformations in SBDD, instead of a single representation. Nevertheless, due to high-costs and technical complexities, experimentally-determined structures for different conformations are only available for few proteins. As discussed in earlier sections, MD simulations are now being used to
collect ensembles of protein structures for SBDD in order to close this gap. Under this MD scheme, the target structure (obtained from PDB or computational modelling) is initially subjected to large-scale MD simulations followed by root mean squared deviation (RMSD) conformational clustering to accumulate all possible conformations of a typical protein structure. Subsequently, statistical analysis methods, such as Principal Component Analysis (PCA), are then employed to transform the original space of correlated variables into a reduced set of independent variables comprising of the most vital dynamics of the system[40-42,79]. This will result in an ensemble of protein structures that can be used in docking-based virtual screening. This MD scheme to account for receptor flexibility is popularly known as the ‘relaxed complex scheme’ (RCS)[42]. RCS has been successfully employed in a number of studies[40-43]. For instance, we have employed extensive MD simulations to conduct an ensemble-based virtual screening against the MDM2 protein[41], a main regulator for p53. We performed over 50 ns MD simulations of the structure of MDM2 using the AMBER99SB force field and NAMD program and sampled 28 distinct conformations of MDM2 for further virtual screening of several ligand databases[41]. The 28 structures included twenty-two structures that comprised ~75% of the apo-trajectory, five structures representing ~80% of the bound-trajectory and a single MDM2 conformation from the MDM2-p53 crystal structure [41]. The study revealed that MDM2 is a highly flexible protein and adopted distinct conformational changes[41], which were effectively captured using MD simulations, as shown in Fig. 4a². In another study, Bowman et al[80] performed MD simulations of p53-MDM2 complex and generated multiple structures of the systems, so as to account protein flexibility in their subsequent docking-based virtual screening. This led to the discovery of five small-molecule inhibitors of the human MDM2-p53 interaction. Particularly, one of the compounds exhibited a Ki of 110 ± 30 nM[80]. These small molecules indeed have distinct scaffolds from nutlin, a known inhibitor of MDM2-p53 interaction[80]. Thus, incorporating RSC approach is able to discover novel therapeutically attractive small molecules. In one of our another studies, we used the MD-based RSC approach to develop a computational atomistic model of a

---

human ether-a-go-go-related (hERG) ion channel [40]. Conformational sampling of the
MD trajectory of hERG resulted in 45 different clusters that made a comprehensive
description of backbone (Fig. 4b) and side-chain dynamics (Fig. 4c) of the inner cavity of
the ion channel[40]. This model serves as a powerful tool to predict hERG blocking and
can be useful in developing safer and more efficient drugs[40].

In combination with other computational approaches, MD simulations can help in
characterizing protein-protein interactions. These types of interactions play important roles
in several biological processes, including signal transduction, cell metabolism and/or
transport. Understanding these interactions can access a new era of drug discovery, hence,
expanding the target space for new and more effective drugs[81]. Although the protein-
protein interfaces are generally large, only selected subset of residues are responsible for
the strong binding at these sites. Such residues, along with the surrounding domains, are
known as hot spots. Protein-protein interactions are also known to possess transient binding
pockets that are not captured in experimentally-determined structures. MD simulation has
become routine in approaches for identifying these hotspots and predicting binding sites
for their regulation. For instance, MD simulations have provided a detailed understanding
of the dimer interface in the HIV 1 protease enzyme, which is characterized by solvent
accessible surface areas and inter-dimeric hydrogen bonds[82]. In a recent study [83], we
employed MD simulations to model and characterize the human programmed death-1 (PD-
1) bound to its two human ligands, PDL-1 and PDL-2. Table 1 lists some of the studies
that have employed MD simulations for various applications (such as accounting protein
flexibility and dynamics, post-docking structural refinements and free energy of binding
calculations) on different target enzymes (or proteins) in the past five years.

MD simulations and post-docking structural refinements

Although docking can predict the optimal placement of a ligand within a receptor’s active
site, not all of the key interactions between the ligand and receptor are usually depicted

---

3Reprinted from Toxicology Letters, 230 (3), Anwar Anwar-Mohamed, Khaled H. Barakat, Rakesh Bhat,
Sergei Y. Noskov, D. Lorne Tyrrell, Jack A. Tuszynski, Michael Houghton, A human ether-a-go-go-related
(hERG) ion channel atomistic model generated by long supercomputer molecular dynamics simulations and
its use in predicting drug cardiotoxicity, 382-392, Copyright (2014), with permission from Elsevier.
accurately. Hence, it is generally recommended to perform MD simulations on the complexes obtained from docking as this can help in optimizing their interactions. For instance, in a previous study[84], molecular docking predicted that sulphonamide derivatives bind effectively into the active site of aldose reductase, which was contrary to the less activity found for these compounds in experiments. In silico refinements of these structures using MD revealed that a water molecule from the exterior migrated to the binding site and interrupted the key interactions between sulphonamide ligands and the receptor. This was identified to be a reason for the reduced activity of the tested compounds in experiments [84]. In another study, MD simulations were used to discern the different docked complexes of propidium and human acetylcholinesterase based on their stability[85]. The most stable structures identified with the help of MD simulations were in excellent correlation with the binding modes proposed by experiments[85]. Similarly, a combination of ensemble-based molecular docking and MD refinements of post-docking complexes helped us reveal for the first time a unique symmetrical binding mode of Daclatasvir (a drug in phase III clinical trials) with the Hepatitis C virus (HCV) NS5A protein and for different HCV genotypes[43], refer to Fig. 54.

MD has made significant contributions in the understanding the structure-properties of G-protein coupled receptors. For instance, a previous study involved post-docking MD simulations to reveal significant dynamic changes in the human CC chemokine receptor 3 (CCR3) and the human muscarinic acetylcholine receptor 3 (hM3R) that influence their ligand binding modes [24]. Especially, MD simulations found a strong H-bond between the docked ligand and N508 residue of hM3R that is key to holding the complex. This was again confirmed by performing MD simulations of N508A mutant hM3R and ligand complex, in which the ligand was found to be pushed to the exit [24]. In another study by Perdih et al[86], the authors employed molecular docking and MD simulations, along with a range of experiments, and identified some of the furan-based benzene mono- and dicarboxylic acid derivatives as potential inhibitors of all four bacterial Mur ligases. The

---

authors initially performed in-vitro testing of seven furan-based benzene-1,3-dicarboxylate derivatives, based on their previous studies and found out that one of the compounds was able to inhibit all Mur ligases in the micromolar range\[86\]. Subsequently, this compound was docked into the active site of MurD enzyme and two different ligand binding modes were identified. Subsequently, the authors performed ~20 ns long MD simulations and interaction energy calculations, so as to further refine the post-docking complex and also identify the best binding mode of the ligand. Finally, based on the results obtained, four novel furan-based benzene-monocarboxylic acid class were discovered to inhibit multiple Mur ligases in the low micromolar range\[86\]. Moreover, one of the newly discovery compounds also exhibited promising antibacterial activity against \textit{S. aureus}\[86\].

Previous studies\[87-89\] have also shown that integrating induced-fit docking (IFD) method along with MD and/or QM/MM simulations can be useful for the efficient description of induced-molecular flexibilities within the protein-ligand complexes and also for accurate binding mode analysis of ligands. For example, in a recent study, Distinto et al\[87\] employed IFD and MD simulations in order to unravel the putative binding modes and activities of 1-(aryliden-2-(4-(4-chlorophenyl)thiazol-2-yl)hydrazines against monoamine oxidase B (MAO-B) enzyme, an attractive target for neurodegenerative diseases. By structural alignment of several X-ray structures of MAO-B co-crystalized with different inhibitors, it was understood that the enzyme adopted induced-fit changes with respect to the bound ligands. Hence, the authors initially performed IFD using the Schrodinger drug discovery suite, during which the side chains near the inhibitor were kept flexible. The results from the IFD showed explained the how ligand binding tend to induce structural changed in the protein. However, many of the compounds showing two binding modes were ranked high in IFD. In order to find out the best binding mode of the inhibitors, the authors performed 3 - 5 ns long MD simulations for both the binding modes of two of the top-ranking compounds from IFD. The MD results followed by the free energy calculations highlighted the significance of fluorine atom interacting with the water near the cofactor and the influence of steric bulkiness of substituents in the arylidene moiety. The authors propose that the pharmacophore features of these experimentally synthesized compounds, developed using combined IFD, MD and free energy calculations, should be
useful for achieving novel high-affinity MAO-B inhibitors for the treatment of neurodegenerative disorders[87].

In another study, Fu et al[88] combined IFD with classical MD simulations, free energy of binding calculations and QM/MM calculations to study the substrate binding to human biliverdin-IXα reductase (hBVR-A) of biliverdin-IXα and four analogues. hBVR-A is a key enzyme in regulating a wide range of cellular processes and is involved in the conversion of biliverdin-IXα to bilirubin-IXα. In this work[88], the authors initially employed the structure of the hBVR-A/NADPH/substrate I complex for the docking of analogs into the binding pocket using the IFD procedure implemented in the Schrodinger program. During the IFD, a tyrosine residue in the active site was treated with flexibility. Subsequently, the best-ranking ternary complex structures from IFD were subjected to classical MD simulations[88]. Multiple snapshots obtained from the MD simulations were used for performing free energy of binding calculations. The predicted free energies of binding for five analogues agreed well with the experimental binding affinities and also helped to identify the best binding pose for the complexes[88]. Finally, the authors investigated the catalytic mechanisms of the ternary complex structure (in this study) by calculating the reaction energy profiles using advanced QM/MM calculations. These advanced calculations were useful to understand the reaction mechanisms of the system studied, which on the long-run should assist in the design of potent hBVR-A inhibitors[88]. Thus, MD serves as an important tool for not only refining the post-docking complexes, but also for revealing more appropriate binding modes of the ligands within the receptor structures.

**MD simulations and predicting the free energy of binding**

Molecular recognition is critical in several biochemical and biological processes [90]. Many biological processes are initiated by specific binding between two interacting entities in the cell. Although docking, combined by MD simulations, can provide a clear image of the shape complementarity between these entities at their binding interface, whether these interactions are significant or realistic requires an additional and essential piece of information, namely the free energy of binding, which is the driving force toward forming
this complex. Calculation of the binding free energy ($\Delta G_{\text{bind}}$), i.e., the free energy difference between the ligand-bound state (complex) and the corresponding unbound states of proteins and ligands, is used to quantify the affinity of a ligand to its target. Assessing the $\Delta G_{\text{bind}}$ of a series of ligands against a particular target can discern those ligands with higher binding affinities with the target. The $\Delta G_{\text{bind}}$ calculations are thus very important in drug design, and normally follow the docking-based virtual screening processes. A number of computational methods, from computationally rigorous thermodynamics pathways approaches to less complex end-point methods, have been developed for $\Delta G_{\text{bind}}$ calculations. The former methods include thermodynamic integration (TI) and free energy perturbation (FEP) methods; while linear interaction energy (LIE), Molecular Mechanics-generalized Born surface area (MM-GBSA), and Molecular Mechanics-Poisson-Boltzmann surface area (MM-PBSA) are end-point methods. Each of these methods has its own strengths and limitations, and their computational requirements and speed are inversely correlated with their accuracy.

TI and FEP methods are thermodynamic pathways approaches that are commonly employed for the calculation of relative binding free energies[91,92]. These methods are mainly based on the application of thermodynamic cycles and thus require the transformation of the system from the initial state to the final state through alchemical changes of the system energy function[91]. These methods involve change of a ligand $A$ into ligand $B$ in two states, such as solvent-only unbound state (of the ligands) and bound state (i.e., ligand-protein complexes). This will provide free energy changes for the unbound states ($\Delta G_{\text{unbound}}$) and bound states ($\Delta G_{\text{bound}}$) of the ligands [91]. It is also possible to mutate ligand $A$ to “nothing”, which in principle can provide absolute free energies of binding. Understandably, these methods (TI and FEP) demand multiple MD simulations and rigorous sampling of ligand, protein and solvent degrees of freedom. As a result, the thermodynamic pathways methods are in general able to provide very accurate estimation of the free energy of binding at the cost of high computational time[93,94]. For instance, the TI method coupled with MD simulations has been employed to identify potential huperzine derivatives with higher binding affinity towards the acetylcholinesterases[95]. Similarly, the FEP approach has also shown to predict more accurate binding free energies.
for a number of inhibitor-enzyme complexes\cite{93,96}. However, estimating the $\Delta G_{\text{bind}}$ values using these methods require large numbers of conformational samples, which in turn inflate the computational costs heavily\cite{93,94}. Given the need for enormous computational resources, these methods have mostly been applied for only small sets of ligand-protein complexes. Nevertheless, with increasing supercomputing capabilities and more improved methods, TI and FEP are gradually being involved in the SBDD pipeline, especially in guiding lead optimization\cite{92,97-100}. For instance, a recent work\cite{100} by a large team of authors, from Schrödinger, Nimbus, Columbia, Yale, and UC-Irvine, show that FEP calculations are able to make highly accurate affinity predictions across a wide range of ligands and targets. This work included fairly diverse sets of targets, such as BACE, CDK2, JNK1, MCL1, p38, PTP1b, Tyk2, and thrombin. The estimated binding free energies reported in this study were in very good agreement with the experiments. Indeed most of the values were within 1 kcal mol$^{-1}$, with only 9 out of 199 ligands studied deviated from their experimental values by over 2 kcal mol$^{-1}$\cite{100}.

A less rigorous alternative to thermodynamic pathways is the end-point approaches, which include methods such as LIE, MM-PBSA, and MM-GBSA. Unlike thermodynamic pathways approaches, these end-point methods sample only structures involved at either ends of the reaction pathways; that is, the free receptors (proteins) and ligands and the final ligand-protein complexes. The $\Delta G_{\text{bind}}$ in this approach can be written as,

$$\Delta G_{\text{bind}} = G_{\text{complex}} - (G_{\text{protein}} + G_{\text{ligand}})$$  (3)

The LIE method, developed by Aqvist et al\cite{101}, considers the process of a ligand ($L$) binding to a protein as a partition problem, in which the free ligand ($F$) from the bulk solvent medium is transferred to a solvated protein environment ($P$). Therefore, two independent MD simulations, one for the complex and the other for solvated ligand, needs to be performed in order to calculate $\Delta G_{\text{bind}}$ using the LIE method. Nonetheless, the reliance of the LIE on the end-points of binding makes it an attractive (and affordable) approach for lead optimization in drug discovery. A number of studies have employed LIE method for the computational analyses of inhibitors against a variety of targets. This includes
benzamide-based thrombin inhibitors[102], inhibitors of Mycobacterium tuberculosis H37Rv cytidine deaminase[103], sertindole analogues to block hERG potassium channel[104] and allophenylnorstatine molecules to inhibit Plm4 enzyme, a target for Plasmodium malariae[105], for instance. The LIE method has been shown to predict binding free energies with a root mean square error (RMSE) < 1 kcal mol\(^{-1}\) compared to the experimental values[106]. As indicated above, the thermodynamic pathways methods are also able to make predictions with a similar, if not better, accuracy range for different targets.

MM-GBSA and MM-PBSA are the two other well-established end-point methods that are most popular in SBDD. The two methods employ an implicit solvent model to account for the solvent molecules and use dielectric continuum models to obtain the electrostatic components for the solvation energy. The MM-PB(GB)SA \(\Delta G_{\text{bind}}\) can be estimated as,

\[
\Delta G_{\text{bind}} = \Delta E_{\text{MM}} + \Delta G_{\text{solv}} - T\Delta S
\]

Here, \(\Delta E_{\text{MM}}\) refers to the molecular mechanical energy and it is the sum of all energies from the bonded and non-bonded interactions. The solvation energy, \(\Delta G_{\text{solv}}\), is the sum of the polar and non-polar contributions of solvation. The polar solvation terms (\(\Delta G_{\text{PB/GB}}\)) are estimated using a Generalized-Born model or a Poisson-Boltzmann solver. The non-polar contributions are computed based on the size of the solvent-accessible surface area (\(\Delta G_{\text{SASA}}\)) in the ligand and protein. The final component of the \(\Delta G_{\text{bind}}\) equation (equation 7) is \(T\Delta S\), which corresponds to the conformational entropy changes in the reaction-product (i.e., protein-ligand complex), upon ligand binding.

The inclusion of conformational entropy (\(T\Delta S\)) in the MMPB(GB)SA calculations, in order to obtain absolute \(\Delta G_{\text{bind}}\), remains challenging. The accurate calculation of (\(T\Delta S\)) is computationally expensive and in many cases, its inclusion does not guarantee better accuracies in the final energies[46,107,108]. Rather, previous studies have shown that accounting for conformational entropy obtained from insufficient MD sampling has adversely affected the calculations[107,109]. For instance, Su et al[107] showed that the
accuracies of their MM-PBSA and MM-GBSA calculations for 16 known benimidazole inhibitors against *F. tularensis* Enoyl-ACP Reductase were significantly affected because of using different number of frames for their enthalpy and entropy calculations. The authors sampled 2400 frames from the MD trajectory used for their enthalpy calculations; however, due to limited computational resources, they only used 48 frames (evenly selected from the trajectory) for the entropy calculations. Therefore, it is important to have large numbers of MD snapshots to derive a reliable estimate of absolute $\Delta G_{\text{bind}}$, which can significantly increase the computational costs. As a result, many studies tend to neglect $T\Delta S$ and use the ‘relative’ $\Delta G_{\text{bind}}$ instead. Relative $\Delta G_{\text{bind}}$ energies can be predicted with a reasonable accuracy, and are generally sufficient to rank a group of compounds against the same target in SBDD [110].

Two strategies are commonly employed in MM-GBSA and MM-PBSA calculations: (i) the three-trajectory scheme and (ii) the single trajectory scheme[20,46,107,111]. In the former, three different MD trajectories that pertain to the ‘apo’ protein, free ligand and the ligand-protein complex (i.e., the end-point structures) are sampled for snapshots. In principle, this three-trajectory scheme provides more accurate results than the single-trajectory approach; however, it demands high computational costs [110,111]. On the other hand, the single trajectory scheme requires only a single MD simulation for the ligand-protein complex, which significantly reduces the required computational time [20,110-112]. Apart from the choice of strategy, there are a number of factors that may affect MMPB(GB)SA calculations, which includes length of simulations, choice of the force field, solute dielectric constants, solvent model, and the net charge of the systems. For instance, it has been argued that employing multiple short and independent MD simulations, instead of one long MD trajectory, can provide better $\Delta G_{\text{bind}}$ predictions [107,108,111,113].

There have been a number of studies that compared and tested the efficiencies of MM-GBSA and MM-PBSA in predicting accurate $\Delta G_{\text{bind}}$ energies for different ligand-protein complexes. Their general conclusion is that the accuracy of these methods tends to be system-dependent. For example, Hou et al.[110] found MM-GBSA to predict accurate
relative $\Delta G_{\text{bind}}$ for 59 ligands against six different protein targets in their study, when compared to MM-PBSA that outperformed the former in making the absolute $\Delta G_{\text{bind}}$ predictions. We have extensively applied MM-PBSA method for a range of studies, including screening and ranking of ligands against ERCC-XPA complex[41], understanding the binding mode of daclatasvir onto the NS5A viral protein [43] and binding of human programmed death-1 of t-cells with its ligands[83]. On the contrary, Oeheme et al[109] concluded that MM-GBSA performed better than the MM-PBSA in computing $\Delta G_{\text{bind}}$ of their ligand-HIV protease systems. Thus, it is clear that neither of these two methods is universally superior and the choice of the method should be made on a case-by-case basis. For example, Jordheim et al[114] combined, MD simulations, virtual screening and MM-PBSA based binding-free energy calculations, along with different experimental techniques, to identify potential inhibitors of ERCC1-XPF protein-protein interactions. The authors performed 20 independent virtual screening runs against the 20 XPF structures present in an NMR ensemble, after their MD equilibration. Top hits from each screening were extracted and then ranked them based on their binding free energies. From these results, 73 compounds were subjected to a range of experiments, including cytotoxicity assays, steady-state fluorescence and synchronous fluorescence experiments, and immunocytochemistry. The hits were evaluated on A549 and HCT116 cancer cells. Finally, one compound was found to exhibit promising activity in all the experiments and was also able to interact with the domain of XPF that is responsible for interacting with ERCC1, thus disrupting the protein-protein interactions. Thus, MD-based binding free energy calculations are helpful in guiding the hit identification stage. However, one of the significant drawbacks of both these methods is their inability to make accurate predictions for ligands with formal charges[109,111,115]. Hence, it is important to improve the existing methods or develop new methods which can account for charged ligands (including tautomers), which form a significant area of drug research.

**MD simulation and solvent dynamics analyses**

Computational analyses of structure and thermodynamic properties of water have recently become a useful tool in SBDD[116-119]. The properties of surface water molecules have been proposed to play important roles in molecular recognition and ligand-protein (and/or
protein-protein) interactions in solution[116,120]. Though small in size, water molecules are involved in a range of interactions, including H-bonds and van der Waals contacts[120]. Due to such interactions, it is often difficult to displace water molecules to facilitate the binding a drug. The energy involved in relocating water molecules between surface layer and bulk water, upon binding of a macromolecule (protein for instance) with another macromolecule and/or ligands, therefore, can significantly impact the overall free energy of binding[116-120]. Hence, the hydration patterns of a binding pocket can offer important insights into the properties of the pocket and also quantify the hydrophobic forces involved in the binding of small-molecule drugs with proteins. There are a number of \textit{in silico} tools that can help in extensive molecular descriptor analyses of solvation from the MD trajectories. These algorithms include WaterMap (from Schrodinger)[121], WaterFLAP (from Molecular Discovery)[122], SZMAP (from OpenEye)[123], GIST (in Amber)[120,124], WatMD (in-house tool of Novartis Inc)[116,125], SPAM (from GlaxoSmithKline)[126], STOW[127], WatClust[127], etc. Some of these methods, WaterMap[121], STOW[127] and WatClust for instance, are based on inhomogeneous fluid solvation theory (IST), where enthalpy is accounted directly from non-bonded interactions and entropy is calculated from a local expansion in terms of correlation functions[121].

For instance, the WaterMap program[121] initially clusters the water molecules (in the MD trajectory) based on their spatial distribution such that they form individual hydration sites. Subsequently the hydration sites are analyzed using IST to determine the enthalpy and entropy properties of water molecules within each site. This method has been successfully employed to gain insights into binding sites for various systems, including peptides binding to PDZ domains[128], the FKB12 protein[129], protease and kinase binding affinity[130,131] and the A2A GPCR[132]. For example, Beuming et al[129] employed WaterMap tool to analyze the hydration sites for a panel of twenty-seven different protein targets across a range of families. Initially, the authors[129] performed ~2 ns long MD simulation for each of the targets and the resultant trajectories were subjected to analyses with the WaterMap program. The results[129] revealed ~31,500 hydration sites in the targets, for which the authors calculated their thermodynamic information (such as free
energy, entropy and enthalpy). The authors further demonstrated that such thermodynamic properties of the hydration sites could be used to identify potential binding sites and evaluate their druggability[129]. It was found that clusters of high-energy solvation sites mostly inclined to be related with binding sites. However, Ramsey et al[120] notes that the IST-based methods are limited to the analyses of high-occupancy hydration sites and they do not describe significantly the hydration structures in low-density regions. As an alternative to these methods, the authors developed a tool named grid IST (or GIST)[120] and implemented it into CPPTRAJ toolset of AmberTools. GIST discretizes the integrals of IST onto a 3D-grid, which fills the binding pocket region and thus covering both high-density and low-density regions[120]. As a result, unlike the IST methods, the GIST is able to offer a smoothed map of water structure and the corresponding thermodynamic information for the complete region of interest. For instance, GIST analyses of the molecular host cucurbit[7]uril have revealed significant information about the hydration structure and thermodynamic properties in this receptor[124]. The results particularly revealed a toroidal region of high density hydration site at the center of the host’s nonpolar cavity. The results[124] also show that this specific hydration site, despite having high density of water molecules, is energetically and entropically not favorable. The authors relate this to the known ability of this receptor to bind external molecules with unusually high affinities[124]. Henceforth, a combination of MD simulations and explicit analyses of solvent dynamics are helpful to advance our knowledge about the effects of water molecules in structural biology and drug design[124].

**Constant pH molecular dynamics**

The ligand-protein complex formation not only leads to conformational changes in the structures of the proteins and/or ligands, but can also impact the pKₐ values of their charged side chains. The most common practice in molecular docking and standard molecular dynamics is to assign fixed protonation states for the protein residues, substrates and ligands, based on prior chemical knowledge. However, it is a known fact that the protonation states of a typical ionisable group involve dynamic processes that can alter the chemical environment during binding. Previous studies[133,134] noted that the pKa values of titratable residues may change due to a number of factors. This includes the solvation of
the group upon ligand-binding, electrostatic interactions between the ligand and protein, and structural reorganizations within ligand-protein complexes after binding. Thus the protonation states of ionisable amino acid residues and non-protein molecules (ligands and substrates) may be subjected to a change during the course of MD simulations. By preserving the protonation states, the MD simulations ignore any binding-induced pKₐ changes within the systems. This missing information can limit our complete understanding of the underlying biological processes.

Constant pH molecular dynamics (or CpHMD), has been developed for the computational prediction of pKa values[135] for ionisable residues in the biological systems under study (refer to Box 2). The early CpHMD approach employed GB solvent as the continuum aqueous environment and Langevin dynamics for the propagation through the non-solvent (or solute) trajectories [136]. However, this approach has been found not so accurate for many systems, particularly, when water molecules play an active role. Alternately, Donnini et al[137] developed a fully atomistic CpHMD method with λ-dynamics approach, which can be carried out in explicit solvents. This method allows for the dynamic change of protonation states of titratable groups, thus being able to predict the possible average protonation states at a given pH. This method samples the relevant configurations of the end states of titration groups, by considering the protonated as λ=0 and deprotonated as λ=1[137]. Given the importance of the protonation states of titratable groups in SBDD, it is suggested that a constant pH MD simulation be performed for the ligand-protein complexes before any production MD simulations are initiated. This way, the protonation states of the ionisable groups in the system can be accurately described.

More recently, there have been significant developments in improving the CpHMD [138-141]. For instance, attempts have been made to improve CpHMD using the replica exchange concept (vide supra). The basic idea is to perform simulations of different replicas at different pH values. After some set number of steps, the pHs are exchanged between the replicas so as to sample a wider range of protonation states[139]. This approach has been shown to greatly improve the convergence rate and accuracy of CpHMD simulations[140].
Limitations of MD

Classical MD simulations remain a valuable tool in drug design. They are helpful in understanding key molecular motions, energetics, ligand-protein interactions, receptor flexibilities, and conformational changes in the molecular systems, which facilitate the identification of potential candidates with higher affinities to targets. However, it is also important to acknowledge that MD also has some potential limitations and pitfalls, most particularly those concerning time limitations, force-field issues and quantum-effects[53,142].

Time limits and the sampling problem

At present times, typical MD simulations are carried out on systems containing hundreds to millions of atoms, and for several nanoseconds to microseconds timescales. Although these are impressive developments in the field (of MD), it is possible that such time limits may not be sufficient to relax the systems to study certain quantities. For instance, a number of physical properties of biological systems, such as protein folding, ligand binding and unbinding processes mostly occur at large timescales that are normally inaccessible using classical mechanics MD simulations. Furthermore, it is known that biological systems can get trapped in deep energy wells of their potential energy surfaces[143], which may result in sampling of insufficient and/or non-relevant conformations even from long MD trajectories[144]. Improper preparation of the initial structure or insufficient equilibration of the initial structure(s) can impact the quality of the MD results. Sampling (or) equilibration of an ensemble of structures, therefore, remains one of the key issues in MD simulations. Such challenges can be tackled by employing alternative strategies. One of the solutions is to apply an enhanced sampling MD approach[46,145], in which an additional bias, such as an external force, is applied to the system in order to explore the different potential energy surfaces. Although this strategy introduces some artefacts from external bias, it is useful to allow large-scale conformational changes in the systems within the affordable computational cost. Several enhanced sampling approaches have been developed, including metadynamics, replica exchange molecular dynamics (REMD), random acceleration molecular dynamics (RAMD), steered molecular dynamics (SMD)
and adaptive bias force steering (ABFS). There are a number of reviews, for example see references[145-147], discussing the applications of these methods in SBDD. Alternatively, coarse-grained MD (CG-MD)[148], which reduces the degrees of freedom in large systems by clustering groups of atoms into CG beads, has been developed to deal with large dynamic changes in more complex macromolecules.

**Force field issues and quantum effects**

Molecular mechanics force field employed in the simulation plays vital roles in defining the structural model of the studied system. Force fields are usually developed by combining available experimental data and the results from high-level *ab initio* calculations on small models that form larger systems, and hence they are fundamentally approximations[53,142]. Furthermore, force fields are parameterized such that they include several atom types describing varied situations of the same atoms (or functional groups). Due to such reasons, the transferability of force fields is restricted. Thus, results of MD simulations are reliable only as long as the potential energy functions (or force fields) mimics the forces experienced by the atoms in the real system under study[142].

Classical MD, because of its capabilities to handle large-size systems using affordable computational resources, has gained extraordinary popularity in SBDD. Classical approximations are mostly well-suited for non-reactive molecular interactions in biological systems[149,150]. However, they are not able to effectively describe the chemical reactions occurring in biological systems. For example, classical MD may not be able to offer great solution for understanding the reaction mechanisms of drug/substrate-protein complexes, chemical processes of proton transfer within active site, and binding/cleaving processes of certain covalently bonded ligands. In such cases, the use of quantum-mechanics (QM), which explicitly models the electrons in the system, becomes essential at the expense of computational time. In order to overcome this challenge, reactive force fields have been developed recently that allows chemical reactivity to be treated to some extent[53,149,150]. In reactive force fields, the interatomic potential defines chemical reactions by implementing a bond-order formulation. Within this scheme, the bond orders in the system are empirically calculated using interatomic distances between atoms during
MD simulation. Whereas, the electronic interactions driving chemical bonding are treated implicitly such that facilitating the modeling of changes in atom connectivity[149,150]. Recent review articles, for example see references [149,150], discuss various applications and challenges of such reactive force fields.

Another important challenge faced by classical MD is accounting electronic polarization, a significant quantum effect[142]. Within the classical MD framework, each atom in the system is assigned with a pre-set partial charge and is maintained throughout the simulation. Nevertheless, this is not always true, as the biomolecules are in general polarizable; meaning that the electron clouds encircling the atoms constantly shift in response to their chemical environment. Thus, it would be effective if the partial changes could be represented as a dynamic parameter, which is not the case with most of the current classical force fields. Realizing the importance of this challenge, there have been significant ongoing efforts to develop robust polarizable force fields for MD simulations[151]. Some of the current generation polarizable force fields include AMOEBA[152,153], CHARMM Drude and AMBER ff02[151]. Indeed, it is important to note that polarizable force fields also have their own challenges and should be used with caution. For example, these polarizable force fields are in general slower than non-polarizable force fields and, as a result, they are more vulnerable to sampling issues. Therefore, polarizable force fields may not be suitable systems where large conformational sampling plays important roles. Though having some weaknesses, the current polarizable force fields are promising. Given the importance of electrostatic interactions in biological systems, and with more developmental efforts underway, polarizable force fields will soon become an inevitable choice for classical MD simulations in future. There are some recent articles that discusses the current status and future directions for polarizable forces and MD simulations [151-153].

**Advanced hybrid QM/MM MD**

Although there have been significant efforts to fix the issues (concerning chemical reactivity and electronic polarization) within the classical MD framework, employing quantum-mechanical (QM) MD, which explicitly models the electrons in the system, has
been an alternative practical strategy in biomolecular simulations and SBDD. QM-MD generates dynamical trajectories by using the forces obtained from the electronic structure calculations that are performed at every time step of simulation. It is, therefore, able to accurately describe any reactions involving significant electronic effects such as electron correlation and electron polarization effects[154] [155]. Nevertheless, QM-MD simulations are extremely computationally intensive, which limits the practicality of applying this approach only to smaller sized systems (~10^2 atoms) and for limited time scales (~10^{-12} s)[156]. Hence, it was extremely important to find a mid-point that offers both ‘the chemical accuracy’ of QM-MD and ‘feasibility’ of MM-MD. To address this problem, Warshel and Levitt[157] introduced a state-of-the-art hybrid MD scheme popularly known as QM/MM. In this approach, a chemically reactive region in ligand-protein complex (mostly binding site residues and bound ligand) are treated with more accurate QM methods, and the rest of the system is described using MM force fields (Fig. 6). To date, a number of QM/MM implementations have been developed[158-160] and applied in many studies that focussed on large drug-protein and/or protein-protein systems. For example, in their recent study, Chen et al[161] employed QM/MM MD and QM/MM GBSA method for studying the interactions of benzamide inhibitors with trypsin. In this study, the authors treated the active site residues of the receptor and the inhibitors with QM methods (B3LYP/6-31G(d), PM3, PM6, and RM1) and the rest of the system with classical ff99SB force field and AMBER program. The study found that binding free energies calculated with the snapshots obtained from QM/MM MD trajectories displayed excellent agreements with experimental values[161]. In another study[162], QM/MM MD simulations revealed that the fourth ligand coordinating with the active site zinc ion in Acutolysin A enzyme is a water molecule, rather than a hydroxide anion, correcting a misconception from the low-resolution X-ray crystal structure. It was also revealed by a study that the QM/MM FEP approach outperformed the conventional FEP scheme in predicting accurate binding free energies for a set of fructose 1,6-bisphosphatase inhibitor [93]. Cui and co-workers[163] showed that a hybrid QM/MM-FEP approach could be used to predict accurate pKa values of biological systems. Thus, QM/MM MD simulations have the ability to offer accurate dynamic information that is significant in understanding the
structure-function relationships of proteins and their interactions with different classes of ligands, the key in drug discovery research.

Nevertheless, it is also important to acknowledge the fact that QM/MM MD simulations also have some clear pitfalls. One of the most important problems in QM/MM simulations is the treatment of the interface region that connects the QM part with that for MM, particularly if they are covalently bonded as in the case of ligand-protein systems. When a complete system is explicitly cut into QM and MM parts, then it will leave the former region with incomplete valences, which can lead to a failed QM treatment[164]. The most common strategy to overcome this issue is to cap the bordering QM residues, which underwent partition, with hydrogen atoms. But such hydrogen capping introduces new atoms into the QM region those were not originally present in the real system, which may lead to some artefacts[164]. Furthermore, QM/MM MD simulations of large protein-ligand systems are still very computationally expensive. Hence, they can only be applied to select systems in drug design, such as for those top-ranking hits filtered from thorough virtual screening and classical MD simulations, where follow-up details about key ligand-protein interactions for pharmacophore modelling are computationally justified.

Perspectives on integrating the computational approaches

Last decade has observed tremendous developments in the field of molecular modelling and drug design methods. As discussed above, a number of modelling and MD-based approaches are available to help in the modern drug design and discovery efforts. Nevertheless, how we integrate these methods, along with other in silico approaches and experiments, is important for increasing our chances of identifying more promising hits from the chemical pool of compounds. Although there are no specific set of rules on how these methods should be combined, extensive knowledge and experience gained over years have provided some logical strategy of implementing them. In Fig. 7, we present a more simplified and practical work-flow that assembles classical MD, binding free energy calculations and QM/MM methods at various stages leading from hit-identification to lead optimization. For instance, the need of classical MD simulations could first arise upon having one (or) more initial three-dimensional structures either from the PDB or through
molecular modelling methods. Because most of these methods are single snapshots of the
target, long classical MD simulations (usually few hundred nanoseconds time scale) are
required so that large conformational changes in the target can be captured during the
simulation. At this stage, the user needs to make a number of cautious choices, such as the
software program, the empirical force-field and simulation parameters that are suitable to
perform stable MD simulations.

Following this stage, different clustering algorithms, such as RMSD-based clustering or
PCA analyses, can be performed to sample the dominant conformations of the target from
the MD trajectory. The target conformations obtained from the MD simulations will serve
the purpose of addressing the protein flexibility concerns during the subsequent virtual
screening procedures. Indeed, there are some computational methods to identify possible
cryptic binding sites from these ensemble of target structures and target them during
screening. Following the docking and scoring, MD simulations can again be performed on
the ligand-protein complexes, in order to refine the complexes and calculate their binding
affinities. At this point, the user must make a number of careful selections, including the
length of MD simulations, force-field for simulation and methods for binding-free
estimations. Usually, it is suggested that short MD trajectories (~1-2 ns long) are collected
for each ligand-target complexes and use them for free energy of binding calculations. The
end-point methods, MM-PBSA or MM-GBSA, are mostly popular for these calculations;
although other methods, such as FEP, are gaining popularity in the field. Once the high-
ranking compounds are identified, they can be experimentally tested using different kinds
of assays. At the current stage, it is hoped that a 5-10% of hit rate (during experimental
testing) can be achieved by incorporating rigorous computational modelling and pre-
screening protocols; although this might not be the case always.

Whatever the results from the experiments are, positive or negative, they can be back-fed
to the computational protocol so as to improve it for subsequent phases of screening. For
example, if the results are negative (meaning no significant hits were identified), then
probably the lengths of initial MD simulations can be increased so as to increase the sample
size for target conformations; and/or increase the chemical search space by increasing the
numbers of compounds in the libraries; and/or refine the parameters in the docking protocols; and/or increase the length of MD simulations for binding free energy calculations; and/or even change the method used for free energy estimations. In the event of obtaining good hits from the experiments, then the user might wish to perform extended MD simulations (now for hundreds of ns) to understand the key dynamic interactions between the hits and the targets. Binding free energy methods (or) other enhanced sampling MD methods can also be applied at this stage so as to gain in-depth knowledge about binding mode(s) of the hits. Based on these information, an effective pharmacophore model or QSAR model (and/or experimental SAR) can be developed and implemented in subsequent screening protocols. When one or more promising hits, those showing attractive inhibition potentials, promising immunological activity and also non-toxic profiles, are identified, then complexes of such hits can be taken forward for more advanced and computationally expensive QM/MM simulations. At this stage, the user must be cautious in defining the QM segment and MM segment in the system and also choose a cost effect (but also accurate) QM model and a suitable MM force field for treating classical segment. The choice of software program is also a key, as using the one that scales well could be helpful to run the QM/MM simulations for large timescales. Such rigorous hybrid simulations can offer extra-ordinary insights about the reaction mechanisms involved between the selected hit(s) and the target(s). Understanding the reaction mechanisms can be useful towards achieving a better lead compound(s). Those leads showing promising in vitro and in vivo activities can be taken to further lead optimization and lengthy and expensive clinical trial stages. Indeed, off-target interactions of drug is yet another important challenge facing the community; and computational methods are also helpful to address this challenge, which is not discussed in this review.

The potentials of combining all the computational methods discussed in this review can be best demonstrated, for instance, by a series of studies[86,165-169] carried out on a bacterial enzyme, namely bacterial MurD ligase. A team of scientists from the National Institute of Chemistry, Slovenia, along with their collaborators, have carried out a number of studies on the enzyme. This includes, studying the domain flexibility using MD simulations followed by drug design efforts[165-167], post-docking refinements of the complexes
using MD approaches[167,170], understanding the reaction mechanism(s) of the identified
hit-enzyme complexes using QM/MM methods[169] and free energy calculations to
understand the binding of inhibitors to the MurD ligase and further drive the design
processes[86,168]. In one of the preliminary studies[165], the authors performed extensive
targeted MD (TMD) simulations in order to gain some insights into substrate binding
process and also the structural changes in the enzyme during the transition(s) between the
experimentally determined closed and open states[165]. In another study[166], the authors
used this information to perform off-path simulation to obtain a relative energy comparison
pathway of the two TMD-generated closing pathways. This study also discerned the
pathway which had high-energy demands to perform the biochemical processes[166]. The
authors claim that the results from their studies agreed well against the experimental
findings[166]. Subsequently, the authors selected three MurD ligase conformations from
their MD simulations and used them for two-stage docking-based virtual screening
study[167]. The screening identified a panel of promising hits, out of which one of them
(an aminothiazole class inhibitor) was confirmed experimentally to act against dual targets,
MurD and MurC. The authors re-docked this inhibitor against all the target structures and
performed extended classical MD simulations to gain atomistic insights into the ligand-
target interactions[167]. The authors also identified another inhibitor class of benzene-1,3-
dicarboxylic acid 2,5-dimethylpyrrole derivatives that showed dual MurD/MurE inhibition
properties[170]. In the follow-up study, the authors performed extended MD simulations
of this inhibitor-MurD complex to explore their geometrical behaviours. Later, they also
performed binding free energy calculations using liner interaction energy (LIE) method
that described the energetic behaviour and binding affinity of the compound[170]. Using
the information gathered from these studies, the authors again developed new
pharmacophore models and performed new phase of virtual screening to only discover
novel set of compounds that showed promising effects in the experiments[170]. Similar
combination of MD and LIE-based binding free energy calculations were also carried out
for Furan-based benzene mono- and dicarboxylic acid derivatives against the bacterial Mur
ligases[86]. In their ongoing computational and experimental efforts to design drugs for
Mur ligases, the authors also performed advanced QM/MM simulations[169], using
B3LYP level of QM theory and CHARMM MM force fields, of the experimental structure
of MurD in the PDB (code: 2UAG). This QM/MM study[169] was useful to understand about the tetrahedral intermediate formation in the enzyme complex, which was not known until then[169]. Hence, the bundle of studies by these authors demonstrate how a series of computational studies (along with experiments) can be set up to advance our knowledge about the structure-properties of a specific target and make progress towards achieving the goal(s) of drug discovery.

Concluding remarks

It has been 38 years since the first molecular dynamics (MD) simulations of bovine pancreatic trypsin inhibitor were carried out for 9.2 picoseconds. Since then, there has been tremendous growth in supercomputing power and significant developments in the accuracy and efficiency of MD-based computational methods. And MD is now well established as an important contributor to drug design and development. With current capacities, MD simulations can be employed for larger biological systems and for microsecond timescales. Such longer classical MD simulations help in effective treatments of the induced-fit effects of the drug binding onto receptors, and can be used to realize optimal drug-receptor binding modes and collect larger conformational samples of the complexes that allow more accurate binding free energy estimations. Alternate versions of classical MD methods, such as CpHMD and enhanced sampling MD approaches, allow tracing chemical changes and other intricate biological events, which normally occur within ligand-protein complexes but are not observed within the practical limits of classical MD simulations. On the other hand, advanced hybrid QM/MM MD methods are useful in revealing the actual reaction mechanisms occurring at the ligand-binding site of the receptor, which are important to design potent ligands that could trigger effective inhibition of the disease targets. Thus MD approaches offer wide range of opportunities and capabilities. Assembling them appropriately with other in silico approaches and experiments can enhance the possibilities of identifying more credible hits that can eventually become effective next-generation drugs to serve human population.
Table 1: Some of the recent studies that employed MD simulations for various applications, such as accounting protein flexibility, post-docking structure refinement and binding free energy calculations, on different targets.

<table>
<thead>
<tr>
<th>Enzyme/Target protein</th>
<th>Reference</th>
<th>Enzyme/Target protein</th>
<th>Reference</th>
<th>Enzyme/Target protein</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleoprotein of Influenza A virus</td>
<td>[171]</td>
<td>Enzyme/Target protein</td>
<td>Reference</td>
<td>Enzyme/Target protein</td>
<td>Reference</td>
</tr>
<tr>
<td>AcrB transporter</td>
<td>[173]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M3 muscarinic Acetylcholine receptor 3</td>
<td>[175]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isomerase</td>
<td>[176]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPCR</td>
<td>[178]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDM2-p53</td>
<td>[41,80]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α-Spectrin SH3 protein</td>
<td>[181]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nalp domain</td>
<td>[183]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hERG</td>
<td>[40]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melanocortin(rep exchng)</td>
<td>[188]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Histone deacetylases</td>
<td>[46,73,191]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glycoprotein</td>
<td>[194]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lysozyme</td>
<td>[196]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mad2</td>
<td>[199]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MurD Ligase</td>
<td>[167]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Giardia duodenalis 14-3-3</td>
<td>[202]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
</tr>
</tbody>
</table>
Figure captions:

Fig. 1: A schematic representation of induced effects of ligand binding to receptor.

Fig. 2: Different categories of methods employed for accounting ligand and receptor flexibilities in molecular docking.

Fig. 3: Different X-ray crystal structures of histone deacetylases 8 (HDAC 8) showing different conformation of binding site pockets. An overlap of the structures is also shown as a ribbon structure.

Fig. 4: Ensembles of structures sampled from long MD trajectories. Twenty-eight structures of MDM2 protein sampled from 50 ns long MD trajectory (a) and forty-five structures of hERG ion channel captured from 500 ns long MD simulation showing flexibilities of backbone region (b) and side-chain dynamics (c). In (a), the structures of holo-, and apo- trajectories of MDM2 protein are shown in green and blue, respectively. Wherein (b) and (c), the colors indicate flexibility of the concerned segments in the dominant conformations. The representative conformation of the target is shown in Red and the other conformations in the clusters are provided in colours ranging between red to blue.

Fig. 5: Binding mode of Daclatasvir with the NS5A protein[43]. The bound drug is shown as a green-colored stick representation and the protein residues are displayed as white sticks (a). The binding sites within the NS5A receptor is also provided as a surface representation (b).
Fig. 6: A QM/MM model of human acetylcholinesterase where residues that can be treated under QM is shown as ball-and-stick and the rest of the systems shown as surface representation and cartoon can be treated with MM force field.

Fig. 7: A simplified and practical workflow for molecular modelling and drug design. This work flow lists a sequence of steps that provides an overview of how MD approaches can be stacked along with other modelling and experimental procedures during the drug design and discovery efforts. In addition, a number of key decisions that needs to be taken during each of the modelling and MD stages are also listed. It is important to insist that this work flow in no-way tries to underestimate the roles of experiments in drug discovery. Rather it only tries to highlight the roles of computational approaches, as discussing experimental techniques in detail is not within the scope of present review.
References


Nair, P. and Miners, J. (2014) Molecular dynamics simulations: from structure function relationships to drug discovery. *In Silico Pharmacology* 2 (1), 1-4


Campbell, A.J. et al. (2014) Ensemble-Based Docking Using Biased Molecular Dynamics. *Journal of Chemical Information and Modeling* 54 (7), 2127-2138


Ma, B. et al. (1999) Folding funnels and binding mechanisms. *Protein Engineering* 12 (9), 713-720


Diskin, R. et al. (2008) A Novel Lipid Binding Site Formed by the MAP Kinase Insert in p38α. *Journal of Molecular Biology* 375 (1), 70-79


Vijulatha, M. et al. (2014) Induced fit docking, pharmacophore modeling, and molecular dynamic simulations on thiazolidinedione derivatives to explore key interactions with Tyr48 in polyol pathway. Journal of Molecular Modeling 20 (7), 1-14


Michel, J. et al. (2009) In Silico Improvement of β3-Peptide Inhibitors of p53•hDM2 and p53•hDMX. Journal of the American Chemical Society 131 (18), 6356-6357


Beuming, T. et al. (2009) High-energy water sites determine peptide binding affinity and specificity of PDZ domains. *Protein Science* 18 (8), 1609-1619


De Vivo, M. et al. (2016) Role of Molecular Dynamics and Related Methods in Drug Discovery. *Journal of Medicinal Chemistry*


Han, Y. et al. (2016) Development, applications and challenges of ReaxFF reactive force field in molecular simulations. *Frontiers of Chemical Science and Engineering* 10 (1), 16-38


170 Perdih, A. et al. (2014) Benzene-1,3-dicarboxylic acid 2,5-dimethylpyrrole derivatives as multiple inhibitors of bacterial Mur ligases (MurC-MurF). *Bioorganic & Medicinal Chemistry* 22 (15), 4124-4134

171 Tarus, B. et al. (2012) Molecular Dynamics Studies of the Nucleoprotein of Influenza A Virus: Role of the Protein Flexibility in RNA Binding. *PLoS ONE* 7 (1), e30038


174 Feig, M. (2016) Local Protein Structure Refinement via Molecular Dynamics Simulations with locPREFMD. *Journal of Chemical Information and Modeling* 56 (7), 1304-1312


Pan, D. et al. (2016) Computational study on the drug resistance mechanism of hepatitis C virus NS5B RNA-dependent RNA polymerase mutants to BMS-791325 by molecular dynamics simulation and binding free energy calculations. *Chemometrics and Intelligent Laboratory Systems* 154, 185-193


Lu, H. et al. (2014) Binding free energies for nicotine analogs inhibiting cytochrome P450 2A6 by a combined use of molecular dynamics simulations and QM/MM-PBSA calculations. *Bioorganic & Medicinal Chemistry* 22 (7), 2149-2156


Mukhametov, A. et al. (2014) Allosteric pocket of the dengue virus (serotype 2) NS2B/NS3 protease: In silico ligand screening and molecular dynamics studies of inhibition. *Journal of Molecular Graphics and Modelling* 52, 103-113


Poongavananam, V. et al. (2014) Binding free energy based structural dynamics analysis of HIV-1 RT RNase H-inhibitor complexes. *Integrative Biology* 6 (11), 1010-1022


Wang, K. et al. (2015) Hierarchical Conformational Analysis of Native Lysozyme Based on Sub-Millisecond Molecular Dynamics Simulations. *PLoS ONE* 10 (6), e0129846


Shao, S. et al. (2014) Dual-inhibitors of STAT5 and STAT3: studies from molecular docking and molecular dynamics simulations. *Journal of Molecular Modeling* 20 (8), 1-17

Li, C. et al. (2014) Molecular Dynamics Simulation on the Conformational Transition of the Mad2 Protein from the Open to the Closed State. *International Journal of Molecular Sciences* 15 (4)


Cau, Y. et al. (2015) Molecular Dynamics Simulations and Structural Analysis of Giardia duodenalis 14-3-3 Protein–Protein Interactions. *Journal of Chemical Information and Modeling* 55 (12), 2611-2622

Wang, Q. et al. (2016) Exploring the mechanism how AF9 recognizes and binds H3K9ac by molecular dynamics simulations and free energy calculations. *Biopolymers* 105 (11), 779-786
### BOX 1: Important quantities in MD analyses

**Root mean square deviation (RMSD)**
RMSD is a measure of the average deviation or distance between the atoms when three-dimensional structures are superimposed on each other. When analysing an MD trajectory, this value (or RMSD) could be a very important quantity that is useful to trace how much the structure that underwent MD simulations has deviated from its’ starting structure.

**Interaction energy**
The interaction energy is the amount of energy that is caused by the interaction(s) between two residues (or objects) and its’ contribution towards the total energy of the system. Interaction energies between different amino acid residues from the target and the bound ligand could make significant impact on the binding affinity of the complex. Thus, identifying the key residues that possess high interaction energy against the ligand is important in binding mode analyses.

**Interaction distance**
It is a minimum distance between two non-bonded residues of proteins or between residues and ligand that could affect each other, thereby impacting the total energy of the system.

**Correlation functions**
They are mathematical descriptors that connect the properties of protein structures with that of their significance. Thus correlation function remains an important tool for protein structure analyses from the MD trajectories.

**Radial distribution function (RDF)**
RDF is a quantity that describes the average radial packaging of atoms in a system and can be calculated by constructing normalized histograms of atom pair distances with respect to an ideal gas.

\[
g(r) = \frac{n(r)}{4\pi r^2 \rho \Delta r}
\]

Where, \(n(r)\) is the number of atoms in a shell of width \(\Delta r\) at distance \(r\) and \(\rho\) is the mean atom density. This quantity can be useful, for instance, to identify how many waters are coordinating with a metal ion in the active site of the protein during the course of MD simulation.

**Hydrogen bond (H-bond)**
The electrostatic force that attracts the hydrogen attached to one electronegative atom to another electronegative atom holding lone pair of electrons. Thus, identifying the number of H-bonds between the bound ligand and its’ surrounding amino acid residues of the protein is one of the key step while analysing the MD trajectories.
**BOX 2: Glossary of terms**

**Virtual screening**
Virtual screening (or *in silico* screening) is a computational approach employed in structure-based drug design to screen a library (or libraries) of small-molecules against the desired protein target in order to rank them based on their affinities to the concerned binding site of the target.

**Molecular docking**
A method to predict the favoured binding orientations between two molecules to form a stable complex

**Scoring function**
Mathematical method to quantify the interactions between two molecules when they are docked together.

**Shape matching**
A sampling method that uses receptor-complementarity as a criterion for identifying the ligand binding conformations

**Stochastic algorithms**
A sampling method that incorporates random changes to the ligand in transitional, rotational and conformational space to identify the most suitable ligand binding conformation

**Systematic search**
A sampling method that utilizes all degrees of freedom to sample the ligand binding conformations

**Induced-fit effect**
Conformational changes in an enzyme triggered by the interactions with (or binding of) small molecules or other proteins.

**Periodic boundary condition (PBC)**
Periodic boundary condition (or PBC) is a method employed in MD simulations to eliminate the issues concerning boundary effects, arising from finite size, by treating the system as infinite with the help of a unit cell.

**Free energy of binding**
Within the context of ligand-protein complex in drug design, the free energy of binding is defined as the free energy difference between the ligand-bound state (complex) and the free unbound states (free protein and free ligand).
**BOX 3: List of abbreviations**

CADD: Computer-aided drug design  
SBDD: Structure-based drug design  
MD: Molecular dynamics  
PDB: Protein data bank  
QM/MM: Quantum Mechanics/Molecular Mechanics  
IFD: Induced-fit docking  
FEP: Free energy perturbation  
LIE: Liner Interaction Energy  
TI: Thermodynamics Integration  
MM-GBSA: Molecular Mechanics-generalized Born surface area  
MM-PBSA: Molecular Mechanics-Poisson-Boltzmann surface area
Molecular modelling (Homology, threading, de novo)

Target structure

Docking & Scoring

Ligand and Protein as flexible bodies

Ideal "best-fit" scenario, where both protein and ligand are flexible.

However, current docking protocols mostly account for only ligand flexibility.
1. Prepared initial structure from PDB (or) molecular modelling approach(es)

**Important steps & decision:**

1. If the structure is available in the PDB and if yes,
   - Whether the structure is complete without any missing residues or segments?
   - Whether multiple conformations of the target are available in the PDB?
2. If the structure is not available, then it has to be modelled, and in this case
   - What modelling approach(es) and software program(s) are needed?
   - If a homology model template structures are available to model the 3D structure of the target?
   - Is the quality of the modeled structure reasonable enough to initiate the next stage?

2. Classical MD Simulation to gain insights into the structure of the target and to identify possible druggable pocket(s)

**Important steps & decision:**

1. Which software program and force-field are suitable for the MD simulation of the concerned target structures. Does your choice is also supported by the previous work(s) in the literature?
2. What parameters (such as cell dimension and temperature) are suitable for the system?
3. What type of water model is appropriate for the system and also compatible with the selected force-field?
4. How long should the system be equilibrated to obtain a fairly good starting configuration for the production runs? This decision is always made by verifying the behavior of the system and other physical quantities, such as temperature, pressure and different energies during the equilibration.
5. What will be the length of production MD simulation and it is sufficient to sample various conformational states of the proteins and also capture the significant dynamics related to ligand binding?

3. Clustering and selection of dominant conformations to account protein flexibility and preparation of ensemble of target structures.

**Important steps & decision:**

1. Which clustering algorithm needs to be used?
2. Has the trajectory from the preceding MD simulations sampled various dominant conformations that is sufficient for accounting protein flexibility in the further drug design efforts?

4. MD simulation and binding free energy calculations

**Important steps & decision:**

1. What software program, force-field and parameters are suitable for the MD simulation? These choices usually remain close to what have been used in MD simulations performed at previous stages, if any.
2. How long should the MD simulations need to be run that could produce fairly accurate binding free energies? Note, as discussed in this article, there have been several debates on the choice of one long MD trajectory or multiple short MD trajectories for this purpose.
3. What binding free energy method is suitable and also affordable for this research? Guidance from literature can help in making this choice.

5. MD simulation of promising hits identified from experimental assays

**Important steps & decision:**

1. What software program, force-field and parameters are suitable for the MD simulation? These choices usually remain close to what have been used in MD simulations performed at previous stages, if any.
2. How long should the MD simulations need to be run to refine the post-docking complexes and also to capture the significant dynamic interactions between the target and the bound ligand.

6. Employ sophisticated binding free energy methods to identify the best binding mode of the hit.

**Important steps & decision:**

1. What binding free energy method is suitable and also affordable for this research? Guidance from literature can help in making this choice.
2. Can the predicted binding mode(s) be related to the range of activity seen in experiments?

7. Advanced QM/MM MD simulations to gain deeper insights about the reaction mechanisms involved between the selected hit(s) and target(s)

**Important steps & decision:**

1. Is the selected ligand-target complexes the best choice to initiate expensive QM/MM MD simulations?
2. Do previous information about similar interactions and their reaction mechanisms available in the literature?
3. What would be the optimal segment of the complex that would require high-level QM treatment?
4. What is the QM/MM protocol and software program available to perform these simulations?
5. Is access to high-performance computers available to carry out these expensive MD simulations?
6. What would be the length of the MD simulations?