Chapter 27

Applications of molecular dynamics simulations in drug discovery

Xubo Lin

Institute of Single Cell Engineering, Key Laboratory of Ministry of Education for Biomechanics and Mechanobiology, Beijing Advanced Innovation Center for Biomedical Engineering, School of Biological Science and Medical Engineering, Beihang University, Beijing, China

1 Introduction

Molecular dynamics (MD) simulation is one of the powerful computer simulation methods for dealing with multibody motions of atoms. The positions, velocities, and forces of atoms/particles are updated every fixed time period (time step) by numerically solving Newton's equations of motion. Forces and potential energies between atoms/particles can be calculated using predefined potential functions (force field), which are the core of MD simulations. Current types of force fields include polarized force fields, all-atom force fields, united-atom force fields, and coarse-grained force fields. Among them, polarized force fields have the highest resolution, which usually introduces additional virtual sites to effectively capture the charge polarization phenomenon. As for all-atom force fields, they adopt a fixed charge model for all atoms. However, the highly vibrating hydrogen atoms restrict the time step of all-atom MD simulations. Hence, in united-atom force fields, hydrogen and carbon atoms in methyl groups and methylene bridges were treated as one interaction site, which significantly reduced the number of atoms and thus calculation. Coarse-grained force fields further map several heavy atoms plus hydrogen atoms into one interaction site with the relative lowest resolution but the highest computing ability. Generally, these force fields of different resolutions are widely used to solve scientific problems on different levels in biophysics and physical chemistry.

In MD simulations, the trajectories of atoms and molecules are recorded in real time, which allows us to visualize the atomic details of the complex biological processes. On the other hand, some macroscopic thermodynamic properties can be determined using statistical physical methods over MD trajectories, which are also measurable in experiments and thus allow direct comparison between MD simulations and experiments. In addition to these advantages, some disadvantages limit the accuracy and applications of MD simulations. First, no force field is absolutely accurate; they need to be continuously improved through new benchmark theoretical calculations and experiments for higher accuracy and broader applications. Second, limited computing ability greatly limits the timescale and length scale of MD simulations, which are improved to some extent by a series of enhanced sampling techniques and multiscale models. Third, although proper selections of algorithms and parameters can reduce system errors, it still cannot completely avoid cumulative errors in numerical integration. In other words, combining with the cellular/molecular experiments, it will be possible to make full use of the advantages of MD simulations to promote drug discovery.[1]

2 Identification of protein conformation ensemble and drug binding site

As early as 1977, the first MD simulation of the small protein bovine pancreatic trypsin inhibitor was performed by McCammon, Gelin, and Karplus,[2] which revealed the fluid-like characters of the protein interior for the first time. The duration of this simulation was only 9.2 ps, but it opened a new era of employing MD simulations in molecular biology and structure-based drug discovery. With the rapid development of the computing ability, the general affordable timescale that can be achieved in all-atom MD simulation is significantly improved to the level of microseconds (μs), making the sampling of protein conformational space much more reliable.

Advances in Protein Molecular and Structural Biology Methods. https://doi.org/10.1016/B978-0-323-90264-9.00027-1

2.1 Identification of protein conformation ensemble

In computational structure-based drug discovery, the structure of the target protein is a prerequisite. However, only a limited number of proteins have experimental crystal structures. If the target protein has no crystal structure, but the crystal structure of its homologous protein is available, homology modeling[3] can be employed to obtain the structure of the target protein. For other proteins, their structures can be predicted using webservers such as Robetta[4] and I-TASSER.[5] With the obtained structures, the subsequent molecular docking usually considers proteins as rigid or semirigid structures, which failed to incorporate the target flexibility and consider critical conformational changes of the target protein accompanied by the binding of small molecules. As mentioned above, MD simulation provides a powerful tool to sample the conformational space of the target protein. Molecular docking based on multiple crystal structures or MD structures, which better captures the flexibility of the target protein and is essential for the target protein with significant structural flexibility, are usually more successful in picking out functional small molecule drug candidates than that based on a single structure.[6–10] However, it sometimes takes significant computational resources for all-atom MD simulations to obtain sufficient protein conformation ensembles for the subsequent ensemble docking. Since coarse-grained models (e.g., Martini[11,12]) can well reproduce the structural and dynamical properties of proteins as atomistic simulations and are much faster than atomistic models, long timescale coarse-grained MD simulations can be used to obtain the representative conformations of the target protein, which can be converted into atomistic models using tools such as *backward.py.*[13]

2.2 Identification of drug binding site

For the ensemble docking, the docking calculations will be repeated for each selected structure. Considering the large databases of the small molecule compounds, global docking requires significant computational power. Hence, it is essential to identify the exact drug binding sites on the surface of these protein structures, which can significantly reduce the computation for the docking calculations. There are many useful bioinformatics tools (e.g., fpocket,[14] ConCavity[15]) that can be used to identify drug binding sites. However, most of these tools are mainly based on the geometry structure of the target protein. Probe-based MD simulation provides an alternative way to uncover drug binding sites from the perspective of interaction energy and statistics. In probe-based MD simulations, probes are usually small molecules (e.g., benzene,[16,17] isopropanol[18]), which are treated as cosolvent molecules. One probe or multiple probes[19,20] can be used in these simulations. The large quantity of probes in the simulation systems enable their sufficient interactions with the target protein and thus the identification of preferred interaction sites. Hence, based on the interaction energy and probe–protein contact statistics, we can predict the drug binding sites robustly. Usually, the properties of these probes are comparable to the chemical functional groups of small-molecule ligands. In certain cases, the replacement of these probes with the corresponding ligands can more directly detect and characterize the expected drug binding sites.[21]

3 Modeling protein-drug interactions

When the structure and drug binding sites of the target protein are determined, high-throughput small molecule screening and validation will be the next critical aspect in the drug discovery process (Fig. 1). Through high-performance computing, we can rank all the tested compounds according to the docking score. However, the top molecules in docking calculations do not necessarily mean the best molecules in action. Because some intermolecular interactions such as solvation effect, entropy effect, and halogen bonding are not accurately described in the scoring functions of molecular docking, it fails to give accurate binding free energies. Hence, results from the docking calculations need to be further validated. MD simulations may play critical roles for the purpose.

3.1 Molecular docking using MD force field functions

As mentioned above, one of the shortcomings of molecular docking is the improper description of intermolecular interactions. Compared with the scoring functions of docking calculations, the all-atom MD force field can deal with the intermolecular interactions much more accurately. Hence, one of the applications of MD simulations in modeling protein–drug interactions is to develop the scoring functions using MD force field functions for molecular docking. For example, Wu et al. developed an MD simulated-annealing-based algorithm using CHARMM force field, CDOCKER[22], to perform molecular docking, where the target protein is kept rigid, and the small molecule ligand is entirely flexible. The final energy minimization is used to refine the docking poses. Compared to the explicit all-atom representation of the target protein, the grid-based treatment does not significantly reduce the docking accuracy but dramatically improves the docking efficiency. Grosdidier et al. launched a new software, EADock,[23] which uses a hybrid evolutionary algorithm with two fitness

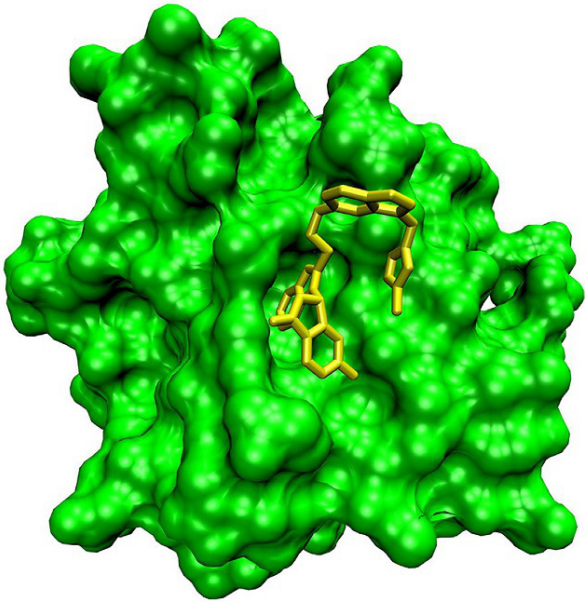

FIG. 1 Protein–drug interactions. Schematic for protein–drug interactions using KRAS as the model protein and F0K as the model ligand (PDB ID: 6GJ8). (No permission required.)

functions to identify binding modes with high accuracy. This software is interfaced with the CHARMM software package to deal with the coordinates and calculate the energies. The RMSD between the docking pose and the crystal structure is lower than 2 Å, which validates the efficiency of the sampling strategy in EADock. It is worth mentioning that the docking calculations using MD force field functions can deal well with the intermolecular interactions, but they fail to capture the entropic changes upon ligand binding, which restricts the final accuracy of the docking calculation.[24]

3.2 Long timescale MD simulations

For classical MD simulations, increasing the simulation time is an easy and effective way to increase the sampling abundance for protein–drug interactions. However, this is very expensive with all-atom models, which most research groups cannot afford. Only limited groups[25,26] can perform μ*s*- or even *s*-scale all-atom unbiased MD simulations to investigate the interactions between the target protein and drug molecules, which provides direct information for the binding kinetics and thermodynamics. These MD-derived observables allow straightforward comparisons between simulations and experiments. A study of this kind offers a valuable benchmark for testing different methods in characterizing drug-binding properties. However, the required enormous computing resources hinder its wide applications.

In order to overcome the shortcoming of long timescale all-atom MD simulations, one solution is to apply a coarse-grained model for this. Compared to the all-atom model, the coarse-grained model usually maps several heavy atoms into one interaction site, which significantly reduces the amount of particles in the simulation and the total computation. Hence, the coarse-grained model provides an excellent alternative way to achieve long timescale and high-throughput MD simulations. Coarse-grained MD simulations even do not need the prior knowledge of the binding site and can be used to study ligand binding kinetics, affinity, and pathways. For example, based on the latest Martini force field,[27] Souza et al. coarse-grained ligands including benzene (target: T4 lysozyme), adenosine/caffeine (target: adenosine A_{2A} receptor, $A_{2A}R$), obeticholic acid (target: farnesoid X receptor, FXR), dasatinib (target: proto-oncogene tyrosineprotein kinase, c-Src), baricitinib (AP2-associated protein kinase 1, AAK1), and performed a series of unbiased millisecond sampling of these ligands and their target proteins.[28] From these long timescale coarse-grained MD simulations, the ligand-binding kinetics as well as the pathways from one binding site to another binding site can be directly obtained from the trajectories. Besides, ligand binding free energies can also be achieved based on the principles of statistical physics. And the back mapping of coarse-grained conformations into atomistic conformations offers a level of accuracy comparable to atomistic simulations. However, an important issue that still needs to resolve is the popularization of ligand/protein docking methods with coarse-grained MD simulations.[29] That is to develop an automated approach for coarse-graining small molecules, forming a coarse-grained small molecule database for coarse-grained molecular docking.

3.3 MD simulations with enhanced sampling methods

For MD simulations, no matter which resolution model (atomistic or coarse-grained models) is used, sufficient sampling over the conformational space is always critical to obtain reliable statistical results. Especially when there is an energy barrier between two states of the simulation system, which is too high to cross, we can hardly sample the high-energy state. This will directly affect the statistical results obtained from MD simulations. To overcome this issue, various enhanced sampling methods have been developed.

3.3.1 Ligand Gaussian accelerated molecular dynamics simulations

Gaussian accelerated molecular dynamics (GaMD) is an enhanced sampling method developed by Miao and McCammon in 2015,[30] which adds a harmonic boost potential following the Gaussian distribution to smoothen the system's potential energy surface for accelerating barrier crossings. By performing cumulant expansion to the second order, GaMD simulations can be reweighted to obtain the actual free energy profiles. One of the notable advantages of this method is no need to predefine reaction coordinates, and thus it has been gradually validated and used in various biomolecular systems.[31]

In order to facilitate the simulations of ligand binding and dissociation processes, the Miao group recently proposed a new algorithm called ligand GaMD (LiGaMD).[32] Compared to the GaMD algorithm, the authors introduce two boost potentials (dual-boost algorithm) to study ligand binding thermodynamics and kinetics. One boost potential is selectively added to the ligand nonbonded interaction potential energy, while the other is applied to the remaining potential energy surface of the whole system. By testing this algorithm to model systems including β-cyclodextrin and trypsin enzymes, they calculated the corresponding binding free energies and kinetic rate constants, which match well with the experimental data. Moreover, in their hundreds-of-nanosecond LiGaMD simulations, they observed repetitive dissociation and binding processes of ligand or guest molecules, which is very difficult in classical MD simulations. In other words, LiGaMD provides a promising approach for quantifying protein–ligand interactions.

3.3.2 Metadynamics

Metadynamics (MetaD)[33] is a very powerful enhanced sampling technique in MD simulations, which can reconstruct the free energy surface as a function of several selected degrees of freedom, which are often called collective variables (CVs). In MetaD simulations, sampling is enhanced by an external history-dependent bias potential, a function of selected CVs. The potential can be decomposed into the sum of Gaussian terms of each CV to discourage the system evolution from revisiting configurations that have already been sampled. Hence, MetaD simulations facilitate the sampling of rare events by pushing the system away from the local free energy minima. In other words, it may promote the identification of new reaction pathways because the simulation system tends to cross the free energy saddle point. However, MetaD simulations have a significant drawback: the choice of proper CVs, which is essential for obtaining a precise free energy landscape quickly. With the rapid development in the past ~20 years (e.g., well-tempered MetaD, a smoothly converging form of the original MetaD algorithm), the current algorithm in MetaD simulations has been an efficient, flexible, and accurate method that is widely used in many research fields including protein–drug interactions.

MetaD simulations can be directly used to perform the docking of small molecules onto fully flexible receptors.[34] In the simulations, ligands enter, overfill, and leave the binding pocket of the receptors. These processes can be adequately sampled to reveal the docking kinetics and binding free energy surfaces, which quantitatively describe the free energy changes during the interactions between some molecules and receptors. This is essential for discovering the binding path of the small molecule from the water solution to the binding pocket of the receptor and from one pocket to another pocket. Besides, by means of the benchmark studies, the authors validated that the free energy surface obtained by MetaD simulations agree well with that obtained by running much longer two-dimensional umbrella sampling simulations. During the interactions between the target protein and the ligand, the effective evaluation of the ligand binding stability will be helpful for drug discovery. For this purpose, an automated binding pose MetaD protocol was proposed.[35] In this method, the ligand is forced to move around its binding pose, and the strength of mobility under the added biased potential is used as an indicator of ligand binding stability.

As mentioned above, MetaD simulations have been successfully applied to soluble proteins. In order to better study transmembrane proteins, funnel MetaD simulations were developed.[36] In funnel MetaD simulations, MetaD-biased potential with funnel-shaped restraint potential applied to the target protein. The MetaD-biased potential still promotes the system evolution to the conformational space that is not sampled. The restraint potential combines a series of cone restraints and a cylindrical restraint to form a funnel, which includes the ligand binding site. When the ligand explores regions inside the funnel area, the system does not feel the funnel potential. If the ligand reaches the edge of the funnel, a repulsive biased

potential is applied to the system to prevent it from visiting regions outside the funnel. This is especially useful for transmembrane proteins, reducing a significant amount of computation. For example, Saleh et al.[37] applied the funnel MetaD simulations to reveal the binding affinity and the transition path of 12 agonists or antagonists in five G-protein-coupled receptors (GPCRs).

3.3.3 Markov state models

Markov state models (MSMs) are a powerful analytical method for dynamic systems such as MD simulations,[38,39] which focus on slow degrees of freedom that dominate the system dynamics, rather than fast degrees of freedom that are less relevant. Briefly, MSMs divide the whole system configuration space into n states, and the MSM itself is an $n \times n$ transition probability matrix. Element in each row of this matrix represents the probability of transitioning from the row-indexed state to the column-indexed state, and the diagonal element represents the probability of staying in the same state. Based on the statistical mechanics, free energies can be calculated using state populations obtained in this matrix, while kinetic information and transition pathways between any two states can be quantified from the conditional pairwise transition probabilities within this matrix. The state populations and the transition probabilities are calculated directly from MD simulations. The MD simulations do not require a very long time because in MSMs, the system is memoryless, and separate trajectories can be integrated together when they occupy common states. However, several short MD simulations are necessary for reliable statistical data for the transition probability matrix. The starting system configurations of these short MD simulations are derived from optimized paths between the initial and final states.

Due to the powerful statistics, MSMs have been widely used to characterize the drug binding kinetics as well as the conformational dynamics of the target proteins. For example, in order to reveal the detailed interactions between serine protease trypsin and small molecule benzamidine, Plattner et al.[40] performed 543 MD simulations with the total simulation time of 149.1 ms. Through MSM analysis over these trajectories, they identified seven metastable conformations with different binding pocket structures and binding/dissociation rates. Linker et al.[41] developed an automated protocol based on MD simulations and MSM analysis to predict the binding sites and modes of fragment-like small molecules. This protocol only uses the three-dimensional structure of the target protein and ligand's chemical structure as the input, which dramatically improves the efficiency of fragment-based drug discovery. Moreover, several software platforms, including MSMBuilder,[42] PyEMMA,[43] and HTMD,[44] have been developed for the easy use of MSM analysis. Besides, MSMs can also be coupled with other methods such as MetaD simulations to provide valuable insights into energetic and dynamical details in protein–drug interactions.[45]

3.3.4 Replica-exchange molecular dynamics

The most direct and straightforward strategy to improve the conformational sampling of the simulation systems is to run multiple replicas simultaneously. However, there are also several disadvantages of this strategy. On the one hand, it will significantly increase the need for computational resources, while on the other hand, multiple independent replicas still need proper coupling schemes to cross the high free energy barriers to obtain reasonable sampling and free energy landscape. The replica-exchange molecular dynamics (REMD) approach is one of these coupling methods and has proven useful in studying protein-drug interactions[46] . REMD simulation of a system starts from the room or physiological temperature, and the subsequent replicas are run at gradually increased temperatures. System coordinates are periodically exchanged between neighboring replicas. A Metropolis acceptance probability, which is a function of their potential energies and temperatures, determines whether the exchange is accepted or not. If the potential energy of the high temperature replica is lower than that of the lower temperature system, the exchange is accepted. Otherwise, the exchange will be refused. Generally, REMD simulation significantly enhances the sampling of the free energy landscape. However, it also modifies the actual dynamics by nonphysical exchanges, thus preventing the direct extraction of the system kinetics. In order to overcome this shortcoming, Stelzl and Hummer mapped the "time-continuous" trajectories from REMD simulations onto a set of discrete states using the transition-based assignment method.[46] Then, they use a maximum likelihood procedure to estimate kinetic rate coefficients based on the statistics of transition events in discontinuous REMD trajectories.

As a powerful parallelization technique, the distributed replica was applied to REMD simulations to boost computing efficiency.[47] Free energy perturbation (FEP) is a useful method to calculate absolute solvation free energies and binding free energies. In order to achieve robust and effective computations, the Roux group combined FEP and distributed REMD simulations.[47] The combined approach significantly improved the sampling efficiency and convergence of free energy computations and was a powerful tool in evaluating the absolute binding free energy, as proved in the camphor-P450 complex system. Kokubo et al. integrated the umbrella sampling with REMD simulations (also known as REUS),[48] which can robustly predict protein–ligand binding structures. Chen et al. integrated REMD and accelerated MD simulations to probe

the effects of disulfide bonds in BACE1 protein on binding three different inhibitors.[49] The obtained free energy landscapes and principal component analysis (PCA) revealed that the breaking of disulfide bonds induced the conformational changes of BACE1 and thus disordered binding poses of these three inhibitors on BACE1.

3.4 MM/PBSA and MM/GBSA binding free energy calculation

Binding free energy is an important thermodynamic parameter for evaluating the binding of drug molecules to the target protein. In principle, the binding free energy of a protein-drug system can be obtained using the formula $\Delta G = -kB\ Tln\Omega$ from MD simulations. *kB* is the Boltzmann's constant, *T* is the temperature of the system, and Ω is the partition function of the system. However, very long classical MD simulation is necessary to achieve the convergence of the binding free energy calculations. If we turn to free energy calculation methods based on biased MD simulations such as umbrellas sampling, free energy perturbation, and thermodynamic integration, the huge demand for computing resources is essential. The accuracy makes these methods useful in the stages of lead optimization, but the computing efficiency limits their applications in large-scale virtual screenings. If we solely rely on molecular docking calculations to evaluate binding free energy, the accuracy may be a problem. Hence, it will be essential to achieve a balance between the accuracy and the efficiency while calculating the binding free energy for high-throughput virtual screening process. For this purpose, end-point free energy calculation methods, which is based on samplings of the final states of a system and thus much less expensive than the classical free energy calculation approaches but much more accurate than most docking calculations, are widely used in structure-based drug design.

One of the most widely used end-point free energy methods is molecular mechanics Poisson–Boltzmann surface area (MM/PBSA) and generalized Born surface area (MM/GBSA).[50] In the MM/PBSA or MM/GBSA approach, the binding free energy for a ligand–protein system can be decomposed into contributions of several different interactions as follows.

$$\Delta G = \Delta H - T\Delta S \tag{1}$$

$$= \Delta E_{MM} + \Delta G_{sol} - T\Delta S \tag{2}$$

$$= \left(\Delta E_{int} + \Delta E_{ele} + \Delta E_{vdW}\right) + \left(\Delta G_{PB/GB} + \gamma \cdot \mathrm{SASA} + b\right) - T\Delta S \tag{3}$$

where ΔE_{int}, ΔE_{ele}, and ΔE_{vdW} represent the changes of internal energies (bond, angle, and dihedral energies), electrostatic energies, and the van der Waals energies upon binding the ligand. The change of solvation energy between the solute and the solvent upon ligand binding includes two parts: the electrostatic (polar) solvation energy $\Delta GPB/GB$ and nonpolar contribution $\gamma \cdot \mathrm{SASA}$. The change of the conformational entropy $-T\Delta S$ can be obtained using normal-mode analysis over a series of system conformations from MD simulations, which is relatively computationally expensive and can be neglected only if the relative binding free energies of two similar molecules are required. Generally, in order to achieve better convergence on the calculations of binding free energy, the PB calculations using a finer grid mesh are significantly time-consuming. As a simple approximation of the PB method, the GB method needs much less computing resources. When performing MM/PB(GB)SA binding free energy calculations, it is worth paying attention to two points. For the first point, MD simulations of the protein–ligand complex need to be performed with an explicit solvent model, which can ensure much more accurate results. For the second point, proper choice of system conformations from MD simulations will be essential to obtain the correct conformational entropy change ($-T\Delta S$).

The straightforward way to perform MM/PB(GB)SA binding free energy calculations is to use the tool *g_mmpbsa,*[51] which is part of the Open Source Drug Discovery (OSDD) consortium and very compatible with the MD software GROMACS. This method has been widely used in high-throughput virtual screening and the analysis of protein–drug interactions. For example, with the SARS-CoV-2 main protease as the target, Wang first performed molecular docking screening of approved drugs and drug candidates in clinical trials.[52] Then, he performed MD simulations and MM/PBSA calculations on the top hits to quickly identify repurposed drugs to combat SARS-CoV-2 based on kinetic information and binding free energies of these candidate drugs. Laurini et al.[53] applied MM/PBSA to quantitatively evaluate the role of each amino acid at the SARS-CoV-2 spike protein/angiotensin-converting enzyme 2 (ACE2) binding interface. The calculated binding free energy data yielded around 92% agreement, which in turn validated the accuracy of MM/PBSA binding free energy calculations. Although the accuracy of the MM/PBSA method to predict the relative binding free energy is acceptable, its ability to predict the absolute binding free energy needs to be improved. The overestimation of absolute binding free energy mainly originates from the calculation of the entropic term. Hence, Huang et al. evaluated the protein–ligand binding affinity by combining MM/PBSA calculations with the interaction entropy method.[54] The latter is responsible for the calculation of the entropy change. The agreement between the computational results with the experimental results proves that this method can effectively improve the ability of the MM/PBSA method in evaluating the absolute binding free energy.

3.5 Alchemical binding free energy calculation

The most widely used alchemical binding free energy calculation methods[55,56] include Thermodynamic Integration (TI), Bennett Acceptance Ratio (BAR), and Free Energy Perturbation (FEP). These methods are more accurate but more time-consuming. Here, we mainly focus on the FEP, which focuses on the change in free energy between two thermodynamic states. Generally, there are two kinds of FEP methods: the absolute FEP (AFEP) and the relative FEP (RFEP). The former calculates the absolute binding free energy of a solvated ligand into the target protein, while the latter calculates the relative binding free energy between two ligands and the target protein. Alchemical binding free energy calculations usually employ unphysical (alchemical) intermediates to estimate the binding free energy of the protein–ligand complex. As shown in Fig. 2, to calculate the difference in binding free energies ($\Delta G_1 - \Delta G_2$) between the target protein and ligand 1/2, two unphysical transformations (atom types were changed directly.) are introduced to form the thermodynamic cycle. Considering that the free energy difference is only determined by the initial and final states rather than the pathway, we can obtain $\Delta G_1 + \Delta G_4 = \Delta G_3 + \Delta G_2$ (two pathways from the state $P + L_1$ to the state PL_2, Fig. 2). That is $\Delta G_1 - \Delta G_2 = \Delta G_3 - \Delta G_4$. In other words, the calculation of $\Delta G_1 - \Delta G_2$ can be converted to the calculation of $\Delta G_3 - \Delta G_4$. ΔG_1 and ΔG_2 are absolute binding free energies, which are computationally expensive. On the contrary, ΔG_3 and ΔG_4 are relatively binding free energies. However, the FEP calculations, only when ligand 1 and ligand 2 are very similar, the calculations of ΔG_3 and ΔG_4 are easy to converge. Hence, in the actual calculations, if ligand 1 and ligand 2 have clear differences, a series of intermediate virtual ligands have to be introduced to achieve the quick convergence of free energy calculations.

3.6 The integration of molecular docking and molecular dynamics simulations

As mentioned above, MD simulations play key roles in many aspects of structure-based drug discovery,[57,58] including generating the conformational ensemble of the target protein, identifying drug binding sites on the target protein, and deriving observables such as the association/dissociation constants and ligand binding free energies, which can be directly measured in the corresponding experiments. On the other hand, molecular docking has become one of the most frequently used methods in high-throughput structure-based drug discovery. The most significant advantage of molecular docking is its ability to quickly predict the protein–ligand binding poses with rough bind affinity values based on its efficient conformational sampling algorithms and scoring functions. Proper integration of these two methods will be essential for achieving a robust and relatively precise protocol for high-throughput computational virtual screening. For ligand identification, high-throughput molecular docking of small molecule libraries may serve as the first round of screening. Then, classical force fields with implicit solvent models and end-point binding free energy calculations (e.g., MM/PBSA or MM/GBSA) can be used to perform further screening on the top hits from molecular docking. For the ligand optimization, long timescale classical MD simulations or biased MD simulations with enhanced sampling methods can be further used to obtain more accurate binding free energies and kinetic information. Besides, MD simulations can be used to validate and optimize the flexibility of both the target protein and docking molecules in flexible molecular docking.[59,60]

The integration of molecular docking and MD simulations in drug discovery is widely used. For example, using a directory of a useful decoy-enhanced (DUD-E) dataset, Guterres and Im performed high-throughput classical MD simulations of protein–ligand complex systems using the top output from Autodock Vina to select active ligands.[61] The ligand binding stability was evaluated using the ligand root-mean-square deviations (RMSD) with the alignment of the target protein. The results demonstrate that even 10 ns all-atom MD simulations can effectively improve the molecular docking results over 56 protein targets and 560 ligands. Besides, for some target proteins, they may significantly change their conformation upon small molecule binding. During the part of molecular docking, the induced fit docking approach will be necessary. Although this approach has been used in many systems, its accuracy is still controversial. In order to improve this, Clark et al. integrated the induced fit docking with MetaD simulations for predicting protein–ligand binding at an affordable computational cost.[62] By testing 42 different protein–ligand systems, the results clearly indicated that this integrated method

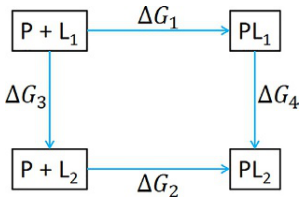

FIG. 2 The thermodynamic cycle. P (target protein), L1 (ligand 1), L2 (ligand 2), PL1 (protein–ligand 1 complex), PL2 (protein–ligand 2 complex). (No permission required.)

could effectively improve the accuracy of the original induced-fit docking approach. Also, global docking and accelerated MD simulations are used in combination to enhance the prediction of the binding between the target protein and peptide small molecules.[63] These examples clearly show the absolute advantages of integrating molecular docking calculation and MD simulations for predicting protein–ligand interactions.

4 Modeling drug-membrane interactions

Except for some drug binding sites (e.g., extracellular domains of membrane proteins), the candidate drug small molecules have to first cross the cell membrane and then bind to the target proteins. Hence, a proper evaluation of the membrane permeability of the obtained candidate small molecules will be essential. In principle, the entrance, translocation, and exit of the small molecules can be directly simulated by classical MD simulations (Fig. 3). However, with CHARMM all-atom force field, it needs as long as μs-scale MD simulations to quantify the membrane permeation process of very simple gas permeants,[64] which is supposed to have high diffusion coefficients. In order to alleviate this situation, coarse-grained models may provide a powerful alternative way to study the membrane permeability of small molecules. However, there are still uncertainties in the minimal time scale of unbiased MD simulations required for the quantification of small molecules' membrane permeability[65] because small molecules outside the lipid membrane are free to diffuse and the time to start the translocation of the lipid membrane is very random. Hence, the combination of the classical MD simulations and biased one-dimensional free energy calculations will be necessary for the purpose. For example, Hofmann et al.[66] performed classical MD simulations with Martini coarse-grained model and umbrella sampling simulations to study the detailed interactions between 105 coarse-grained compounds and 6 lipid membranes. That is 630 drug-membrane combinations. The transfer free energy $G(z)$ can be obtained from umbrella sampling simulations using the distance(z) between the small molecule and the center of the lipid membrane along the membrane normal as the reaction coordinate. The local diffusion coefficient $D(z)$ can be calculated through the autocorrelation function of the random force imposed on the small molecule. Then, the permeability coefficient P of the small molecule can be estimated by one-dimensional Smoluchowski equation $P^{-1} = \int \frac{e^{G(z)/k_B T}}{D(z)} dz$. The detailed database of small molecule structures, their dynamics in the lipid membrane, and membrane permeability will provide useful insights into the structure–property relationships governing drug–membrane permeability.

On the other hand, machine learning methods are also powerful tools to predict the membrane permeability of small molecules, which previously mainly use experimental data for the training set. Recently, Bennett et al.[67] evaluated the enhancement of atomistic MD simulations with different machine learning methods to predict the drug permeability from the water phase to the cyclohexane phase. They firstly obtained transfer-free energy profiles of 15,000 small molecules, which then served as the training set for different machine learning models such as the spatial graph neural network model (SG-CNN), the 3D-convolutional neural network (3D-CNN), and shallow learning using extended circular fingerprints (ECFP). The results indicated that these models yielded different prediction accuracy for transfer-free energies, with SG-CNN achieving the highest accuracy (SG-CNN ≈ 3D-CNN ≫ ECFP). Through the combination of MD simulations and machine learning methods, the authors demonstrated the tight correlation between properties of small molecules and their transfer-free energies. This is very practical for discovering candidate drugs targeting the intracellular targets or the transmembrane domains of membrane proteins.

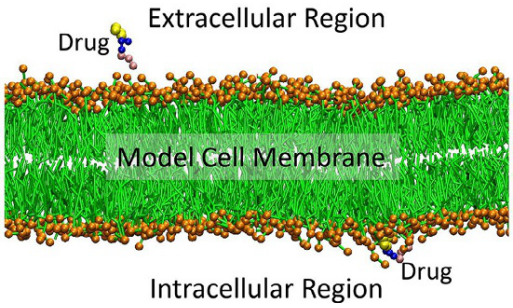

FIG. 3 Drug–membrane interactions. The binding of the drug molecule to cell membrane is shown. *(No permission required.)*

5 Conclusion and future perspectives

Human health still faces the grand challenges of many diseases. Effective and less expensive drugs to combat these diseases are in high demand. Hence, drug discovery plays a vital role in promoting better health and better life. However, drug discovery and development often take over 10 years, which consists of fundamental preclinical research and several phases of clinical trial research. Various computer-aided drug screening and design methods can effectively reduce the needed time to identify potential candidate drugs during the process of fundamental preclinical research. In this chapter, we reviewed the possible applications of MD simulations in different aspects of drug discovery. In the preparation of high-throughput virtual screening, MD simulations can generate the conformational ensemble of the target protein, identify the drug binding sites on the target protein, and optimize the structure of small molecules. In the process of high-throughput virtual screening, docking calculations can be performed using MD force fields rather than the classical scoring functions. End-point binding free energy calculations such as MM/PBSA or MM/GBSA, which are based on MD simulations and less expensive, can be integrated with docking calculations to improve the accuracy of the virtual screening without significantly affecting its efficiency. Besides, coarse-grained MD simulations can be directly used for high-throughput screening. In the post-processing stage of high-throughput virtual screening, the top hits can be further reevaluated and re-ranked by either long MD simulations with enhanced sampling methods or various more precise binding free energy calculation approaches. MD-derived observables during the process enable the direct comparison with the molecular experiments. Finally, MD simulation provides a valuable tool to evaluate drug permeability, which is vital for drugs targeting the intracellular targets or the transmembrane domains of membrane proteins.

As mentioned above, MD simulations have shown broad applications in drug discovery. However, every method has its advantages and also disadvantages. The integration of different methods will help maximize their advantages and may minimize their disadvantages. For example, the integration of molecular docking and MD simulations can effectively improve the screening accuracy. The combination of different MD simulation methods can also promote more precise and comprehensive results on the kinetic information and binding free energies of the small molecules on the target protein. On the other hand, artificial intelligence[68,69] has attracted much attention in the past several years due to its potential in drug discovery. It has been proved that MD simulations can generate the training sets for optimizing the artificial intelligence models for drug discovery, and in turn, artificial intelligence can be used to improve the accuracy of MD force fields. In the future, the cross-integration of MD simulations with other calculation algorithms for drug discovery will be an important research direction.

References

1. Hollingsworth SA, Dror RO. Molecular dynamics simulation for all. *Neuron.* 2018;99(6):1129–1143. https://doi.org/10.1016/j.neuron.2018.08.011.
2. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. *Nature.* 1977;267(5612):585–590. https://doi.org/10.1038/267585a0.
3. Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics.* 2016;54(1):5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3.
4. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* 2004;32(Suppl 2):W526–W531. https://doi.org/10.1093/nar/gkh468%.
5. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER suite: protein structure and function prediction. *Nat Methods.* 2015;12(1):7–8. https://doi.org/10.1038/nmeth.3213.
6. Amaro RE, Baudry J, Chodera J, et al. Ensemble docking in drug discovery. *Biophys J.* 2018;114(10):2271–2278. https://doi.org/10.1016/j.bpj.2018.02.038.
7. Evangelista Falcon W, Ellingson SR, Smith JC, Baudry J. Ensemble docking in drug discovery: how many protein configurations from molecular dynamics simulations are needed to reproduce known ligand binding? *J Phys Chem B.* 2019;123(25):5189–5195. https://doi.org/10.1021/acs.jpcb.8b11491.
8. Gupta AK, Wang X, Pagba CV, et al. Multi-target, ensemble-based virtual screening yields novel allosteric KRAS inhibitors at high success rate. *Chem Biol Drug Des.* 2019;94(2):1441–1456. https://doi.org/10.1111/cbdd.13519.
9. Acharya A, Agarwal R, Baker MB, et al. Supercomputer-based ensemble docking drug discovery pipeline with application to Covid-19. *J Chem Inf Model.* 2020;60(12):5832–5852. https://doi.org/10.1021/acs.jcim.0c01010.
10. Parks JM, Smith JC. How to discover antiviral drugs quickly. *N Engl J Med.* 2020;382(23):2261–2264. https://doi.org/10.1056/NEJMcibr2007042.
11. Marrink SJ, Risselada HJ, Yefimov S, Tieleman DP, de Vries AH. The MARTINI force field: coarse grained model for biomolecular simulations. *J Phys Chem B.* 2007;111(27):7812–7824. https://doi.org/10.1021/jp071097f.
12. de Jong DH, Singh G, Bennett WFD, et al. Improved parameters for the Martini coarse-grained protein force field. *J Chem Theory Comput.* 2013;9(1):687–697. https://doi.org/10.1021/ct300646g.
13. Wassenaar TA, Pluhackova K, Böckmann RA, Marrink SJ, Tieleman DP. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J Chem Theory Comput.* 2014;10(2):676–690. https://doi.org/10.1021/ct400617g.

14. Schmidtke P, Le Guilloux V, Maupetit J, Tufféry P. fpocket: online tools for protein ensemble pocket detection and tracking. *Nucleic Acids Res.* 2010;38(Suppl 2):W582–W589. https://doi.org/10.1093/nar/gkq383%.
15. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput Biol.* 2009;5(12). https://doi.org/10.1371/journal.pcbi.1000585, e1000585.
16. Tan YS, Reeks J, Brown CJ, et al. Benzene probes in molecular dynamics simulations reveal novel binding sites for ligand design. *J Phys Chem Lett.* 2016;7(17):3452–3457. https://doi.org/10.1021/acs.jpclett.6b01525.
17. Zuzic L, Marzinek JK, Warwicker J, Bond PJ. A benzene-mapping approach for uncovering cryptic pockets in membrane-bound proteins. *J Chem Theory Comput.* 2020;16(9):5948–5959. https://doi.org/10.1021/acs.jctc.0c00370.
18. Prakash P, Hancock JF, Gorfe AA. Binding hotspots on K-ras: consensus ligand binding sites and other reactive regions from probe-based molecular dynamics analysis. *Proteins.* 2015;83(5):898–909. https://doi.org/10.1002/prot.24786.
19. Sayyed-Ahmad A, Gorfe AA. Mixed-probe simulation and probe-derived surface topography map analysis for ligand binding site identification. *J Chem Theory Comput.* 2017;13(4):1851–1861. https://doi.org/10.1021/acs.jctc.7b00130.
20. Graham SE, Leja N, Carlson HA. MixMD probeview: robust binding site prediction from cosolvent simulations. *J Chem Inf Model.* 2018;58(7):1426–1433. https://doi.org/10.1021/acs.jcim.8b00265.
21. Tan YS, Verma CS. Straightforward incorporation of multiple ligand types into molecular dynamics simulations for efficient binding site detection and characterization. *J Chem Theory Comput.* 2020;16(10):6633–6644. https://doi.org/10.1021/acs.jctc.0c00405.
22. Wu G, Robertson DH, Brooks III CL, Vieth M. Detailed analysis of grid-based molecular docking: a case study of CDOCKER—a CHARMm-based MD docking algorithm. *J Comput Chem.* 2003;24(13):1549–1562. https://doi.org/10.1002/jcc.10306.
23. Grosdidier A, Zoete V, Michielin O. EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins.* 2007;67(4):1010–1025. https://doi.org/10.1002/prot.21367.
24. Yin S, Biedermannova L, Vondrasek J, Dokholyan NV. MedusaScore: an accurate force field-based scoring function for virtual drug screening. *J Chem Inf Model.* 2008;48(8):1656–1662. https://doi.org/10.1021/ci8001167.
25. Pan AC, Xu H, Palpant T, Shaw DE. Quantitative characterization of the binding and unbinding of Millimolar drug fragments with molecular dynamics simulations. *J Chem Theory Comput.* 2017;13(7):3372–3377. https://doi.org/10.1021/acs.jctc.7b00172.
26. Wolf S, Lickert B, Bray S, Stock G. Multisecond ligand dissociation dynamics from atomistic simulations. *Nat Commun.* 2020;11(1):2918. https://doi.org/10.1038/s41467-020-16655-1.
27. Souza PCT, Alessandri R, Barnoud J, et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat Methods.* 2021;18(4):382–388. https://doi.org/10.1038/s41592-021-01098-3.
28. Souza PCT, Thallmair S, Conflitti P, et al. Protein–ligand binding with the coarse-grained Martini model. *Nat Commun.* 2020;11(1):3714. https://doi.org/10.1038/s41467-020-17437-5.
29. Souza PCT, Limongelli V, Wu S, Marrink SJ, Monticelli L. Perspectives on high-throughput ligand/protein docking with Martini MD simulations. *Front Mol Biosci.* 2021;8(199):657222. https://doi.org/10.3389/fmolb.2021.657222.
30. Miao Y, Feher VA, McCammon JA. Gaussian accelerated molecular dynamics: unconstrained enhanced sampling and free energy calculation. *J Chem Theory Comput.* 2015;11(8):3584–3595. https://doi.org/10.1021/acs.jctc.5b00436.
31. Wang J, Arantes PR, Bhattarai A, et al. Gaussian accelerated molecular dynamics: principles and applications. *WIREs Comput Mol Sci.* 2021. https://doi.org/10.1002/wcms.1521.
32. Miao Y, Bhattarai A, Wang J. Ligand Gaussian accelerated molecular dynamics (LiGaMD): characterization of ligand binding thermodynamics and kinetics. *J Chem Theory Comput.* 2020;16(9):5526–5547. https://doi.org/10.1021/acs.jctc.0c00395.
33. Barducci A, Bonomi M, Parrinello M. Metadynamics. *WIREs Comput Mol Sci.* 2011;1(5):826–843. https://doi.org/10.1002/wcms.31.
34. Gervasio FL, Laio A, Parrinello M. Flexible docking in solution using metadynamics. *J Am Chem Soc.* 2005;127(8):2600–2607. https://doi.org/10.1021/ja0445950.
35. Fusani L, Palmer DS, Somers DO, Wall ID. Exploring ligand stability in protein crystal structures using binding pose metadynamics. *J Chem Inf Model.* 2020;60(3):1528–1539. https://doi.org/10.1021/acs.jcim.9b00843.
36. Raniolo S, Limongelli V. Ligand binding free-energy calculations with funnel metadynamics. *Nat Protoc.* 2020;15(9):2837–2866. https://doi.org/10.1038/s41596-020-0342-4.
37. Saleh N, Ibrahim P, Saladino G, Gervasio FL, Clark T. An efficient metadynamics-based protocol to model the binding affinity and the transition state ensemble of G-protein-coupled receptor ligands. *J Chem Inf Model.* 2017;57(5):1210–1217. https://doi.org/10.1021/acs.jcim.6b00772.
38. Wang W, Cao S, Zhu L, Huang X. Constructing Markov state models to elucidate the functional conformational changes of complex biomolecules. *WIREs Comput Mol Sci.* 2018;8(1). https://doi.org/10.1002/wcms.1343, e1343.
39. Husic BE, Pande VS. Markov state models: from an art to a science. *J Am Chem Soc.* 2018;140(7):2386–2396. https://doi.org/10.1021/jacs.7b12191.
40. Plattner N, Noé F. Protein conformational plasticity and complex ligand-binding kinetics explored by atomistic simulations and Markov models. *Nat Commun.* 2015;6(1):7653. https://doi.org/10.1038/ncomms8653.
41. Linker SM, Magarkar A, Köfinger J, Hummer G, Seeliger D. Fragment binding pose predictions using unbiased simulations and Markov-state models. *J Chem Theory Comput.* 2019;15(9):4974–4981. https://doi.org/10.1021/acs.jctc.9b00069.
42. Harrigan MP, Sultan MM, Hernández CX, et al. MSMBuilder: statistical models for biomolecular dynamics. *Biophys J.* 2017;112(1):10–15. https://doi.org/10.1016/j.bpj.2016.10.042.
43. Scherer MK, Trendelkamp-Schroer B, Paul F, et al. PyEMMA 2: a software package for estimation, validation, and analysis of Markov models. *J Chem Theory Comput.* 2015;11(11):5525–5542. https://doi.org/10.1021/acs.jctc.5b00743.

44. Doerr S, Harvey MJ, Noé F, De Fabritiis G. HTMD: high-throughput molecular dynamics for molecular discovery. *J Chem Theory Comput.* 2016;12(4):1845–1852. https://doi.org/10.1021/acs.jctc.6b00049.
45. Bernetti M, Masetti M, Recanatini M, Amaro RE, Cavalli A. An integrated Markov state model and path metadynamics approach to characterize drug binding processes. *J Chem Theory Comput.* 2019;15(10):5689–5702. https://doi.org/10.1021/acs.jctc.9b00450.
46. Stelzl LS, Hummer G. Kinetics from replica exchange molecular dynamics simulations. *J Chem Theory Comput.* 2017;13(8):3927–3935. https://doi.org/10.1021/acs.jctc.7b00372.
47. Jiang W, Hodoscek M, Roux B. Computation of absolute hydration and binding free energy with free energy perturbation distributed replica-exchange molecular dynamics. *J Chem Theory Comput.* 2009;5(10):2583–2588. https://doi.org/10.1021/ct900223z.
48. Kokubo H, Tanaka T, Okamoto Y. Prediction of protein–ligand binding structures by replica-exchange umbrella sampling simulations: application to kinase systems. *J Chem Theory Comput.* 2013;9(10):4660–4671. https://doi.org/10.1021/ct4004383.
49. Chen J, Yin B, Wang W, Sun H. Effects of disulfide bonds on binding of inhibitors to β-amyloid cleaving enzyme 1 decoded by multiple replica accelerated molecular dynamics simulations. *ACS Chem Neurosci.* 2020;11(12):1811–1826. https://doi.org/10.1021/acschemneuro.0c00234.
50. Wang E, Sun H, Wang J, et al. End-point binding free energy calculation with MM/PBSA and MM/GBSA: strategies and applications in drug design. *Chem Rev.* 2019;119(16):9478–9508. https://doi.org/10.1021/acs.chemrev.9b00055.
51. Kumari R, Kumar R, Lynn A. g_mmpbsa—a GROMACS tool for high-throughput MM-PBSA calculations. *J Chem Inf Model.* 2014;54(7):1951–1962. https://doi.org/10.1021/ci500020m.
52. Wang J. Fast identification of possible drug treatment of coronavirus disease-19 (COVID-19) through computational drug repurposing study. *J Chem Inf Model.* 2020;60(6):3277–3286. https://doi.org/10.1021/acs.jcim.0c00179.
53. Laurini E, Marson D, Aulic S, Fermeglia A, Pricl S. Computational mutagenesis at the SARS-CoV-2 spike protein/angiotensin-converting enzyme 2 binding Interface: comparison with experimental evidence. *ACS Nano.* 2021. https://doi.org/10.1021/acsnano.0c10833.
54. Huang K, Luo S, Cong Y, Zhong S, Zhang JZH, Duan L. An accurate free energy estimator: based on MM/PBSA combined with interaction entropy for protein–ligand binding affinity. *Nanoscale.* 2020;12(19):10737–10750. https://doi.org/10.1039/C9NR10638C.
55. Lee T-S, Allen BK, Giese TJ, et al. Alchemical binding free energy calculations in AMBER20: advances and best practices for drug discovery. *J Chem Inf Model.* 2020;60(11):5595–5623. https://doi.org/10.1021/acs.jcim.0c00613.
56. Song LF, Merz KM. Evolution of alchemical free energy methods in drug discovery. *J Chem Inf Model.* 2020;60(11):5308–5318. https://doi.org/10.1021/acs.jcim.0c00547.
57. De Vivo M, Masetti M, Bottegoni G, Cavalli A. Role of molecular dynamics and related methods in drug discovery. *J Med Chem.* 2016;59(9):4035–4061. https://doi.org/10.1021/acs.jmedchem.5b01684.
58. Shukla R, Tripathi T. In: Singh S, ed. *Molecular Dynamics Simulation in Drug Discovery: Opportunities and Challenges.* Singapore: Springer; 2021:295–316. https://doi.org/10.1007/978-981-15-8936-2_12.
59. Śledź P, Caflisch A. Protein structure-based drug design: from docking to molecular dynamics. *Curr Opin Struct Biol.* 2018;48:93–102. https://doi.org/10.1016/j.sbi.2017.10.010.
60. Shukla R, Tripathi T. Molecular dynamics simulation of protein and protein-ligand complexes. In: *Computer-Aided Drug Design.* Singapore: Springer; 2020:133–161. https://doi.org/10.1007/978-981-15-6815-2_7.
61. Guterres H, Im W. Improving protein-ligand docking results with high-throughput molecular dynamics simulations. *J Chem Inf Model.* 2020;60(4):2189–2198. https://doi.org/10.1021/acs.jcim.0c00057.
62. Clark AJ, Tiwary P, Borrelli K, et al. Prediction of protein–ligand binding poses via a combination of induced fit docking and metadynamics simulations. *J Chem Theory Comput.* 2016;12(6):2990–2998. https://doi.org/10.1021/acs.jctc.6b00201.
63. Wang J, Alekseenko A, Kozakov D, Miao Y. Improved modeling of peptide-protein binding through global docking and accelerated molecular dynamics simulations. *Front Mol Biosci.* 2019;6:112. https://doi.org/10.3389/fmolb.2019.00112.
64. Venable RM, Krämer A, Pastor RW. Molecular dynamics simulations of membrane permeability. *Chem Rev.* 2019;119(9):5954–5997. https://doi.org/10.1021/acs.chemrev.8b00486.
65. Krämer A, Ghysels A, Wang E, et al. Membrane permeability of small molecules from unbiased molecular dynamics simulations. *J Chem Phys.* 2020;153(12):124107. https://doi.org/10.1063/5.0013429.
66. Hoffmann C, Centi A, Menichetti R, Bereau T. Molecular dynamics trajectories for 630 coarse-grained drug-membrane permeations. *Sci Data.* 2020;7(1):51. https://doi.org/10.1038/s41597-020-0391-0.
67. Bennett WFD, He S, Bilodeau CL, et al. Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning. *J Chem Inf Model.* 2020;60(11):5375–5381. https://doi.org/10.1021/acs.jcim.0c00318.
68. Díaz Ó, Dalton JAR, Giraldo J. Artificial intelligence: a novel approach for drug discovery. *Trends Pharmacol Sci.* 2019;40(8):550–551. https://doi.org/10.1016/j.tips.2019.06.005.
69. Marchetti F, Moroni E, Pandini A, Colombo G. Machine learning prediction of allosteric drug activity from molecular dynamics. *J Phys Chem Lett.* 2021;3724–3732. https://doi.org/10.1021/acs.jpclett.1c00045.

This page intentionally left blank

Chapter 28

Envisaging the conformational space of proteins by coupling machine learning and molecular dynamics

Murali Aarthy and Sanjeev Kumar Singh

Computer Aided Drug Design and Molecular Modeling Lab, Department of Bioinformatics, Alagappa University, Karaikudi, Tamil Nadu, India

1 Introduction

Proteins, the major biological subunit of the human body, exist in multiple conformations with rigid structural formation. This formation is considered the assembly of dynamic conformations distributed across the free energy landscape based on the Boltzmann –weighted probability of occurrence.[1,2] About 60 years ago, Feynman stated that life is not only about the organization of atoms but also about the *jiggling and wiggling of atoms*.[3] The significant prototype of structural biology states that the three-dimensional (3D) fold of the protein is determined by the sequence, whereas the detonation of the structural data in the earlier decades was dramatically lingered with the classical view endorsing Feynman's prediction.[4] Proteins are dynamic objects with arrangements ranging from the subrotameric side-chain movement to the domain movement, which associates inherently to its own function. There are various computational techniques to study and understand the dynamics of the proteins, including the molecular dynamics (MD) simulation and the normal mode analysis.[5] The conformational space differs for each and every type of protein in terms of its side-chain and backbone residues. The research group of McGregor analyzed the relationship between the side-chain and the secondary structure of the globular proteins and reported the accurate distributions of the dihedral angles of the side-chains obtained from the proteins determined with the high resolution.[6] The most important mechanism involved in protein regulation and function is conformational change. The stability of proteins and the conformational sub-ensembles is highly essential to understand the regulation of protein dynamics. The dynamic nature of the protein owes difficulty in the observation of the conformational changes at the molecular level in wet-lab experiments, whereas the computational methods determine the timescale convoluted.[7–10]

The structural hierarchy of proteins is differentiated into the primary, secondary, tertiary, and quaternary structures. Among these, the tertiary structure strongly determines the function of the protein. The irregular folding of the protein leads to the loss of function and results in several diseases.[2] The role of protein folding and the determination of structure prediction is critical for drug design strategies and the biotechnology sectors.[11] The protein conformation is defined as the arrangement of the amino acids or the essential atoms in the 3D spatial structure. The energy functions acquired have the energy potential for different types of contacts formed during the conformation.[12] The determination of a large number of protein structures in the past few decades showed incredible progress in the structure-resolving methods to analyze protein with high flexibility and complexity.[4] The protein domain is the unit of protein structure formed during folding and possesses critical roles in the function.[13] In general, the mechanical degrees of freedom present in the molecular system with the effective conformational degrees as measured by the root-mean-squared Cartesian distances among all the conformations represent the vital property for assessment. Each protein attains a unique conformation, which is treated as the specific point in the high dimensional Euclidean space.[14] The protein possesses various conformational in different temperatures, pH environments, and ligand binding.[15] Further, for every protein, the system to identify and study the stability is different. The systems for membrane proteins, protein–ligand complex, the analysis of free energy, the identification of potential mean forces are different. The analysis of the conformational spaces acquired with the help of simulation is provided in Fig. 1, which depicts the insights on the process.

Advances in Protein Molecular and Structural Biology Methods. https://doi.org/10.1016/B978-0-323-90264-9.00028-3