Send Orders for Reprints to reprints@benthamscience.ae

Current Topics in Medicinal Chemistry, 2017, *17*, 1-9

REVIEW ARTICLE

Drug Discovery and Molecular Dynamics: Methods, Applications and Perspective Beyond the Second Timescale

Gerard Martínez-Rosell[a], Toni Giorgino[b] and Matt J. Harvey[c] and Gianni de Fabritiis[*,d]

[a]*Computational Biophysics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader 88, 08003, Barcelona, Spain;* [b]*Institute of Neurosciences, National Research Council of Italy (IN-CNR), Padua, Italy;* [c]*Acellera, Barcelona Biomedical Research Park (PRBB), C/Doctor Aiguader 88, 08003 Barcelona, Spain;* [d]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, Barcelona 08010, Spain*

ARTICLE HISTORY

Received: September 05, 2016
Revised: November 14, 2016
Accepted: November 15, 2016

DOI:
10.2174/1568026617666170414142549

Abstract: Bio-molecular dynamics (MD) simulations based on graphical processing units (GPUs) were first released to the public in the early 2009 with the code ACEMD. Almost 8 years after, applications now encompass a broad range of molecular studies, while throughput improvements have opened the way to millisecond sampling timescales. Based on an extrapolation of the amount of sampling in published literature, the second timescale will be reached by the year 2022, and therefore we predict that molecular dynamics is going to become one of the main tools in drug discovery in both academia and industry. Here, we review successful applications in the drug discovery domain developed over these recent years of GPU-based MD. We also retrospectively analyse limitations that have been overcome over the years and give a perspective on challenges that remain to be addressed.

Keywords: Molecular dynamics, simulation, drug discovery, ACEMD, perspective, review.

1. INTRODUCTION

Simulation of molecular systems can be performed at a variety of different levels of theory and resolutions. This review takes into consideration all-atom classical molecular dynamics simulations (MD). Classical MD represents a pragmatic compromise between physical fidelity of model and computational efficiency, and is able to reproduce experimentally-observable properties of condensed-phase, biomolecular systems, with its main limitation being the inability to represent chemical reactions. For a thorough introduction to the essentials of classical MD we refer the reader to [1].

Despite the low algorithmic complexity of MD in comparison to quantum chemistry methods, the computational cost is such that high performance computing (HPC) systems have been required to perform simulations of sufficient length to approach biologically relevant timescales [2]. The size and specialisation of the parallel HPC systems required has made MD sampling of even small biologically-interesting systems very costly in terms of Euro per simulated time. Consequently, much technical effort has been invested in developing specialized hardware, such as Anton supercomputer [3], and simulation software optimised to maximise performance on these machines.

*Address correspondence to this author at the Department of Computational Biophysics, University Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader 88, 08003, Barcelona, Spain;
Tel: +34 933160537; E-mail: gianni.defabritiis@upf.edu

In the latter half of last decade, developments in the computer graphics technology sector resulted in the introduction to the HPC field of a new class of processor with radically different characteristics to conventional CPUs. The characteristics of these processors, termed GPUs (graphics processing units), make them highly amenable to certain classes of scientific computation, in particular those such as MD which contain a high degree of intrinsic parallelism. Although there are many established MD codes these all embody design decisions that are optimal for classical parallel HPC systems, which left them requiring extensive re-engineering to exploit GPUs. The most efficient GPU MD codes are those such as ACEMD [4] and recent versions of PMEMD [5], OpenMM [6] and Desmond [7], all of which have been designed and optimised specifically for the architecture of GPUs.

These processors have been proven effective for MD that it is possible to achieve simulation rates on a single GPU that would have previously required large HPC resource. Although GPUs have not yielded an improvement in attainable single simulation performance, the fact that they can deliver HPC-class performance on commodity hardware has yielded a significant (>10x) reduction in the Euro/simulated time cost (Fig. **1A**). Consequently, it has been possible to run tens of thousands of simulations routinely (*e.g.* using distributed computing projects like GPUGRID [8]) and methods to reconstruct the information provided by these many short simulations have also improved, arguably being now superior to a single long simulation data of equivalent sampling time [8, 9].

1568-0266/17 $58.00+.00 © 2017 Bentham Science Publishers

Within this review we look at how the advent of GPUs has changed the research done in the last years. To do so we focus on work performed on GPU hardware with ACEMD, because it offers the longest historical data having been introduced in 2008. Indeed, GPU performance as measured in floating-point operations per second (GFLOPS) **(**Fig. **1B)** has doubled approximately every 3 years, with ACEMD performance increasing by approximately 2.1x in that interval, reflecting general GPU architectural maturation and code improvements.

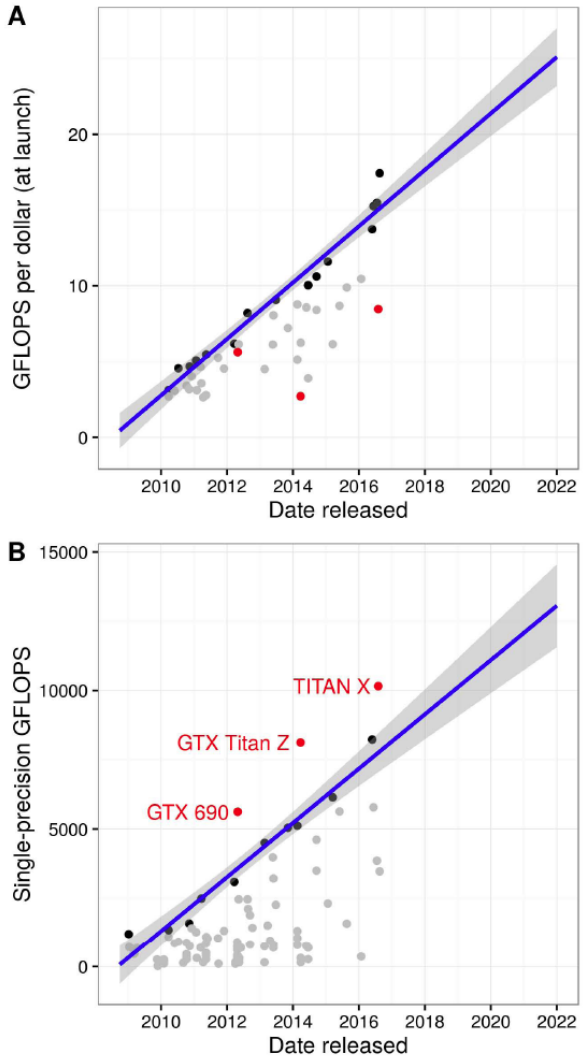

Fig. (1). Total computing throughput (**A**) and performance per dollar (**B**) versus date of release of NVIDIA graphic cards with programmable computing capabilities (CUDA). Unit of performance is 10^9 single-precision floating point operations per second, or GFLOPS. Linear fits and confidence intervals are indicated for the best performing models (black points). Red points are "flagship" models, *i.e.* early next-generation releases, not included in the fit due to relatively unfavorable cost-performance ratios.

2. APPLICATIONS OF MD IN DRUG DISCOVERY

In-silico techniques, among them molecular dynamics (MD), can be useful both to guide and rationalize each of the steps of the drug discovery pipeline [11]**.** Specifically, MD can help to understand the dynamics of the target, such as the molecular mechanisms underlying a pathology, and can be useful to support and drive the results from hit discovery, hit-to-lead and lead optimization by assessing receptor-ligand binding poses, stability and dynamics of the binding poses, receptor-ligand binding affinities and even kinetics (k_{on} and k_{off})[12]. However, MD has still not been widely included in the mainstream drug discovery pipeline. Major advances in hardware, software and force-field accuracy are steadily introducing MD as an effective technique to assess the ability of a ligand to bind a target as well as characterizing the dynamics of the target.

A literature review of ACEMD drug discovery applications over the last 8 years reveals that the main use cases of MD in the field of drug discovery principally include, (a) the conformational characterization of targets and understanding of key "druggable" molecular mechanisms, (b) the generation of protein conformations for ensemble docking, (c) the postprocessing of docked or crystallographic poses, (d) the reproduction of thermodynamic and kinetic properties by means of full ligand-protein binding events and (e) the calculation of free energies using biased MD methods.

The impact of the technological development, both in terms of computational cost reduction and methods improvement, on the kind of studies MD has been able to address until now is enormous. While is still true that the progress of the field is bound to the previous accumulated scientific knowledge and the ability to ask the right questions, MD is also particularly dependent on the technological advancement, which limits the biological processes timescales we are able to sample. In this sense, the computational cost reduction has been remarkable over the last years and one is able to extrapolate single GPU simulation rates in the order of the μs/day by 2022 for systems of intermediate size (ca 50k atoms including solvent). When further coupled to a computing infrastructure that delivers access to large numbers of GPUs, such as GPU-based HPC machinery, or a distributed computing network like GPUGRID [8], we extrapolate that by 2022, MD-based studies will employ aggregate sampling on the second timescale (Fig. **2**).

2.1. Early Protein-ion Binding Studies

The first studies on ligand binding using GPUs involved the simplest form of ligands: ions. Early in 2010, an all-atom microsecond-long unbiased simulation was able to reconstruct full binding events of sodium ions to the GPCR D_2 receptor embedded in a membrane [13]. The binding was described as happening in a two-step manner, the first step involving four extracellular negatively charged residues that produced a favorable environment for sodium, and then a second step where the ions visited three internal binding locations. Further 25 nanosecond-long metadynamics simulations were produced to assess the binding free energy profile. The location of the ion and position of the side chain around it was identical to a crystal structure published several years later [14].

Computationally cheaper methods like steered molecular dynamics (SMD) were also used early in 2011 to assess the binding free energy profile of potassium ions permeation through Gramicidin A [15]. Steered Molecular Dynamics is a biased method that consists of applying an external force to

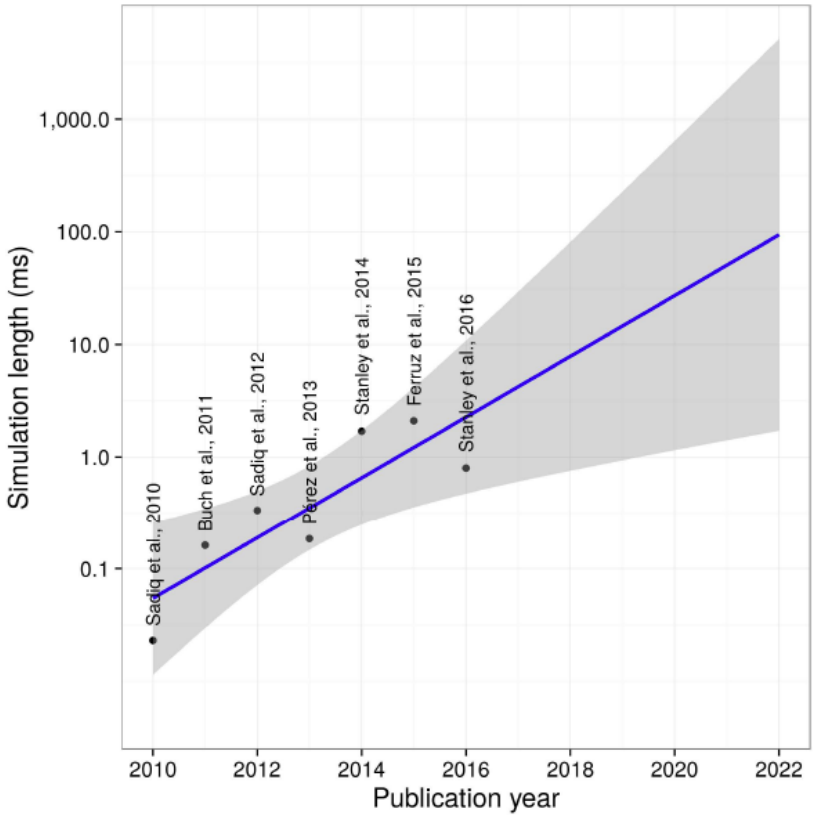

Fig. (2). Approximate total aggregate sampled time for high-throughput all-atom molecular dynamics studies published in years 2010-2016 (log scale). An exponential function (solid line with 95% confidence interval) was fit to the largest studies of each year (black dots).

pull out the ligand along a reaction coordinate, for example the distance between the ligand and the binding cavity or a vector describing a putative unbinding pathway. In practice, the ligand is attached to a spring with a given force constant and the center of the spring is moved at a constant speed in its way out from the bound pose to the bulk. By measuring the forces and work applied to pull out the ligand one is able to reconstruct a free energy profile. The novelty of this work compared to other SMD studies is that an extensive high throughput sampling approach was followed, obtaining an accurate estimation of the statistical uncertainty thanks to the 2000 simulation pulls produced (4 μs of total simulation data).

2.2. Protein-drug Binding Free Energy Calculation by Means of Metadynamics

The high complexity of drug-protein binding processes was initially tackled in terms of free energy methods such as metadynamics [16]. These methods are characterized by using extra biasing forces to, for example, drive the system away from local minima or to force the exploration of a specific region of the free energy surface. By using extra forces, one is able to sample processes that would otherwise take an unattainable computing time of unbiased simulations. Given the limited amount of computing resources at the time, these were very appealing methods to tackle drug-protein binding processes.

In particular, metadynamics relies on the definition of collective variables, *i.e.* a set of order parameters relevant to the process under study, such as protein-ligand distance in the case of the ligand-binding, certain torsion angles, hydrogen bonds distance, molecule orientation/rotation, *etc*. By progressively adding Gaussian-shaped potentials to these CVs, one is able to let the system drift away from the original coordinates and jump across energetic barriers until all minima are "filled" with a bias potential. Once the local minima are filled, if the CVs are well chosen, the system will diffuse freely among all states spanned by the CV. By adding up the applied Gaussians one can reconstruct the bias potential and by changing its sign, the free energy landscape can be obtained.

Metadynamics was applied in early 2010 to calculate the binding free energy of a congeneric series of CDK2 inhibitors [17] using ACEMD in combination with the PLUMED [18] plugin. In this work Fidelak *et al.* used (a) the distance between the binding cavity and the center of masses of the ligand and (b) a dihedral angle between two protein reference points and the ligand as CV. By adding small Gaussian potentials to these variables, they are able to produce simulations of undocking and redocking events, which is an indicator that the CV are well chosen and the runs are fully converged. An additional metadynamics run using path collective variables (PCV) [19] along the minimum-energy unbinding pathway, allowed Branduardi *et al.* to compute a free energy surface for the unbinding process, as well as the value of the binding free energy obtained by integrating the FES.

Note that, although the computational resources for MD are nowadays much more accessible to the scientific com-

munity, metadynamics applications are still relevant, such as the recent 2016 study of molecular diffusion through pores[20], of extreme importance in ADME optimization[21] or in the rationalization of drug absorption-related resistances. D'Agostino *et al.* studied the translocation of the antibiotic Meropenem through OmpF using as collective variables the extracellular-intracellular distance and the orientation of the ligand. They observed that only one ligand orientation, with a positive group first, results in actual translocation. Furthermore, they identify three key residues of great importance for the internalization of the antibiotic and proposed that their mutation could be a possible source of resistance by reducing the influx rate.

2.3. Protein-peptide Binding Free Energy Calculation by Means of Umbrella Sampling

Umbrella sampling (US) is an enhanced sampling scheme first used by Torrie and Valleau in the 1970s that consists in defining one or more reaction coordinates and simulating the system in equally spaced windows along these coordinates [22]. The reaction coordinate in the case of ligand binding, in its most simplistic form, can be the 1D unbinding pathway. In order to ensure that the system samples exhaustively a specific space of the reaction coordinate, a potential is applied to restrain the movement of the system and ensure each window is equally explored. This extra force applied is usually a harmonic potential (umbrella). By measuring the oscillation of the system along the reaction coordinate for each window, we are able to reconstruct an energy profile (potential of mean force - PMF) that will highlight the energetic barriers present in the chosen reaction coordinate. The weighted histogram analysis method (WHAM)[23] is the most widely used method to recover the PMF although other strategies can also be used [24].

Protein-peptide binding was first tackled using ACEMD in the early 2010 where it was possible to reproduce the binding affinity of the SH2 domain for the tetrapeptide pY-EEI [8]. Due to the long timescales required to reproduce spontaneous binding events, umbrella sampling (US) was employed (20.5 μs of total data). This work was a proof of concept that a volunteer distributed computing network such as GPUGRID [8] is able to produce high-throughput all-atom molecular dynamics simulations equivalent to expensive high performance computing (HPC) resources.

Later work in 2011 on the same system was not only able to reproduce the experimental binding free energy but also defined an optimal parameter set [25]. The assumption behind this work is that there is an optimal combination of force constant and window width to reproduce a binding free energy in the least amount of simulation sampling possible. Note that while in this study the umbrella sampling potential was applied to the center of mass of the peptide, in other more complex implementations such as the one by Woo and Roux [26], several potentials are applied to restrain diverse degrees of freedom including ligand conformation, orientation and radial translation along the reaction coordinate

2.4. Full Reconstruction of Protein-drug Binding Events

In 2011 sufficient computing resources were available to tackle complex events such as protein-drug binding in an unbiased manner, ie letting the systems evolve freely without adding any biasing force. A number of thermodynamic and kinetic properties could be measured from these simulations, namely: (a) ligand binding pose, (b) binding kinetics (k_{on} and k_{off}), (c) binding free energy and (d) equilibrium state distribution.

The first system to be modeled using this approach was benzamidine-trypsin. This prototypical and still relevant benchmark system includes trypsin, a protease, and benzamidine, a very small molecule with quick binding kinetics (k_{on}), which makes it ideal to reproduce binding events within affordable computing time. A pioneering large-scale study by Buch *et al.* used ACEMD to produce 495 simulations of 100 ns, 187 of which produced binding events within 2 Å [9]. The observed events enabled the computation of both thermodynamic (binding affinity) and kinetic (binding rates) properties by spatially clustering the ligand coordinates along the simulations and building a transition model called Markov State Model (MSM)[27]. MSM enable the calculation of probabilities of bulk-bound transitions at a molecular resolution, from which macroscopic binding free energies and rates can then be inferred.

A similar work was carried out later in 2014 involving the enzyme ß-lactamase and the small fragment carboxythiophene, in which binding poses and kinetics were reproduced within an aggregate simulation time of 148 microseconds [28]. Fragments are small molecules, usually with a molecular weight lower than 300 Da, which usually bind in a promiscuous way due to the small size and the relatively low affinity; this makes the binding difficult to be captured experimentally. Therefore, this work provided further evidence that the technology is useful and ready to help rationalize drug design endeavors, especially during the fragment-to-lead phase.

2.5. Full Reconstruction of Protein-peptide Binding Events

The success demonstration of the reconstruction of full drug-protein binding processes and the access to high-throughput computing resources led to pursue a new milestone in complexity: the full reconstruction of recognition process between proteins and peptides.

The work of Giorgino *et al.* in 2012 provided a simple illustration of the fact that short simulations can extensively sample a process occurring in a timescale longer than a single simulation length [29]. An ensemble of 772 independent replicas was used to recover the full association process between the flexible phosphorylated tetrapeptide pYEEI and its highly specific SH2 partner. Despite the time scale of the "search" process ($k_{on} \sim 10^6$ M^{-1} s^{-1}, [P·L]~20 mM) is relatively long with respect to the length of individual simulations (200 ns), the availability of independent replicas (>150 μs total sampled time) coupled with the relative rapidity of the association itself (rare uncorrelated events) provided sufficient statistics for the direct computation of the association rate.

2.6. Protein-protein Binding Studies

A special type of protein-ligand binding is when the ligand is another protein. Some protein-protein interaction

(PPI) processes, such as antigen-antibody recognition, are based on large interaction surfaces and involve rearrangement of structural elements or side-chains at either or both surfaces. Therefore, the complexity in protein-protein interactions is much higher due to an increase in the number of degrees of freedom and that the conformational rearrangements necessary for binding usually require larger timescales. This is why by 2012 biased methods like umbrella sampling started to be used to explore protein-protein binding, as a step beyond the use of simple MD runs to check complex stabilities.

One of the first applications of US in the study of binding between two extended protein-protein binding was developed by Buch *et al.* in 2013. In this study, our group calculated the affinity of the cetuximab, a chimeric (mouse/human) IgG1 anti-EGFR monoclonal antibody for EGFR, in presence and absence of the S468R resistance-inducing mutation. US showed how the mutation simultaneously reduces the affinity for the antibody Cetuximab, and increases that for the endogenous ligand, thus shifting the competitive balance between the two and diminishing the therapeutic effectiveness of the former [30].

Selent *et al.* also studied protein-protein binding by producing short MD simulations of 20 ns to refine a predicted complex between survivin/CDK4 [31].

2.7. MD as a Drug-protein Complex Post-processing Tool

Thanks to the success of the first relatively simple drug-protein models like benzamidine-trypsin, applications of MD in drug discovery increased. Specifically, pharmacologically relevant systems such as the GPCRs started receiving great interest, partially motivated by a flourishing number of crystallographic structures becoming available. However, the bigger size of the compounds under study and the high number of them made the full binding pathway reconstruction approach unattainable due to sampling limitations. Instead, the strategy followed consisted in producing short MD runs to validate and refine ligand-receptor binding poses produced by docking or obtained from crystal structures.

The advantages of using MD after a docking prediction are multiple: (**a**) it allows us to study the ligand dynamics in the binding pocket and (**b**) it allows us to reveal the role of water molecules within the binding site, which can be important contributors to the ligand binding free energy [32].

Sabbadin *et al.* followed this approach in their work on GPCR-ligand complexes in 2014 [33]. They used ACEMD to produce simulations of the A_{2A} adenosine receptor with ligands placed in cocrystallized and decoy poses and calculated individual electrostatic and hydrophobic contributions to the interaction energy of each protein residue contacting the ligand. These "interaction energy fingerprints" (IEF) were able to discern the bioactive conformation from the decoy ones and are proposed as an alternative to the conventional scoring functions.

Finally, is worth mentioning our 2014 study on Linear Interaction Energies (LIE) as an alternative method to rank ligands [34]. 1549 compounds including real binders and decoys were simulated in two states: (a) docked to a trypsin binding cavity and (b) alone in solvent. By computing the difference of interaction energies between bound and unbound states, we were able to predict binding free energies. Although the LIE method proved effective, did not yield results better than high quality docking algorithms such as GOLD [35] and GLIDE [36].

2.8. Smarter Sampling Schemes Allow Multiple Full Ligand-protein Binding Pathway Reconstructions within Affordable Time

Early work on benzamidine-trypsin established the foundations for what later would become known as high throughput molecular dynamics (HTMD), a new paradigm where ligand-receptor binding poses, affinities and kinetics can be reproduced within affordable time by producing multiple parallel MD simulations [2]. However, late in 2013 it became evident that to fully take advantage of this unbiased high-throughput approach, smarter and automatic sampling schemes should be designed to reduce the computational cost.

Inspired on previous work by Vijay Pande group [37]–[39], a method called adaptive sampling [40] was released in early 2014. It consists in an on-the-fly learning scheme that reduces by an order of magnitude the necessary sampling time to reproduce the benzamidine-trypsin binding affinity. This protocol periodically produces and analyses simulations by building a MSM and respawning simulations from "hot-spots" such as ligand metastable poses in a completely automated way.

An additional enhanced sampling approach also using ACEMD called supervised MD (suMD) was developed by the group of Stefano Moro [41]. The rationale behind this protocol is to produce very short simulations and "respawn" from those where the ligand got closer to the binding site. Therefore, sampling focuses to recover a putative binding pathway within nanosecond-scale aggregate time. Differently from adaptive sampling, this technique requires the prior knowledge of a binding site. This is similar in spirit to the FAST method [42].

As shown in our work by Ferruz *et al.* in 2015, the usage of adaptive sampling proved effective to reproduce the binding poses of 12 out of 15 and kinetics for 4 out of 6 fragments against protease factor Xa (2.1 ms of total simulation time) [43]. The technology also revealed secondary poses, usually hidden to other techniques such as crystallography.

2.9. Full Protein-lipid Binding Pathway Reconstruction through the Membrane

Protein-lipid binding is a poorly-explored process at the atomic level. The fact that the lipidic ligands bind to their targets through the membrane instead from the solvent like previous non-lipidic soluble ligands, adds a layer of complexity in the production and analysis of simulations.

In 2016, Stanley *et al.* were able to resolve the binding of the lipid inhibitor ML056 to the sphingosine-1-phosphate receptor 1 ($S1P_1R$) using unbiased molecular dynamics simulations with an aggregate sampling of over 800 μs [44]. The pathway was described to be a multistage process involving a first rate-limiting step consisting in the ligand diffusing in the bilayer leaflet to the "membrane vestibule" at

the top of TM7 and then the ligand moving through a channel formed by TM1, TM7 and the N-terminal of the receptor to the orthosteric binding cavity, where it finally adopts the crystallographic pose within 1 Å. The presence of multiple lipid ligands, instead of a single ligand like in previous work, required the use of a special Markov State Model by defining a volumetric map based on the lipid phosphate occupancy. Thanks to this model, the three key slowest processes and their respective relaxation timescales were identified: (1) flipping of the ligand between bilayers (~100 μs), (2) transition from the upper leaflet to the "membrane vestibule" (1–10 μs) and (3) transition from the vestibule to the bound pose (~500 ns). As the lipid flipping is not involved in the binding process, the rate-limiting step was defined to be the diffusion from the leaflet to the vestibule.

2.10. Multibody Binding Studies

In a recent work, Ferruz *et al.*, a complex multi-cofactor and substrate binding to *myo*-inositol monophosphatase (IMPase) and its underlying dynamic interplay was studied by means of all-atom MD simulations and using MSM and adaptive sampling scheme [45]. IMPase is an enzyme that depends on 3 Mg^{2+} ions as cofactors for its catalytic activity - their binding poses are known but the way cofactors and substrate cooperate is unknown. Thanks to extensive all-atom MD simulations (0.8ms of total simulation time) a mechanism is proposed by which the first two ions bind with a very quick k_{on} and remain bound for very long timescales while the third ion binds with very slow k_{on}. Finally, the substrate can bind to IMPase bound to two or three ions, being the later scenario orders of magnitude faster.

This multi body pathway reconstruction is an example of how the protein-ligand binding field has evolved since its humble beginnings with simple protein-ion systems. It represents well the complexity level accessible by our current technology and opens the door to future allosterism and cooperativity studies currently unattainable.

2.11. Conformational Rearrangements and Unstructured Proteins

An important feature of MD is its ability to sample the conformational space of macromolecules in solution. While crystallography-derived structures provide essentially static conformations, and magnetic resonance (NMR) experiment provide a limited range of dynamic information, the conformational space sampled by MD is largely determined by the computing resources available. Pioneering studies have indeed been able to capture full and repeated folding and unfolding of domains, albeit on selected fast-folding targets [45,46]. Extended sampling is especially important for targets undergoing large conformational changes during their activity. Here, we highlight some work done using ACEMD that represents some of the contributions and the potential of MD to (a) pinpoint transient events and (b) uncover pockets and mechanisms susceptible to be targeted by drugs acting as agonist, antagonist, or allosteric modulators.

For example, Sadiq *et al.* [48] were able to describe the self-association of immature HIV-1 protease to its extended amino-terminus recognition motif. Maturation of the HIV virus is a complex process which occurs through the repeated auto-cleavage of its GagPol polypeptide precursor chain. The association, preliminary to the catalytic cleavage, requires a "wide opening" of the HIV dimer flaps; the self-binding process was observed and its rate estimated *via* Markov-state analysis.

On the other hand, the emergence of antibiotic resistance and the relenting pace of the discovery of novel antibiotics are motivating the study of complex bacterial membranes and transporters. Ceccarelli and Winterhalter [49] studied the translocation of antibiotics through four mutants of the OmpC protein pores found in clinical isolates of Escherichia coli. Metadynamics-based simulations provided free energy profiles of translocation and the corresponding hydrogen-bonding network. Also on the topic of bacterial membranes, Kong *et al.* [50] studied the multi-microsecond dynamics of export through the Wza transporter of K30 oligosaccharides (part of bacteria's protective layer) of various lengths. Sattelle *et al.* [51] used multi-microsecond simulations to study the dynamics of oligosaccharides in water, modeled in the GLYCAM force field [52].

Computational techniques are especially promising for the rationalization of the relatively underexplored area of allosteric communication, *i.e.* inter-protein interactions allowing molecules to modulate the activity of a target even though their interaction site does not coincide with the active site (orthosteric pocket) of the protein [53]. For instance, the previously mentioned metadynamics study of Selent *et al.* [13] showed an allosteric effect of a sodium ion on a known allosteric modulation site (Asp2.50) in the dopaminergic D2 receptor. More generally, GPCRs and gated ion channels have a prominent interest in drug discovery; the complexity of gating motions, coupled with the membrane environment, makes the molecular description of inter-state transitions very challenging.

An area of computational research which is being opened by advances in both computing power and force field accuracy [54] is that of intrinsically disordered proteins (IDP). IDP are a class of biomolecules which are devoid - to varying degrees - of an ordered structure in at least some conditions. They are involved in numerous biological processes, and especially in promiscuous recognition, some of them undergoing fold-upon-bind transitions [54,55]. Unfortunately, the absence of a defined tertiary structure hinders the development of active drugs against them following structure-based drug design (SBDD) approaches. The huge dimensionality of the phase space of IDP domains makes the characterization of IDPs especially challenging; an important step towards the analysis of their kinetics was the introduction of time-lagged independent component analysis (tICA), which projects the data along coordinates spanning "slow processes" usually corresponding the more interesting collective degrees of freedom, thus enabling a meaningful discretization of the configuration space [57]. Stanley *et al.* used tICA and large-scale MSM to show how the phosphorylation of S133 in the unstructured kinase inducible domain (KID) alters the rate of exchanges between an ordered and a disordered transient state, thus again showing a dynamic-equilibrium shift mechanism, here involving a transition state [58].

3. PAST, PRESENT AND FUTURE CHALLENGES OF MD

The state of the art in classical MD has its roots in research performed in the '70s. It is thanks to extensive and detailed refinements in force fields conducted in the last decades, and still ongoing, that we can extract kinetic information and extrapolate timescales even for conformations outside of the typical "stability basins" of backbones and side chains. Force-fields are still ongoing refinements, and this is especially important because they are now covering non-protein macromolecules, as we have seen in the review: small molecules [59], lipids [58], and proteoglycans [51], ions, *etc*.

In this context, the development of a large-scale architecture specialized for all-atom MD simulations, such as the Anton[3] and Anton 2[61, p. 2], has provided an impulse to the development and validation of force-fields. While Anton machines are not available for general purchase or use, the availability of continuous-trajectory runs several orders of magnitudes longer that what was previously available created "natural" benchmark sets for challenging tasks such as gating channels conformational transitions [62] and folding-unfolding transitions[63]. Verifying that force-field are continuously refined, transferrable between systems, and valid in long-term extrapolations is an invaluable reassurance on the ability of large-scale MD simulations, regardless of the chosen infrastructure, to provide reliable results (in terms of both thermodynamics and kinetics) in the long timescales implied by high-throughput MD experiments.

The parameterization of small chemical entities is especially important for drug design; force fields have been addressing the need to faithfully transfer behaviors which can be described with high-level quantum approaches into classical counterparts used in MD. Given the combinatorial explosion of atom types in small molecules, the number of parameters to be fitted is much higher. Recent force fields [64]–[66] and tools [67], [68] are now providing systematic protocols for addressing this issue.

Another limitation of classical MD is the commonly made assumption of constant protonation states for titratable groups. This may be a good approximation when conformational changes are limited (software such as PROPKA [69] and PDB2PQR [70] can predict titration states and optimize H bonding networks), but "static" approaches are bound to fail *e.g.* for residues seeing large changes in solvation shells. So-called constant-pH simulations, although not yet mainstream, have been devised to lift this limitation [71]. Similarly, efforts are underway towards force fields accounting for charge polarization [72, 73]; more in general, chemical reactions can be accounted by hybrid QM-MM schemes [74].

Another obstacle to the widespread adoption of MD is the material process of preparing systems for simulation. PDB files must undergo several pre-processing steps before a simulation-ready system is obtained, namely *in silico* solvation, parameterization, titration, and so on. These steps can be performed manually in interactive software like VMD, PyMol and Chimera, but interactive manipulation is not reproducible (*e.g.* replicated for different molecules or by independent researchers). Platforms like HTMD [75] are being developed to overcome the "ad-hoc" approach to system preparation making it more (a) streamlined, (b) self-documenting and (c) reproducible. Similar considerations apply to the post-production steps, *i.e.* when large amounts of molecular trajectories have to be analyzed to extract the structural, thermodynamic or kinetic quantities of interest; manual inspection is ruled out, and even the creation of custom analysis scripts is cumbersome and error prone, especially in low-level languages. The availability of structural-biology specific analysis primitives high-throughput software overcomes these difficulties by streamlining and standardizing the analysis steps, encoding them into self-contained, self-documenting and easily re-runnable documents [76].

The predicted computational cost reduction and the development of new analysis tools should be able to foster research lines involving models with either longer timescales or bigger size. In terms of timescales, effort is being focused, for example, in reproducing protein-protein binding with unbiased simulations, which has been classically hindered by sampling and analysis limitations. As computation cost decreases, one can expect the scientific community becoming ready to face new research challenges such as more complex pathway studies or even all-atom mesoscale simulations.

A higher accessibility to computation time, as well as the advent of accurate analysis tools and generated experience in the field should motivate the scientific community to move from the classical perspective studies, where mere reproduction is pursued to validate the effectiveness of the method, to prospective applications yielding actual predictions that can serve to the experimentalists to guide and rationalize drug discovery endeavors.

CONFLICT OF INTEREST

The authors declare the following competing financial interest(s): G.D.F. and M.J.H. are shareholders of Acellera Ltd.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] A. R. Leach, Molecular Modelling: Principles and Applications. Pearson Education, 2001.

[2] M. J. Harvey and G. De Fabritiis, "High-throughput molecular dynamics: the powerful new tool for drug discovery," Drug Discov. Today, vol. 17, no. 19–20, pp. 1059–1062, Oct. 2012.

[3] D. E. Shaw et al., "Anton, a special-purpose machine for molecular dynamics simulation," Commun. ACM, vol. 51, no. 7, p. 91, Jul. 2008.

[4] M. J. Harvey, G. Giupponi, and G. De Fabritiis, "ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale," J. Chem. Theory Comput., vol. 5, no. 6, pp. 1632–1639, Jun. 2009.

[5] R. Salomon-Ferrer, A. W. Götz, D. Poole, S. Le Grand, and R. C. Walker, "Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald," J. Chem. Theory Comput., vol. 9, no. 9, pp. 3878–3888, Sep. 2013.

[6] P. Eastman et al., "OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation," J. Chem. Theory Comput., vol. 9, no. 1, pp. 461–469, Jan. 2013.

[7] K. J. Bowers et al., "Scalable algorithms for molecular dynamics simulations on commodity clusters," in Proceedings of the 2006 ACM/IEEE conference on Supercomputing, Tampa, Florida, 2006, p. 84.

[8] I. Buch, M. J. Harvey, T. Giorgino, D. P. Anderson, and G. De Fabritiis, "High-Throughput All-Atom Molecular Dynamics Simulations Using Distributed Computing," J. Chem. Inf. Model., vol. 50, no. 3, pp. 397–403, Mar. 2010.

[9] I. Buch, T. Giorgino, and G. De Fabritiis, "Complete reconstruction of an enzyme-inhibitor binding process by molecular dynamics simulations," Proc. Natl. Acad. Sci., vol. 108, no. 25, pp. 10184–10189, Jun. 2011.

[10] F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich, and T. R. Weikl, "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations," Proc. Natl. Acad. Sci., vol. 106, no. 45, pp. 19011–19016, Nov. 2009.

[11] J. Hughes, S. Rees, S. Kalindjian, and K. Philpott, "Principles of early drug discovery," Br. J. Pharmacol., vol. 162, no. 6, pp. 1239–1249, Mar. 2011.

[12] M. De Vivo, M. Masetti, G. Bottegoni, and A. Cavalli, "Role of Molecular Dynamics and Related Methods in Drug Discovery," J. Med. Chem., vol. 59, no. 9, pp. 4035–4061, May 2016.

[13] J. Selent, F. Sanz, M. Pastor, and G. De Fabritiis, "Induced Effects of Sodium Ions on Dopaminergic G-Protein Coupled Receptors," PLoS Comput Biol, vol. 6, no. 8, p. e1000884, 2010.

[14] W. Liu et al., "Structural Basis for Allosteric Regulation of GPCRs by Sodium Ions," Science, vol. 337, no. 6091, pp. 232–236, Jul. 2012.

[15] T. Giorgino and G. De Fabritiis, "A High-Throughput Steered Molecular Dynamics Study on the Free Energy Profile of Ion Permeation through Gramicidin A," J Chem Theory Comput, vol. 7, no. 6, pp. 1943–1950, 2011.

[16] A. Laio and F. L. Gervasio, "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science," Rep. Prog. Phys., vol. 71, no. 12, p. 126601, 2008.

[17] J. Fidelak, J. Juraszek, D. Branduardi, M. Bianciotto, and F. L. Gervasio, "Free-Energy-Based Methods for Binding Profile Determination in a Congeneric Series of CDK2 Inhibitors," J. Phys. Chem. B, vol. 114, no. 29, pp. 9516–9524, Jul. 2010.

[18] M. Bonomi et al., "PLUMED: A portable plugin for free-energy calculations with molecular dynamics," Comput. Phys. Commun., vol. 180, no. 10, pp. 1961–1972, Oct. 2009.

[19] D. Branduardi, F. L. Gervasio, and M. Parrinello, "From A to B in free energy space," J. Chem. Phys., vol. 126, no. 5, pp. 54103-54103–10, Feb. 2007.

[20] T. D'Agostino, S. Salis, and M. Ceccarelli, "A kinetic model for molecular diffusion through pores," Biochim. Biophys. Acta BBA - Biomembr., vol. 1858, no. 7, Part B, pp. 1772–1777, Jul. 2016.

[21] G. W. Caldwell, Z. Yan, W. Tang, M. Dasgupta, and B. Hasting, "ADME optimization and toxicity assessment in early- and late-phase drug discovery," Curr. Top. Med. Chem., vol. 9, no. 11, pp. 965–980, 2009.

[22] G. M. Torrie and J. P. Valleau, "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling," J. Comput. Phys., vol. 23, no. 2, pp. 187–199, 1977.

[23] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, "THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method," J. Comput. Chem., vol. 13, no. 8, pp. 1011–1021, Oct. 1992.

[24] J. Kästner, "Umbrella sampling," Wiley Interdiscip. Rev. Comput. Mol. Sci., vol. 1, no. 6, pp. 932–942, Nov. 2011.

[25] I. Buch, S. K. Sadiq, and G. De Fabritiis, "Optimized Potential of Mean Force Calculations for Standard Binding Free Energies," J. Chem. Theory Comput., vol. 7, no. 6, pp. 1765–1772, Jun. 2011.

[26] H.-J. Woo and B. Roux, "Calculation of absolute protein-ligand binding free energy from computer simulations.," Proc. Natl. Acad. Sci. U. S. A., vol. 102, no. 19, pp. 6825–6830, Maggio 2005.

[27] F. Noé and S. Fischer, "Transition networks for modeling the kinetics of conformational change in macromolecules," Curr. Opin. Struct. Biol., vol. 18, no. 2, pp. 154–162, Apr. 2008.

[28] P. Bisignano, S. Doerr, M. J. Harvey, A. D. Favia, A. Cavalli, and G. De Fabritiis, "Kinetic Characterization of Fragment Binding in AmpC β-Lactamase by High-Throughput Molecular Simulations," J. Chem. Inf. Model., vol. 54, no. 2, pp. 362–366, Feb. 2014.

[29] T. Giorgino, I. Buch, and G. De Fabritiis, "Visualizing the Induced Binding of SH2-Phosphopeptide," J Chem Theory Comput, vol. 8, no. 4, pp. 1171–1175, 2012.

[30] I. Buch, N. Ferruz, and G. De Fabritiis, "Computational modeling of an epidermal growth factor receptor single-mutation resistance to cetuximab in colorectal cancer treatment," J. Chem. Inf. Model., vol. 53, no. 12, pp. 3123–3126, Dec. 2013.

[31] J. Selent, A. A. Kaczor, R. Guixà-González, P. Carrió, M. Pastor, and C. Obiol-Pardo, "Rational design of the survivin/CDK4 complex by combining protein–protein docking and molecular dynamics simulations," J. Mol. Model., vol. 19, no. 4, pp. 1507–1514, Dec. 2012.

[32] S. B. A. de Beer, N. P. E. Vermeulen, and C. Oostenbrink, "The role of water molecules in computational drug design," Curr. Top. Med. Chem., vol. 10, no. 1, pp. 55–66, 2010.

[33] D. Sabbadin, A. Ciancetta, and S. Moro, "Bridging Molecular Docking to Membrane Molecular Dynamics To Investigate GPCR–Ligand Recognition: The Human A2A Adenosine Receptor as a Key Study," J. Chem. Inf. Model., vol. 54, no. 1, pp. 169–183, Jan. 2014.

[34] G. Lauro, N. Ferruz, S. Fulle, M. J. Harvey, P. W. Finn, and G. De Fabritiis, "Reranking docking poses using molecular simulations and approximate free energy methods," J. Chem. Inf. Model., vol. 54, no. 8, pp. 2185–2189, Aug. 2014.

[35] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking1," J. Mol. Biol., vol. 267, no. 3, pp. 727–748, Apr. 1997.

[36] R. A. Friesner et al., "Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy," J. Med. Chem., vol. 47, no. 7, pp. 1739–1749, Mar. 2004.

[37] N. Singhal and V. S. Pande, "Error analysis and efficient sampling in Markovian state models for molecular dynamics," J. Chem. Phys., vol. 123, no. 20, p. 204909, Nov. 2005.

[38] N. S. Hinrichs and V. S. Pande, "Calculation of the distribution of eigenvalues and eigenvectors in Markovian state models for molecular dynamics," J. Chem. Phys., vol. 126, no. 24, p. 244101, Jun. 2007.

[39] J. K. Weber and V. S. Pande, "Characterization and Rapid Sampling of Protein Folding Markov State Model Topologies," J. Chem. Theory Comput., vol. 7, no. 10, pp. 3405–3411, Oct. 2011.

[40] S. Doerr and G. De Fabritiis, "On-the-Fly Learning and Sampling of Ligand Binding by High-Throughput Molecular Simulations," J. Chem. Theory Comput., vol. 10, no. 5, pp. 2064–2069, May 2014.

[41] D. Sabbadin and S. Moro, "Supervised molecular dynamics (SuMD) as a helpful tool to depict GPCR-ligand recognition pathway in a nanosecond time scale," J. Chem. Inf. Model., vol. 54, no. 2, pp. 372–376, Feb. 2014.

[42] M. I. Zimmerman and G. R. Bowman, "FAST Conformational Searches by Balancing Exploration/Exploitation Trade-Offs," J. Chem. Theory Comput., vol. 11, no. 12, pp. 5747–5757, Dec. 2015.

[43] N. Ferruz, M. J. Harvey, J. Mestres, and G. De Fabritiis, "Insights from Fragment Hit Binding Assays by Molecular Simulations," J. Chem. Inf. Model., vol. 55, no. 10, pp. 2200–2205, Oct. 2015.

[44] N. Stanley, L. Pardo, and G. D. Fabritiis, "The pathway of ligand entry from the membrane bilayer to a lipid G protein-coupled receptor," Sci. Rep., vol. 6, Mar. 2016.

[45] N. Ferruz, G. Tresadern, A. Pineda-Lucena, and G. De Fabritiis, "Multibody cofactor and substrate molecular recognition in the myo-inositol monophosphatase enzyme," Sci. Rep., vol. 6, p. 30275, 2016.

[46] D. L. Ensign, P. M. Kasson, and V. S. Pande, "Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece," J. Mol. Biol., vol. 374, no. 3, pp. 806–816, Nov. 2007.

[47] V. A. Voelz, G. R. Bowman, K. Beauchamp, and V. S. Pande, "Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1-39)," J. Am. Chem. Soc., vol. 132, no. 5, pp. 1526–1528, Feb. 2010.

[48] S. K. Sadiq, F. Noé, and G. D. Fabritiis, "Kinetic characterization of the critical step in HIV-1 protease maturation," Proc. Natl. Acad. Sci., Nov. 2012.

[49] H. Bajaj et al., "Molecular Basis of Filtering Carbapenems by Porins from β-Lactam-resistant Clinical Strains of Escherichia coli," J. Biol. Chem., vol. 291, no. 6, pp. 2837–2847, Feb. 2016.

[50] L. Kong, A. Almond, H. Bayley, and B. G. Davis, "Chemical polyglycosylation and nanolitre detection enables single-molecule reca-

pitulation of bacterial sugar export," Nat. Chem., vol. 8, no. 5, pp. 461–469, May 2016.
[51] B. M. Sattelle and A. Almond, "Shaping up for structural glycomics: a predictive protocol for oligosaccharide conformational analysis applied to N-linked glycans," Carbohydr. Res., vol. 383, pp. 34–42, Jan. 2014.
[52] K. N. Kirschner et al., "GLYCAM06: A generalizable biomolecular force field. Carbohydrates," J. Comput. Chem., vol. 29, no. 4, pp. 622–655, Mar. 2008.
[53] R. Nussinov and C.-J. Tsai, "Allostery in Disease and in Drug Discovery," Cell, vol. 153, no. 2, pp. 293–305, Apr. 2013.
[54] J. Henriques, C. Cragnell, and M. Skepö, "Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment," J. Chem. Theory Comput., vol. 11, no. 7, pp. 3420–3431, Jul. 2015.
[55] P. Tompa, "Unstructural biology coming of age," Curr. Opin. Struct. Biol., vol. 21, no. 3, pp. 419–425, Jun. 2011.
[56] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," Nat Rev Mol Cell Biol, vol. 6, no. 3, pp. 197–208, Mar. 2005.
[57] G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, "Identification of slow molecular order parameters for Markov model construction," J. Chem. Phys., vol. 139, no. 1, p. 15102, Jul. 2013.
[58] N. Stanley, S. Esteban-Martín, and G. De Fabritiis, "Kinetic modulation of a disordered protein domain by phosphorylation," Nat. Commun., vol. 5, Oct. 2014.
[59] J. B. Klauda, V. Monje, T. Kim, and W. Im, "Improving the CHARMM Force Field for Polyunsaturated Fatty Acid Chains," J. Phys. Chem. B, 2012.
[60] J. Wang, W. Wang, P. A. Kollman, and D. A. Case, "Automatic atom type and bond type perception in molecular mechanical calculations," J. Mol. Graph. Model., vol. 25, no. 2, pp. 247–260, Oct. 2006.
[61] D. E. Shaw et al., "Anton 2: Raising the Bar for Performance and Programmability in a Special-purpose Molecular Dynamics Supercomputer," in Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, Piscataway, NJ, USA, 2014, pp. 41–53.
[62] M. Ø. Jensen, V. Jogini, D. W. Borhani, A. E. Leffler, R. O. Dror, and D. E. Shaw, "Mechanism of voltage gating in potassium channels," Science, vol. 336, no. 6078, pp. 229–233, Apr. 2012.
[63] S. Piana, K. Lindorff-Larsen, and D. E. Shaw, "How robust are protein folding simulations with respect to force field parameterization?," Biophys. J., vol. 100, no. 9, pp. L47-49, May 2011.
[64] J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman, and D. A. Case, "Development and testing of a general amber force field," J. Comput. Chem., vol. 25, no. 9, pp. 1157–1174, Jul. 2004.
[65] K. Vanommeslaeghe and A. D. MacKerell, "Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing," J. Chem. Inf. Model., vol. 52, no. 12, pp. 3144–3154, Dec. 2012.
[66] D. L. Mobley, É. Dumont, J. D. Chodera, and K. A. Dill, "Comparison of Charge Models for Fixed-Charge Force Fields: Small-Molecule Hydration Free Energies in Explicit Solvent," J. Phys. Chem. B, vol. 111, no. 9, pp. 2242–2254, Mar. 2007.
[67] C. G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J. C. Gumbart, "Rapid parameterization of small molecules using the force field toolkit," J. Comput. Chem., vol. 34, no. 32, pp. 2757–2770, Dicembre 2013.
[68] J. D. Yesselman, D. J. Price, J. L. Knight, and C. L. Brooks, "MATCH: an atom-typing toolset for molecular mechanics force fields," J. Comput. Chem., vol. 33, no. 2, pp. 189–202, Jan. 2012.
[69] M. H. M. Olsson, C. R. Søndergaard, M. Rostkowski, and J. H. Jensen, "PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions," J. Chem. Theory Comput., vol. 7, no. 2, pp. 525–537, Feb. 2011.
[70] T. J. Dolinsky et al., "PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations," Nucleic Acids Res., vol. 35, no. suppl 2, pp. W522–W525, Jul. 2007.
[71] S. Donnini, R. T. Ullmann, G. Groenhof, and H. Grubmüller, "Charge-Neutral Constant pH Molecular Dynamics Simulations Using a Parsimonious Proton Buffer," J. Chem. Theory Comput., vol. 12, no. 3, pp. 1040–1051, Mar. 2016.
[72] Y. Shi et al., "Polarizable Atomic Multipole-Based AMOEBA Force Field for Proteins," J. Chem. Theory Comput., vol. 9, no. 9, pp. 4046–4063, Sep. 2013.
[73] C. M. Baker, "Polarizable force fields for molecular dynamics simulations of biomolecules," Wiley Interdiscip. Rev. Comput. Mol. Sci., vol. 5, no. 2, pp. 241–254, Mar. 2015.
[74] H. M. Senn and W. Thiel, "QM/MM Methods for Biomolecular Systems," Angew. Chem. Int. Ed., vol. 48, no. 7, pp. 1198–1229, Feb. 2009.
[75] S. Doerr, M. J. Harvey, F. Noé, and G. De Fabritiis, "HTMD: High-Throughput Molecular Dynamics for Molecular Discovery," J. Chem. Theory Comput., vol. 12, no. 4, pp. 1845–1852, Apr. 2016.
[76] F. Perez and B. E. Granger, "IPython: A System for Interactive Scientific Computing," Comput. Sci. Eng., vol. 9, no. 3, pp. 21–29, May 2007.

DISCLAIMER: The above article has been published in Epub (ahead of print) on the basis of the materials provided by the author. The Editorial Department reserves the right to make minor modifications for further improvement of the manuscript.

PMID: 28413955