Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Simulations meet machine learning in structural biology

Adrià Pérez[1], Gerard Martínez-Rosell[1] and Gianni De Fabritiis[1,2]

Classical molecular dynamics (MD) simulations will be able to reach sampling in the second timescale within five years, producing petabytes of simulation data at current force field accuracy. Notwithstanding this, MD will still be in the regime of low-throughput, high-latency predictions with average accuracy. We envisage that machine learning (ML) will be able to solve both the accuracy and time-to-prediction problem by learning predictive models using expensive simulation data. The synergies between classical, quantum simulations and ML methods, such as artificial neural networks, have the potential to drastically reshape the way we make predictions in computational structural biology and drug discovery.

Addresses
[1] Computational Biophysiscs Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), Doctor Aiguader 88, 08003 Barcelona, Spain
[2] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, Barcelona 08010, Spain

Corresponding author: De Fabritiis, Gianni (gianni.defabritiis@upf.edu)

Current Opinion in Structural Biology 2018, **49**:139–144

This review comes from a themed issue on **Theory and simulation**

Edited by **Kresten Lindorff-Larsen** and **Robert Best**

https://doi.org/10.1016/j.sbi.2018.02.004

Introduction

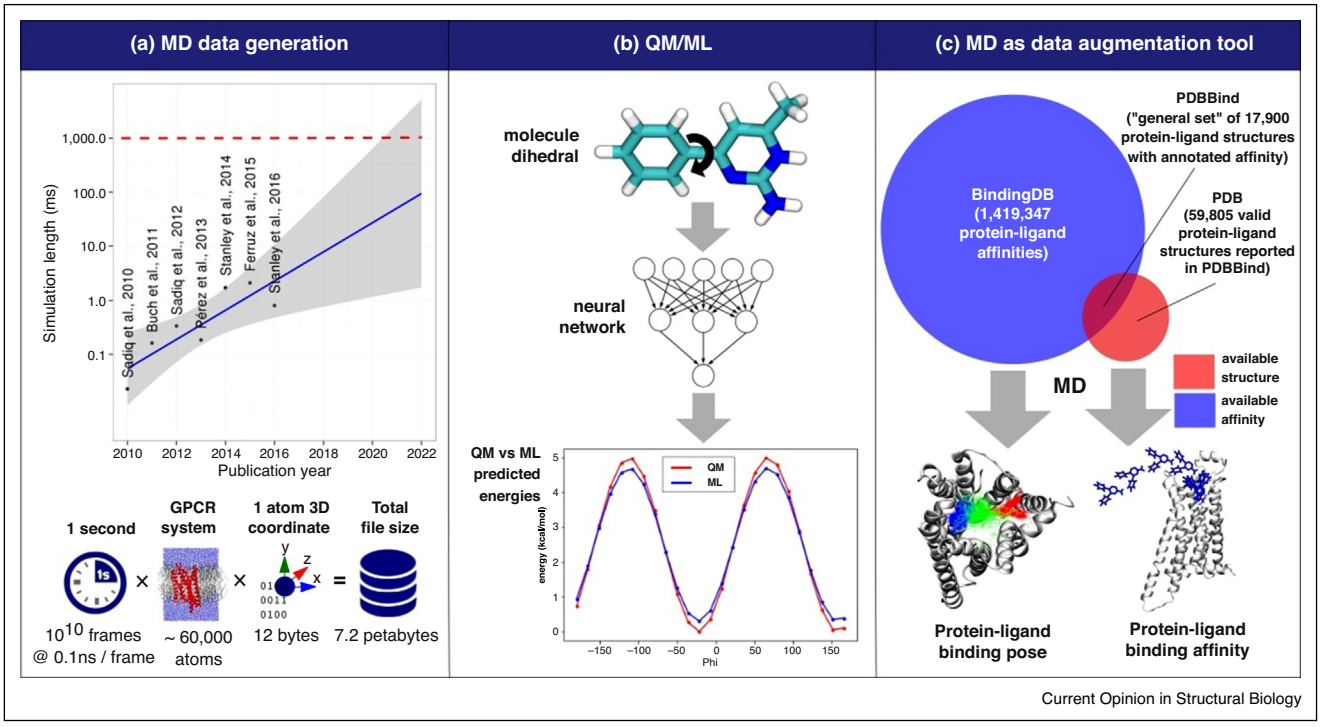
Molecular dynamics (MD) simulations are one of the predominant techniques to study protein dynamics. MD is often used to capture dynamical processes of proteins across different timescales with atomistic details in order to rationalize biological phenomena. Despite the potential to become a surrogate model of real protein dynamics, some important issues still remain to be solved, mainly: high computational cost and sampling limitations [1] and force field accuracy [2–4].

Classical MD simulations constitute a balance between accuracy and efficiency. For example, quantum-level phenomena such as enzymatic reactions, polarizability and proton transfers are neglected in exchange for computational speed. Commonly used force fields, based on a parameterization of a closed form potential, are fast to compute, but use approximations that forfeit accuracy. The extent to which these limitations may affect the validity of the results depends on the system and the biological question at hand. Quantum mechanics (QM) calculations can be used to obtain an accurate description of a molecule, but are computationally demanding and very limited in terms of sampling. Ideally, one would like to simulate at quantum level accuracy, which describes the physics and chemistry precisely, but at the sampling regime of current classical simulations.

The first simulation of protein dynamics dates from 1977 and consisted of a 9.2 ps trajectory of the bovine pancreatic trypsin inhibitor (BPTI) in vacuum [5]. In 2010, [6] reported a 1 ms trajectory of the same protein in explicit solvent, which constitutes a 100 million increase in trajectory length compared to the first simulation. In 30 years, MD simulations have increased sampling capabilities over 8 orders of magnitude, with increasing accuracy in the force fields [2–4]. In the last 10 years, MD has evolved from single simulation [7–9] to high-throughput molecular dynamics studies [10–15,16•], where hundreds of microseconds of simulations are computed in independent trajectories to obtain converged statistics. Software and hardware innovations, such as the implementation of MD codes for GPUs [17–20], distributed computing projects like Folding@home [21], GPUGRID [22] and the development of special-purpose supercomputers like ANTON [23], are steadily decreasing the computational cost of molecular simulations. Additionally, the development of adaptive sampling schemes has introduced more efficient ways to sample conformational space, decreasing the amount of simulations needed [24–26].

In a recent review we estimated that MD would reach seconds of aggregated sampling using commodity hardware by 2022 [27] (Figure 1a), generating petabytes of simulation data. For instance, the file size of one second of simulation data of a 60 000-atom system (e.g. a GPCR system) at 0.1 ns per frame is 7.2 Petabytes (reduced to a third using compressed trajectory file formats). This amount of data constitutes a valuable source of information, but the knowledge extracted from it is mainly used to rationalize a particular protein system at hand, not to generalize it to other systems. In this review, we envision a paradigm change in the near future where expensive simulations (QM and MD) are not used to predict but to learn models, so that further predictions can be drawn using ML approaches. By doing so, the large computational cost required by simulations becomes justifiable, in particular if the results are more accurate by the use of more expensive simulation methods.

Figure 1



Overview of a combined simulation and machine learning approach. **(a)** MD data generation is expected to reach the second aggregated timescale by 2022 and an output files size of several petabytes by 2022 based on a trend of maximum aggregated time per paper per year using the ACEMD software. Chart adapted from [27]. Referenced publications correspond to [12,13,15,29,56–58]. **(b)**. A first example of ML replacing QM to predict dihedral energies given a neural network trained with QM simulations. **(c)**. An example of data augmentation by MD: augment protein–ligand binding poses for a set of protein–ligand pairs with unknown binding mode; augment binding affinity data for a set of resolved protein–ligand complex structures of unknown affinities.

Machine learning applied to structural biology

ML approaches are not new in simulation analysis. For instance, the common analysis pipeline for MD simulations involves dimensionality reduction [28–33]) and clustering algorithms.

In the last few years, ML applications have grown exponentially. One of the main factors driving this growth is the broad popularization of a particular type of ML called deep neural networks [34,35]. An artificial neural network (NN) is a simple mathematical framework organized in layers, each of them performing a matrix multiplication and a non-linear function of the input variables $\boldsymbol{x}$. The output of a single neuron ϕ in each layer is given by $\phi = f(w^t x + b)$, where w are learnable weights, b is a bias and f is some nonlinear function. NNs can have several to hundred of nested layers and in such cases is called "deep". Given enough parameters, a NN is capable of interpolating any continuous function [36,37].

The application of NN models in computational biology is steadily increasing [38]. For instance, the Merck molecular activity challenge demonstrated the potential of deep neural network models in drug discovery [39]. DeepTox [40] is a deep learning-based model for toxicity prediction of compounds, winning the Tox21 toxicology prediction challenge in 2014 by a large margin. Variational autoencoders [41], a generative flavor of deep NNs, were recently applied to convert discrete representations of molecules to and from a multidimensional continuous representation [42], allowing for efficient search and optimization through open-ended spaces of chemical compounds. Additionally, autoencoders have also been used for dimensionality reduction in MD [43–45]. VAMPnets [46] fit a Markov state model from the system specific molecular simulation data. NNs have also been used to reproduce the free-energy surface of molecules [47]. Deep convolutional neural networks (CNN) [48] have become increasingly popular due to its performance in machine vision, a property that has been exploited by us and others to apply it on structural biology by treating proteins as 3D images. CNNs have been used for ligand binding site detection [49•], ligand pose prediction [50], ligand active/inactive classification [51], ligand binding affinity prediction [52•] and protein design [53]. Also, the DeepChem software [54•] and the MoleculeNet challenge [55] provide multiple featurization algorithms and access to relevant QSAR prediction datasets.