Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Supercomputing in the biological sciences: Toward Zettascale and Yottascale simulations

Karissa Sanbonmatsu[1,2]

Molecular simulations of biological systems tend to be significantly more compute-intensive than those in materials science and astrophysics, due to important contributions of long-range electrostatic forces and large numbers of time steps (>1E9) required. Simulations of biomolecular complexes of microseconds to milliseconds are considered state-of-the-art today. However, these time scales are miniscule in comparison to physiological time scales relevant to molecular machine activity, drug action, and elongation cycles for protein synthesis, RNA synthesis, and DNA synthesis (seconds to days). While an exascale supercomputer has simulated an entire virus for nanoseconds, this supercomputer would need to be 10 billion times faster to simulate that virus for 3 hours of physiological time, demonstrating the insatiable need for computing power. With growing interest in computational drug design from the pharmaceutical sector, the biological sciences are positioned to be an industry driver in computing.

Addresses
[1] Los Alamos National Laboratory, United States
[2] New Mexico Consortium, New Mexico

Corresponding author: Sanbonmatsu, Karissa (kys@lanl.gov)

Current Opinion in Structural Biology 2024, **88**:102889

This review comes from a themed issue on **Theory and Simulation/Computational Methods (2024)**

Edited by **Paolo Carloni** and **Karissa Y. Sanbonmatsu**

For complete overview of the section, please refer the article collection - Theory and Simulation/Computational Methods (2024)

Available online 19 August 2024

https://doi.org/10.1016/j.sbi.2024.102889

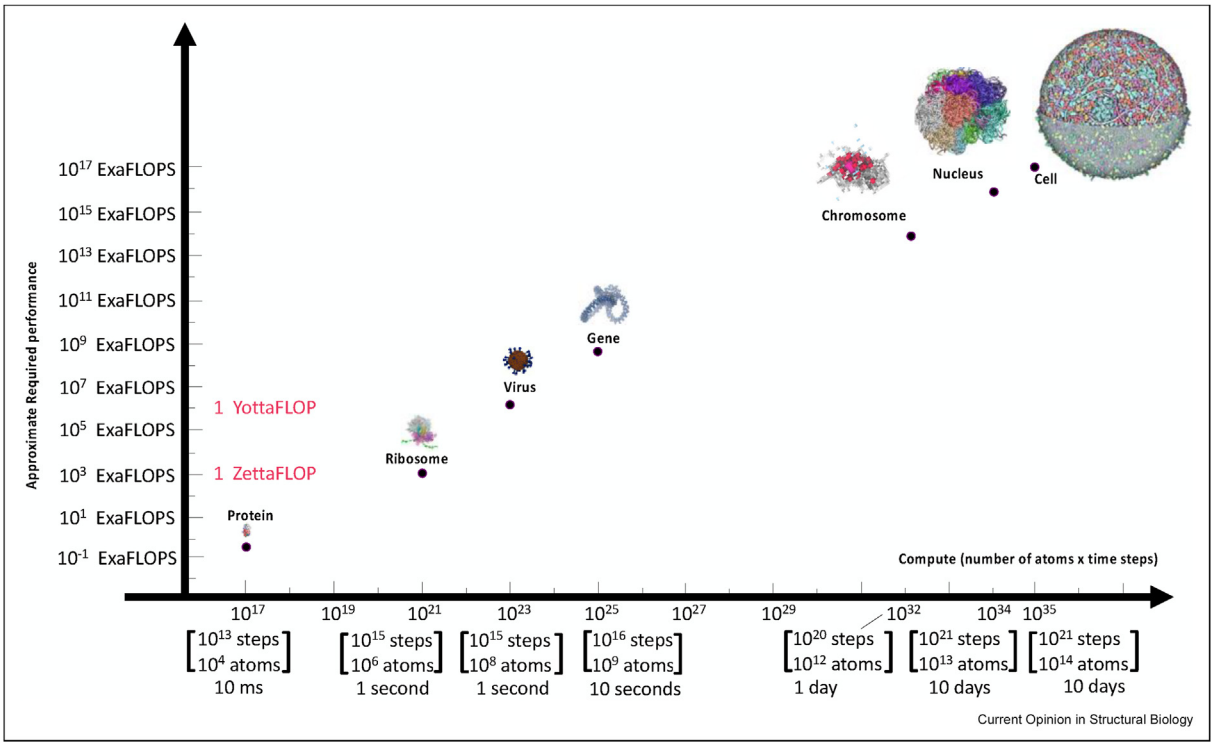
Introduction

Biological systems present some of the most demanding, compute-intensive high-performance computing applications. For example, mechanistic understanding of viruses and molecular machines, as well as computational drug-design efforts, often requires calculations of free energies. Free-energy calculations require enormous amounts of conformational sampling to achieve equilibrium thermodynamics [1]. Even modest amounts of sampling (e.g. 1 ms of physiological time) require 10^{12} time steps. Due to the electrostatic charges present, long-range electrostatic forces play important roles (e.g. every nucleotide of DNA and RNA is charged) [2]. Thus, biological simulations are often much more intensive than material science applications, which typically do not include long-range electrostatic interactions. While cosmological calculations have long-range gravitational forces, biological simulations require orders of magnitude more time steps. Additional factors of complexity in biological systems, such as the fact that many processes are far from equilibrium and that chemical reactions can be critical (requiring quantum mechanical calculations), further complicate these systems. If we neglect chemical reactions and nonequilibrium effects, we estimate that simulating 1 s of physiological time for the human genome (in the case of 23 chromosomes) would require at least 10^4 yottaFLOPs (1 YF = 10^{24} FLOPs) or 10^{10} exaFLOPs. While these calculations are far beyond the scope of current platforms, they provide a roadmap for the way forward in biomolecular simulation and demonstrate that the biological sciences have an insatiable demand for computing and will continue to do so for the coming decades (Figure 1). In Figure 1, system sizes are taken from actual all-atom simulations (protein, ribosome, virus, gene) or from the length scale of coarse grained simulations (chromosome, nucleus, cell). FLOP values are extrapolated from actual values (protein) or from high performance computing (HPC) platforms on which molecular dynamics simulations were performed (ribosome, gene). With growing interest in computational drug design from large pharmaceutical companies, it is reasonable to suggest that biomolecular computation may drive the computing industry in the future.

As we do not currently have 10 billion exascale supercomputers at our disposal, many have adopted an integrated biology approach, using constraints from a wide variety of experimental data to dramatically expand the accessible time scales. This approach is not ab initio and limits the scope of predictions that can be made. However, as the accuracy of free-energy calculations is directly related to the amount of conformational sampling achieved, which we know is inadequate by many orders of magnitude today, today's ab initio approaches

Given the role as Guest Editor, Karissa Y. Sanbonmatsu had no involvement in the peer review of the article and has no access to information regarding its peer-review. Full responsibility for the editorial process of this article was delegated to Paolo Carloni.

Figure 1



Approximate required performance vs. compute, in units of time steps multiplied by the number of atoms. Simulating 10 days of physiological time of a cell in atomistic detail in explicit solvent would take a 10 ExaFLOPS supercomputer approximately 10^{16} years, or 10^{6} times the lifetime of our Sun.

are also severely limited due to limitations in compute power, even using the world's most powerful supercomputers, as detailed earlier. Thus, in many situations, an integrated biology approach, which incorporates experimental data as constraints, represents the most accurate and descriptive approach available. A variety of computational tools have been used to characterize large biological systems, including all-atom explicit solvent simulations, long-time-scale simulations, enhanced sampling simulations, and coarse-grained and reduced-model simulations. In the following, we review efforts in each of these areas. This review is by no means comprehensive and due to limited space, does not mention many important studies in the field. The review uses a few illustrative examples to highlight methods and topics that have emerged in the push toward larger and larger systems and simulations.

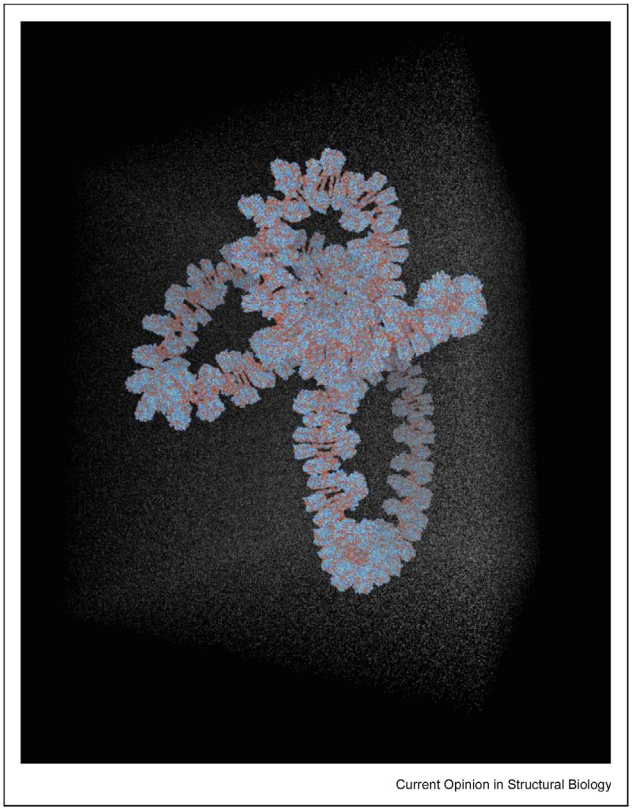
Structural biology studies are the basis of atomistic molecular dynamics simulations of macromolecular complexes

Molecular dynamics simulations often use high-resolution structures of large biomolecular complexes generated by X-ray crystallography or cryogenic electron microscopy (cryo-EM). There are too many examples to name in this review. A few specific examples of such systems include virus particles, the ribosome, chromatin complexes, and membrane-receptor complexes, such as RyR1, glutamate, and STRA6 [3–9]. While crystallographic structures were predominantly used in molecular dynamics simulations since their inception into the 2000s, the resolution revolution in cryo-EM made a large number of structures of large macromolecular complexes available in the 2010s to the present day.

Equilibrium-explicit-solvent molecular dynamics simulations of large biomolecular systems

Equilibrium molecular dynamics simulations in explicit solvent produce information about short-time-scale fluctuations (hundreds of nanoseconds to milliseconds), hydrogen-bond networks, solvent structure, ion distributions, electrostatic potentials, and ligand interactions. Schulten et al. and Grubmuller et al. were pioneers of large-scale biomolecular simulations, performing simulations of over 100,000 atoms of fibronectin and F1-ATP synthase [10,11]. Since these early studies, successive landmarks were established by simulations of over a million atoms of the ribosome, flagellum, and satellite tobacco mosaic virus [12–16]. Most recently, the first simulations over a billion atoms have been performed of a chromatin system (GATA4 gene, 427 nucleosomes), the SARS-CoV-2 particle, and large swaths of cytoplasm, including hundreds of proteins [17–19] (Figure 2). As

Figure 2



Explicit solvent simulation of GATA4 gene including 83 kb and 427 nucleosomes requires 1 billion atoms.

discussed in the following, the challenges involved in simulating large systems are compounded by the fact that larger systems require more sampling than do smaller systems for equilibration.

Long-time-scale simulations

Long-time-scale simulations represent another important frontier of molecular simulation. Here, simulations of single proteins are performed for longer time scales (milliseconds) to characterize fluctuations on these time scales and help validate short-time-scale (hundreds of nanoseconds to microseconds) simulations. The first microsecond simulation was performed by Duan et al. [20], in which a single protein was simulated. Later, with the aid of the Anton platform, millisecond simulations were performed [21], where a single protein was simulated. More recently, simulations exceeding milliseconds have been performed. As the simulation time scale increases by significant leaps, different behaviors are observed, underscoring the challenges involved in assessing convergence and establishing equilibrium. For example, suppose one performs a simulation of 10 ms, but the relevant physiological timescale is 10 min. One could then perform 100 more 10 ms simulations, and obtaining consistency between the 100 simulations, conclude that the simulations have converged according to various metrics. However, if 10 years later, a new supercomputer is developed capable of performing simulations of 10 s, these new simulations may produce results that conflict with the 10-ms simulations. This pattern of declaring convergence could proceed until 100 simulations of 10 min can be performed, decades in the future. In short, with limited compute power, and correspondingly limited conformational sampling and time scales, it is difficult to assess ergodicity and whether one has truly achieved thermodynamic equilibrium.

Enhanced sampling methods

A key area of development in molecular simulation has been enhanced sampling algorithms. Typically, conformational sampling is artificially enhanced in very specific ways, such as careful heating by small amounts of heat in specific circumstances or by lowering force-field interaction energies of very specific interactions at specific times [22]. These artificial enhancements are carefully quantified, followed, and tracked. At the end of the study, the additions of heat, temperature, or energy are often subtracted, yielding equilibrium free-energy landscapes [1].

Some of the earliest work in enhanced sampling originated in the spin-glass community in condensed-matter physics, with seminal papers from Hansmann and Okamoto, who applied these techniques to applications in protein folding [23–28]. Hansmann, Okamoto and Sugita pioneered replica-exchange simulations, a technique widely used today, where many replicate simulations are performed at slightly different temperatures, and small temperature jumps are given to each replicate by exchanging temperatures with its neighbor in the temperature space [22,23]. This gives a ~25-fold boost in conformational sampling and Boltzmann equilibrium sampling and easily folds small proteins and peptides [29]. The method has also been applied to RNA systems such as ribosome–antibiotic complexes, single-stranded RNA, and riboswitches [30–32]. Many variants of this algorithm have been developed that do not melt the system and are more suitable for large biomolecular complexes. These include Replica exchange umbrella sampling (REUS), Hamilton replica, resolution replica, protonation replica, and sparse replica [33–38]. In one case, a multiscale study was produced with a smart-resolution replica-exchange technique [37]. More recently, weighted histogram methods and artificial intelligence (AI) methods have been used [39,40].

Metadynamics is another popular and useful technique, made convenient in the PLUMED framework [41–43]. Here, basins in the energy landscape are gradually filled by Gaussians, which are carefully tracked. In the end, the Gaussians are subtracted, giving the free-energy landscape. A key aspect of many of these methods is

determining the most useful collective variable or order parameter to describe a conformational change of interest that best characterizes a transition. Another emerging area is that of enhanced-sampling kinetic approaches [44—46].

While enormous resources in the pharmaceutical community have been devoted to accurately characterizing the enthalpic component of the free energy of drug binding, much less attention has been paid to the entropic component, which critically depends on obtaining tremendous amounts of conformational sampling [47,48]. The free energy consists of the difference between the enthalpic and entropic terms. These terms are often large and opposite in sign, nearly canceling. Thus, to obtain an accurate free energy of binding, one must absolutely obtain an accurate entropic component of the free energy. Conformational sampling and accurate entropy estimation remains a key challenge in drug design today.

While significantly less expensive than conventional molecular dynamics simulations, these methods are nonetheless quite expensive, complex, and difficult to implement for very large systems. One of the open areas in large-scale simulation is to apply these very useful methods to very large systems.

Coarse-grained simulations

Molecular dynamics simulations of biomolecules considered 'more accurate' are often those that are considered the 'most realistic' and include the most effects possible, such as atomistic detail, explicit water, electrostatics, polarization, explicit ions, the buffer closest to the biochemical experiment, and the most physiologically relevant complex. However, these are sometimes not the most useful simulations as they are (i) very limited in their access to relevant time scales and (ii) have corresponding inaccurate free-energy estimates due to the lack of conformational sampling, especially for very large systems (equilibrium conformational sampling is required for the entropic component of the free energy). The current state-of-the-art for explicit solvent atomistic molecular dynamics simulations for relatively large systems grossly undersamples their dynamics, even using the fastest supercomputers available. For a million-atom system such as the ribosome (2—5 million atoms, depending on the specific ribosome complex), the state-of-the-art is at best 1 ms of physiological time. As relevant rates are 100—800 ms, at least 1 s of aggregate sampling is required to have a minimum number of replicates for statistics. As shown in Figure 1, this would require a 1 ZettaFLOP supercomputer (1000 ExaFLOPS). Since the fastest machine available is $\sim$ 1—2 exaFLOPS, even a one-millisecond simulation of the ribosome would be 1000-fold undersampled, yielding grossly inaccurate free energies. Even assuming a 25-fold boost for enhanced sampling algorithms, the system would still be 40-fold undersampled. This problem is compounded when simulating even larger systems. Not only do these have correspondingly more atoms but they require much longer time scales to equilibrate (e.g. the GATA4 and COVID billion-atom simulations sample only a few nanoseconds but likely require much longer than 1 s for equilibration).

These examples underscore the advantages of using coarse-grained and reduced-model simulations (protein shown in Figure 1 was simulated using reduced-model simulations to achieve agreement with single-molecule Forster resonance energy transfer (FRET) [49]). For example, native-contact-based simulations of tRNA accommodation into the ribosome during transfer RNA (tRNA) selection are able to simulate events with rates slower than 100 ms with excellent statistics (hundreds of trajectories to one thousand trajectories of accommodation) [50,51]. While these simpler models lack the complexities included in explicit-solvent atomistic molecular dynamics simulations, they have vastly superior sampling and assessments of conformational sampling and entropy. Coarse-grained strategies follow a fundamental tenet of theoretical physics: the best practice is to understand zeroth-order effects first, and only after this, move on to more complicated higher-order effects. A good physicist must know how to neglect what is unimportant and focus on what is important by making the most effective approximations and simplifications. One should be reminded that Newton obtained quite accurate approximations of the Earth's orbit around the Sun by neglecting the Earth's ocean, atmosphere, asymmetry, all life on Earth, and approximating the Earth as a single pseudo particle, a very coarse-grained model indeed [52—54]. Newton's highly simplistic model (a very coarse-grained model with simplistic force field, in molecular dynamics simulation parlance) could predict the orbital periods and relative distances of planets from the Sun to within a fraction of a percent and accurately predicted eclipses and the return of Halley's Comet. Once these fundamental zeroth-order effects are understood, perturbation theory can be used to model higher-order effects, such as tidal effects of the Moon and variations in orbits. These examples highlight the approach of choosing the most appropriate level of coarse-graining for the task at hand and for the particular question of interest. In biological and biomedical applications, time scales range from femtoseconds to weeks, and length scales range from angstroms to meters, spanning more than ten orders of magnitude. To deal with this complexity, an appropriate level of coarse-graining must be chosen.

In addition to saving vast amounts of compute time, coarse-grained simulations have the advantage of being very amenable to the incorporation of experimental data as they can access much longer time scales and much larger systems that are more relevant to experimental studies. This makes these approaches ideal for integrated biology studies. For example, many conformational changes of the ribosome occur with rates of hundreds of milliseconds. A

proper study with excellent statistics might require a hundred of these hundred-millisecond trajectories (e.g. as in single-molecule FRET studies), summing to aggregate sampling of 10 s, at least four orders of magnitude longer than a millisecond simulation of the ribosome. However, with proper coarse graining, such simulations are possible [55—57]. These simulations use force fields that are 'less accurate' in the conventional sense, with pseudo particles that represent several or many atoms, corrections to compensate for lack of explicit water molecules, and less accurate electrostatics. However, these same coarse-grained simulations are, in a certain specific sense, more accurate than all-atom explicit-solvent simulations, in that the coarse-grained simulations obtain vastly more conformational sampling than explicit solvent simulations and therefore yield more accurate entropic components of the free energy, which critically depends on the amount of conformational space sampled. These methods obtain much more conformational sampling relative to explicit-solvent approaches, especially for large systems [58—60]. Several coarse-grained simulations of virus particles have been performed, including the SARS-CoV-2 virion [61—65]. Coarse-grained approaches to chromatin have also explored very large systems with various levels of coarse-graining, describing nucleosomes and linker DNA with various levels of detail [66—72].

Most importantly, many coarse-grain and reduced-model simulations are 'top-down' and use experimental data as constraints, guaranteeing consistency with experimental data from the outset. The approach is often used in integrated structural biology studies, where molecular simulations bring together disparate types of experimental data into a coherent study of mechanism [73—76]. This approach is often more of an interpolation between experimental data points than an extrapolation into unknown parameter regimes. While a naïve observer might ask, "what can one learn from such a simulation where the 'answer' is hardwired into the simulation?" such interpolative simulations do not suffer from concerns about experimental validation (as the validation is designed into the simulation from the outset) and yield interesting and useful insights related to transitions between states, such as energy landscapes, transition pathways, and transition intermediates, as well as information about mechanism such as order of events. The approach also allows the testing of different possible mechanisms while maintaining consistency with experimental data.

A key area of "top-down" simulations that use experimental data from the outset are called structure-based simulations, also known native-contact models or as Go-like models, after the pioneer of the method, Professor Nobuhiro Go [77,78]. Here, the potential is based on the native contacts present in a known structure. These potentials were used by Onuchic et al. to work out the protein-folding-funnel free-energy landscape [79,80], where simulations beginning with a completed disordered protein could be rapidly folded into the native folded state. These approaches have been modernized in atomistic, off-lattice versions with electrostatics and ions. They have also been repurposed to study large-scale conformational changes of macromolecular complexes, using the final state in a conformational change as the native state and beginning with a known conformation preceding the conformational change, such as a PRE to POST conformational change [50,51,81]. In the case of the ribosome, structure-based approaches have been used to investigate accommodation of tRNA into the ribosome during tRNA selection for cognate and near-cognate tRNAs, integrating an X-ray crystallography, cryo-EM, single-molecule FRET, and single-molecule FRET data [50,51]. There have been several top-down, integrative structural biologies related to COVID-19 [61,82].

In the direction of very coarse-grained simulation, a new area of simulation has emerged, inspired by simulations in polymer physics and the invention of a high-throughput sequencing method called Hi-C, which experimentally obtains contact maps of chromosomes by crosslinking, digestion, and sequencing. Here, beads-on-springs polymer models are used to approximate chromosomes, with a single bead typically representing 100 kb of DNA (in explicit solvent, simulating a single bead would require approximately one billion atoms). Mirny et al. performed ab initio block copolymer simulations [83,84]. Onuchic, Wolynes, and Zhang applied maximum-entropy techniques to produce three-dimensional (3D) models and gain insight [85,86].

Lappala developed a top-down method (4DHiC) to directly incorporate large amounts of experimental data to directly visualize the Hi-C data in 3D, revealing compartment mixing during various states of X-inactivation (shown in Figure 1 as 'Chromosome') [87]. The nucleus model shown in Figure 1 was produced by Ankush Singhal using a similar method.

Toward whole-cell simulations

A final area of coarse graining includes efforts to simulate an entire cell (image shown in Figure 1 as 'Cell' published previously [88]). Starting in the 1990s with simple zero-dimension models based on reaction kinetics, the field has evolved to 3D models of the cell with time evolution [89,90]. Several efforts are making great progress in this area [88,91—93].

Conclusions

Molecular simulations of biological systems tend to be significantly more compute-intensive than those in material science and astrophysics, due to important contributions of long-range electrostatic forces and large numbers of time steps ($>10^9$) required for microsecond simulations of proteins, DNA, RNA, and macromolecular complexes, including viruses, chromatin, ribosomes, and

other molecular machines. Although simulation durations ranging from microseconds to milliseconds are often considered the state-of-the-art for explicit-solvent molecular dynamics simulations today; these time scales are miniscule in comparison to physiological time scales relevant to molecular machine activity, drug action, and elongation cycles for protein synthesis, RNA synthesis, and DNA synthesis, ranging from hundreds of milliseconds to days. AI is being implemented in molecular dynamics simulations to improve sampling of biologically relevant events, enhancing collective variable selection or optimizing force-field potentials. However, for AI to yield necessary gains of many orders of magnitude, we first need to provide such enormously long simulation trajectories of large systems to train on, which will likely not be accessible for many decades. While the largest and fastest supercomputers on Earth can simulate a virus for nanoseconds on an exascale supercomputer, these platforms are at least twelve orders of magnitude away from performing a single simulation of a virus for 10 s, not to mention the requirement for statistical replicates. These issues are compounded by the fact that larger systems require much more conformational sampling to reach equilibrium, making the task of scaling up to full chromosomes, nuclei, and cells formidable. With growing interest in computational drug design from the pharmaceutical sector, the biological sciences are positioned to be an industry driver in computing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data were used for the research described in the article.

Acknowledgements

The author is grateful to Ankush Singhal for the image of the nucleus and to Emvia Calixte and Tyler Reddy for the image of the SARS-CoV-2 virion coarse-grained simulation. The author is grateful to generous support from Los Alamos National Laboratory Institutional Computing for allocations on LANL Q, LANL RoadRunner, LANL Trinity, LANL Chicoma, and LANL Venado. The author thanks the generous support from NIH (RO1-GM110310 and R01-GM072686), NSF (MCB 2337393), LANL LDRD and and by the U.S. Department of Energy, Office of Science, through the Biological and Environmental Research (BER) and the Advanced Scientific Computing Research (ASCR) programs under contract number 89233218CNA000001 to Los Alamos National Laboratory (Triad National Security, LLC).

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Rosta E, Hummer G: **Free energies from dynamic weighted histogram analysis using unbiased Markov state model**. *J Chem Theor Comput* 2015, **11**:276–285.

2. * Braun E, Gilmer J, Mayes HB, Mobley DL, Monroe JI, Prasad S, Zuckerman DM: **Best practices for foundations in molecular simulations [article v1.0]**. *Living J Comput Mol Sci* 2019, **1**.
This is an excellent practical guide to molecular dynamics simulations. The article starts with the basics and delves into the finer points of setting up and performing molecular dynamics simulations.

3. Frank J, Agrawal RK: **A ratchet-like inter-subunit reorganization of the ribosome during translocation**. *Nature* 2000, **406**: 318–322.

4. Pokidysheva E, Zhang Y, Battisti AJ, Bator-Kelly CM, Chipman PR, Xiao C, Gregorio GG, Hendrickson WA, Kuhn RJ, Rossmann MG: **Cryo-EM reconstruction of dengue virus in complex with the carbohydrate recognition domain of DC-SIGN**. *Cell* 2006, **124**:485–493.

5. des Georges A, Clarke OB, Zalk R, Yuan Q, Condon KJ, Grassucci RA, Hendrickson WA, Marks AR, Frank J: **Structural basis for gating and activation of RyR1**. *Cell* 2016, **167**: 145–157 e117.

6. Chen Y, Clarke OB, Kim J, Stowe S, Kim YK, Assur Z, Cavalier M, Godoy-Ruiz R, von Alpen DC, Manzini C, *et al.*: **Structure of the STRA6 receptor for retinol uptake**. *Science* 2016, **353**.

7. Frank J, Radermacher M, Wagenknecht T, Verschoor A: **Studying ribosome structure by electron microscopy and computer-image processing**. *Methods Enzymol* 1988, **164**:3–35.

8. ** Bhattacharjee S, Feng X, Maji S, Dadhwal P, Zhang Z, Brown ZP, Frank J: **Time resolution in cryo-EM using a PDMS-based microfluidic chip assembly and its application to the study of HflX-mediated ribosome recycling**. *Cell* 2024, **187**:782–796 e723.
The authors present a ground-breaking method for time-resolved cryo-EM that leverages microfluidics. The new device operates from 10 to 1000 ms and follows reactions with spatial resolution of 3 Å, presenting a template for future research in cryo-EM.

9. Twomey EC, Yelshanskaya MV, Grassucci RA, Frank J, Sobolevsky AI: **Elucidation of AMPA receptor-stargazin complexes by cryo-electron microscopy**. *Science* 2016, **353**:83–86.

10. Gao M, Craig D, Vogel V, Schulten K: **Identifying unfolding intermediates of FN-III(10) by steered molecular dynamics**. *J Mol Biol* 2002, **323**:939–950.

11. Bockmann RA, Grubmuller H: **Nanoseconds molecular dynamics simulation of primary mechanical energy transfer steps in F1-ATP synthase**. *Nat Struct Biol* 2002, **9**:198–202.

12. Sanbonmatsu KY, Joseph S, Tung CS: **Simulating movement of tRNA into the ribosome during decoding**. *Proc Natl Acad Sci U S A* 2005, **102**:15854–15859.

13. Villa E, Sengupta J, Trabuco LG, LeBarron J, Baxter WT, Shaikh TR, Grassucci RA, Nissen P, Ehrenberg M, Schulten K, Frank J: **Ribosome-induced changes in elongation factor Tu conformation control GTP hydrolysis**. *Proc Natl Acad Sci U S A* 2009, **106**:1063–1068.

14. Kitao A, Yonekura K, Maki-Yonekura S, Samatey FA, Imada K, Namba K, Go N: **Switch interactions control energy frustration and multiple flagellar filament structures**. *Proc Natl Acad Sci USA* 2006, **103**:4894–4899.

15. Bock LV, Blau C, Schroder GF, Davydov II, Fischer N, Stark H, Rodnina MV, Vaiana AC, Grubmuller H: **Energy barriers and driving forces in tRNA translocation through the ribosome**. *Nat Struct Mol Biol* 2013, **20**:1390–1396.

16. Freddolino PL, Arkhipov AS, Larson SB, McPherson A, Schulten K: **Molecular dynamics simulations of the complete satellite tobacco mosaic virus**. *Structure* 2006, **14**:437–449.

17. * Jung J, Nishima W, Daniels M, Bascom G, Kobayashi C, Adedoyin A, Wall M, Lappala A, Phillips D, Fischer W, *et al.*: **Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations**. *J Comput Chem* 2019, **40**:1919–1930.
The authors present the first billion atom simulation, performing explicit solvent simulations of the GATA4 gene, including 427 nucleosomes. The unique parallelization scheme of GENESIS allowed scaling to ~500,000 cores.

18. ** Dommer A, Casalino L, Kearns F, Rosenfeld M, Wauer N, Ahn SH, Russo J, Oliveira S, Morris C, Bogetti A, *et al.*:

#COVIDisAirborne: AI-enabled multiscale computational microscopy of delta SARS-CoV-2 in a respiratory aerosol. *Int J High Perform Comput Appl* 2023, **37**:28–44.
The authors perform the largest simulation of a virus to date, at over one billion atoms. In this integrated data-driven study, the authors explore the dynamics of aerosolized viruses.

19. Jung J, Kobayashi C, Kasahara K, Tan C, Kuroda A, Minami K, Ishiduki S, Nishiki T, Inoue H, Ishikawa Y, *et al.*: **New parallel computing algorithm of molecular dynamics for extremely huge scale biological systems**. *J Comput Chem* 2021, **42**:231–241.

20. Duan Y, Kollman PA: **Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution**. *Science* 1998, **282**:740–744.

21. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W: Atomic-level characterization of the structural dynamics of proteins. Science 330:341-346.

22. Sugita Y, Okamoto Y: **Replica-exchange molecular dynamics method for protein folding**. *Chem Phys Lett* 1999, **314**: 141–151.

23. Hansmann U: **Parallel tempering algorithm for conformational studies of biological molecules**. *Chem Phys Lett* 1997, **281**: 140–150.

24. Hansmann U: **Generalized ensembles: a new way of simulating proteins**. *PHYSICA A 1st Asia-Pacific and the 4th Taipei International Workshop/Symposium on Statistical Physics - New Directions in Statistical Physics (StatPhys-Taipei-1997)* 1998, **254**:15–23.

25. Hansmann U: **Effective way for determination of multicanonical weights**. *Phys Rev* 1997, **56**:6200–6203.

26. Hansmann U: **Protein-folding simulations in generalized ensembles**. *INTERNATIONAL JOURNAL OF QUANTUM CHEMISTRY Symposium on Theory and Computations in Molecular and Materials Sciences Biology and Pharmacology* 2002, **90**: 1515–1521.

27. Hansmann U, Okamoto Y: **New Monte Carlo algorithms for protein folding**. *Curr Opin Struct Biol* 1999, **9**:177–183.

28. Okamoto Y, Hansmann U: **Thermodynamics of helix-coil transitions studied by multicanonical algorithms**. *J Phys Chem* 1995, **99**:11276–11287.

29. Sanbonmatsu KY, Garcia AE: **Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics**. *Proteins* 2002, **46**:225–234.

30. Eichhorn CD, Feng J, Suddala KC, Walter NG, Brooks 3rd CL, Al-Hashimi HM: **Unraveling the structural complexity in a single-stranded RNA tail: implications for efficient ligand binding in the prequeuosine riboswitch**. *Nucleic Acids Res* 2012, **40**: 1345–1355.

31. Kuhrova P, Best RB, Bottaro S, Bussi G, Sponer J, Otyepka M, Banas P: **Computer folding of RNA tetraloops: identification of key force field deficiencies**. *J Chem Theor Comput* 2016, **12**: 4534–4548.

32. Vaiana AC, Sanbonmatsu KY: **Stochastic gating and drug-ribosome interactions**. *J Mol Biol* 2009, **386**:648–661.

33. Mitsutake A, Sugita Y, Okamoto Y: **Generalized-ensemble algorithms for molecular simulations of biopolymers**. *Biopolymers* 2001, **60**:96–123.

34. Yoda T, Sugita Y, Okamoto Y: **Protein folding simulations by generalized-ensemble algorithms**. *Adv Exp Med Biol* 2014, **805**:1–27.

35. Kokubo H, Tanaka T, Okamoto Y: **Ab initio prediction of protein-ligand binding structures by replica-exchange umbrella sampling simulations**. *J Comput Chem* 2011, **32**:2810–2821.

36. Itoh SG, Damjanovic A, Brooks BR: **pH replica-exchange method based on discrete protonation states**. *Proteins* 2011, **79**:3420–3436.

37. Liu P, Voth GA: **Smart resolution replica exchange: an efficient algorithm for exploring complex energy landscapes**. *J Chem Phys* 2007, **126**, 045106.

38. Roe DR, Bergonzo C, Cheatham 3rd TE: **Evaluation of enhanced sampling provided by accelerated molecular dynamics with Hamiltonian replica exchange methods**. *J Phys Chem B* 2014, **118**:3543–3552.

39. Ojha AA, Thakur S, Ahn SH, Amaro RE: **DeepWEST: deep learning of kinetic models with the weighted ensemble simulation toolkit for enhanced sampling**. *J Chem Theor Comput* 2023, **19**:1342–1359.

40. Galvelis R, Sugita Y: **Neural network and nearest neighbor algorithms for enhancing sampling of molecular dynamics**. *J Chem Theor Comput* 2017, **13**:2489–2500.

41. Bussi G, Tribello GA: **Analyzing and biasing simulations with PLUMED**. *Methods Mol Biol* 2019, **2022**:529–578.

42. Branduardi D, Bussi G, Parrinello M: **Metadynamics with adaptive Gaussians**. *J Chem Theor Comput* 2012, **8**:2247–2254.

43. Prajapati JD, Onuchic JN, Sanbonmatsu KY: **Exploring the energy landscape of riboswitches using collective variables based on tertiary contacts**. *J Mol Biol* 2022, **434**, 167788.

44. Hummer G, Szabo A: **Optimal dimensionality reduction of multistate kinetic and markov-state models**. *J Phys Chem B* 2015, **119**:9029–9037.

45. Weber JK, Pande VS: **Characterization and rapid sampling of protein folding Markov state model topologies**. *J Chem Theor Comput* 2011, **7**:3405–3411.

46. Prinz JH, Chodera JD, Pande VS, Swope WC, Smith JC, Noe F: **Optimal use of data in parallel tempering simulations for the construction of discrete-state Markov models of biomolecular dynamics**. *J Chem Phys* 2011, **134**, 244108.

47. Sinko W, Lindert S, McCammon JA: **Accounting for receptor flexibility and enhanced sampling methods in computer-aided drug design**. *Chem Biol Drug Des* 2013, **81**:41–49.

48. Borrego-Sanchez A, Debnath J, Parrinello M: **Kinetics of drug release from clay using enhanced sampling methods**. *Pharmaceutics* 2022, **14**.

49. Girodat D, Pati AK, Terry DS, Blanchard SC, Sanbonmatsu KY: **Quantitative comparison between sub-millisecond time resolution single-molecule FRET measurements and 10-second molecular simulations of a biosensor protein**. *PLoS Comput Biol* 2020, **16**, e1008293.

50. Whitford PC, Geggier P, Altman RB, Blanchard SC, Onuchic JN, Sanbonmatsu KY: **Accommodation of aminoacyl-tRNA into the ribosome involves reversible excursions along multiple pathways**. *RNA* 2010, **16**:1196–1204.

51. * Girodat D, Wieden HJ, Blanchard SC, Sanbonmatsu KY: **Geometric alignment of aminoacyl-tRNA relative to catalytic centers of the ribosome underpins accurate mRNA decoding**. *Nat Commun* 2023, **14**:5582.
The authors use reduced model force fields (atomistic native-contact based potentials) and performs hundreds of simulations of spontaneous tRNA movement into the ribosome for cognate and near-cognate tRNA. They find that the physical size of both tRNA and ribosome are key determinants of the tRNA selection fidelity mechanism.

52. Newton I: Philosophiae naturalis principia mathematica. London: Royal Society; 1687.

53. Newton I, Cohen IB, Whitman AM: *The Principia : mathematical principles of natural philosophy*. Berkeley: University of California Press; 1999.

54. Newton I: *Newton's philosophy of nature: selections from his writings*. New York,: Hafner Pub. Co.; 1953.

55. Bradley R, Radhakrishnan R: **Coarse-grained models for protein-cell membrane interactions**. *Polymers* 2013, **5**: 890–936.

56. Zhang Z, Voth GA: **Coarse-grained representations of large biomolecular complexes from low-resolution structural data**. *J Chem Theor Comput* 2010, **6**:2990–3002.

57. Portillo-Ledesma S, Li Z, Schlick T: **Genome modeling: from chromatin fibers to genes**. *Curr Opin Struct Biol* 2023, **78**, 102506.

58. Zheng W: **Coarse-grained modeling of the structural states and transition underlying the powerstroke of dynein motor domain**. *J Chem Phys* 2012, **136**, 155103.

59. Zhang Z, Sanbonmatsu KY, Voth GA: **Key intermolecular interactions in the E. coli 70S ribosome revealed by coarse-grained analysis**. *J Am Chem Soc* 2011, **133**:16828–16838.

60. Yoo J, Cui Q: **Curvature generation and pressure profile modulation in membrane by lysolipids: insights from coarse-grained simulations**. *Biophys J* 2009, **97**:2267–2276.

61. * Yu A, Pak AJ, He P, Monje-Galvan V, Casalino L, Gaieb Z, Dommer AC, Amaro RE, Voth GA: **A multiscale coarse-grained model of the SARS-CoV-2 virion**. *Biophys J* 2021, **120**: 1097–1104.

The authors present an excellent coarse-grained model of the SARS-CoV-2 virion, integrating crystallography and cryo-EM data.

62. Reddy T, Shorthouse D, Parton DL, Jefferys E, Fowler PW, Chavent M, Baaden M, Sansom MSP: **Nothing to sneeze at: a dynamic and integrative computational model of an influenza A virion**. *Structure* 2015, **23**:584–597.

63. Reddy T, Sansom MS: **The role of the membrane in the structure and biophysical robustness of the dengue virion envelope**. *Structure* 2016, **24**:375–382.

64. Hagan MF, Zandi R: **Recent advances in coarse-grained modeling of virus assembly**. *Curr Opin Virol* 2016, **18**:36–43.

65. Arkhipov A, Freddolino PL, Schulten K: **Stability and dynamics of virus capsids described by coarse-grained modeling**. *Structure* 2006, **14**:1767–1777.

66. Dans PD, Walther J, Gomez H, Orozco M: **Multiscale simulation of DNA**. *Curr Opin Struct Biol* 2016, **37**:29–45.

67. Gavin D: **Bascom CGMaTS: mesoscale modeling reveals formation of an epigenetically driven HOXC gene hub**. *Proc Natl Acad Sci U S A* 2019, **116**:4955–4962.

68. Bascom GD, Schlick T: **Chromatin fiber folding directed by cooperative histone tail acetylation and linker histone binding**. *Biophys J* 2018, **114**:2376–2385.

69. Perisic O, Schlick T: **Dependence of the linker histone and chromatin condensation on the nucleosome environment**. *J Phys Chem B* 2017, **121**:7823–7832.

70. Huertas J, Woods EJ, Collepardo-Guevara R: **Multiscale modelling of chromatin organisation: resolving nucleosomes at near-atomistic resolution inside genes**. *Curr Opin Cell Biol* 2022, **75**, 102067.

71. Farr SE, Woods EJ, Joseph JA, Garaizar A, Collepardo-Guevara R: **Nucleosome plasticity is a critical element of chromatin liquid-liquid phase separation and multivalent nucleosome interactions**. *Nat Commun* 2021, **12**:2883.

72. Bishop TC, Zhmudsky OO: **Mechanical model of the nucleosome and chromatin**. *J Biomol Struct Dyn* 2002, **19**:877–887.

73. Yuan S, Xia L, Wang C, Wu F, Zhang B, Pan C, Fan Z, Lei X, Stevens RC, Sali A, *et al.*: **Conformational dynamics of the activated GLP-1 receptor-G(s) complex revealed by cross-linking mass spectrometry and integrative structure modeling**. *ACS Cent Sci* 2023, **9**:992–1007.

74. * Hancock M, Peulen TO, Webb B, Poon B, Fraser JS, Adams P, Sali A: **Integration of software tools for integrative modeling of biomolecular systems**. *J Struct Biol* 2022, **214**, 107841.

The authors present Integrative Modeling Platform, designed to help users perform integrated structural biology models. They include an excellent discussion about software interoperability and its role in integrative structure biology.

75. Sali A: **From integrative structural biology to cell biology**. *J Biol Chem* 2021, **296**, 100743.

76. Rout MP, Sali A: **Principles for integrative structural biology studies**. *Cell* 2019, **177**:1384–1403.

77. Taketomi H, Ueda Y, Go N: **Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions**. *Int J Pept Protein Res* 1975, **7**: 445–459.

78. * Takada S: **Go model revisited**. *Biophys Physicobiol* 2019, **16**: 248–255.

This is an excellent article about Go models and the latest variations on this theme.

79. Leopold PE, Montal M, Onuchic JN: **Protein folding funnels: a kinetic approach to the sequence-structure relationship**. *Proc Natl Acad Sci USA* 1992, **89**:8721–8725.

80. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG: **Funnels, pathways, and the energy landscape of protein folding: a synthesis**. *Proteins* 1995, **21**:167–195.

81. Feng J, Walter NG, Brooks 3rd CL: **Cooperative and directional folding of the preQ1 riboswitch aptamer domain**. *J Am Chem Soc* 2011, **133**:4196–4199.

82. Miner JC, Fenimore PW, Fischer WM, McMahon BH, Sanbonmatsu KY, Tung CS: **Integrative structural studies of the SARS-CoV-2 spike protein during the fusion process (2022)**. *Curr Res Struct Biol* 2022, **4**:220–230.

83. Fudenberg G, Mirny LA: **Higher-order chromatin structure: bridging physics and biology**. *Curr Opin Genet Dev* 2012, **22**: 115–124.

84. Fudenberg G, Getz G, Meyerson M, Mirny LA: **High order chromatin architecture shapes the landscape of chromosomal alterations in cancer**. *Nat Biotechnol* 2011, **29**: 1109–1114.

85. Di Pierro M, Zhang B, Aiden EL, Wolynes PG, Onuchic JN: **Transferable model for chromosome architecture**. *Proc Natl Acad Sci U S A* 2016, **113**:12168–12173.

86. Liu S, Athreya A, Lao Z, Zhang B: **From nucleosomes to compartments: physicochemical interactions underlying chromatin organization**. *Annu Rev Biophys* 2024.

87. Lappala A, Wang CY, Kriz A, Michalk H, Tan K, Lee JT, Sanbonmatsu KY: **Four-dimensional chromosome reconstruction elucidates the spatiotemporal reorganization of the mammalian X chromosome**. *Proc Natl Acad Sci U S A* 2021, **118**.

88. * Jewett AI, Stelter D, Lambert J, Saladi SM, Roscioni OM, Ricci M, Autin L, Maritan M, Bashusqeh SM, Keyes T, *et al.*: **Moltemplate: a tool for coarse-grained modeling of complex biological matter and soft condensed matter physics**. *J Mol Biol* 2021, **433**, 166841.

The authors present coarse-grained models of many large systems, including a whole cell.

89. Tomita M: **Whole-cell simulation: a grand challenge of the 21st century**. *Trends Biotechnol* 2001, **19**:205–210.

90. Tomita M, Hashimoto K, Takahashi K, Shimizu T, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, Hutchison CA: **E-CELL: software environment for whole cell simulation**. *Genome Inform Ser Workshop Genome Inform* 1997, **8**: 147–155.

91. ** Stevens JA, Grunewald F, van Tilburg PAM, Konig M, Gilbert BR, Brier TA, Thornburg ZR, Luthey-Schulten Z, Marrink SJ: **Molecular dynamics simulation of an entire cell**. *Front Chem* 2023, **11**, 1106495.

The authors describe how an integrative approach can be employed to model an entire cell (the minimal cell, JCVI-syn3A). This excellent article details many of the challenges and issues with such simulations.

92. Fisher CP, Plant NJ, Moore JB, Kierzek AM: **QSSPN: dynamic simulation of molecular interaction networks describing gene regulation, signalling and whole-cell metabolism in human cells**. *Bioinformatics* 2013, **29**:3181–3190.

93. Das B, Mitra P: **High-performance whole-cell simulation exploiting modular cell biology principles**. *J Chem Inf Model* 2021, **61**:1481–1492.