Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

The ugly, bad, and good stories of large-scale biomolecular simulations

Chitrak Gupta[1,4], Daipayan Sarkar[1,2], D. Peter Tieleman[3] and Abhishek Singharoy[1,4]

Abstract

Molecular modeling of large biomolecular assemblies exemplifies a disruptive area holding both promises and contentions. Propelled by peta and exascale computing, several simulation methodologies have now matured into user-friendly tools that are successfully employed for modeling viruses, membranous nano-constructs, and key pieces of the genetic machinery. We present three unifying biophysical themes that emanate from some of the most recent multi-million atom simulation endeavors. Despite connecting molecular changes with phenotypic outcomes, the quality measures of these simulations remain questionable. We discuss the existing and upcoming strategies for constructing representative ensembles of large systems, how new computing technologies will boost this area, and make a point that integrative modeling guided by experimental data is the future of biomolecular computations.

Addresses

[1] School of Molecular Sciences, Center for Applied Structural Discovery, Arizona State University at Tempe, Tempe, AZ, 85282, USA
[2] MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI, 48824-1319, USA
[3] Centre for Molecular Simulation and Department of Biological Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada
[4] Biodesign Institute, Tempe, AZ, 85281, USA

Corresponding author: Singharoy, Abhishek (asinghar@asu.edu)
(Gupta C.), (Sarkar D.), (Tieleman D.P.), (Singharoy A.)

Current Opinion in Structural Biology 2022, **73**:102338

This review comes from a themed issue on **Macromolecular Assemblies**

Edited by **Alan Brown** and **Franca Fraternali**

For complete overview of the section, please refer the article collection - Macromolecular Assemblies

Available online 1 March 2022

https://doi.org/10.1016/j.sbi.2022.102338

Abbreviations

MD, Molecular Dynamics; CG, Coarse Grained; ms, millisecond; μs, microsecond; ns, nanosecond; fs, femtosecond.

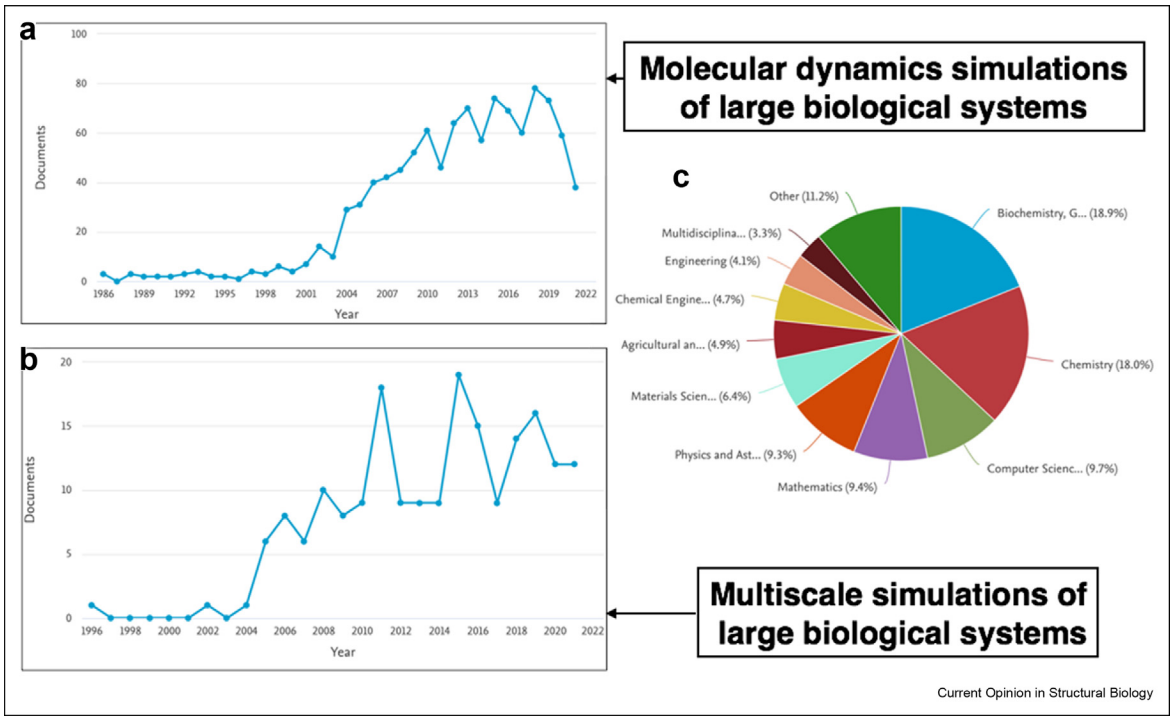
Living cells are brimming with activity of numerous macromolecular complexes. Bolstered by groundbreaking advances in high-resolution imaging and cryo-electron microscopy, and most recently in deep learning [1,2], the atomic structures of many complexes are now accessible. However, structures alone provide limited information about the function of these complexes, in particular how they interact within large-scale networks. Supported by a parallel revolution in multiscale algorithms and computing hardware, molecular simulations of nanometer to sub-micron sized systems are now conceivable. By leveraging all-atom (AA), coarse-grained (CG), and multi-physics molecular dynamics (MD) simulations, the last decade has seen "computational microscopy" of heterogeneous multi-subunit macromolecular complexes, including the ribosome [3], proteasome [4], nuclear pore complex [5], virus capsids (reviewed in the study by Goh et al. [6]), an entire gene [7], crowded environments inside bacterial cells [8], budding of mitochondrial membrane [9], and energy transduction across a photosynthetic organelle [10], and the list continues to grow (Figure 1).

A story of hits and misses

Despite delivering biological insights across the molecular, meso, and even up to phenotypic scales [10], the merits of large-system simulations are worth reflecting on. A discrepancy remains between biological and simulatable timescales, which is aptly summarized in the equation: $Simulation\ time \approx (\alpha N \ln N)\, T_{real}/\Delta\nu$; $\alpha \approx 10^5$ [11]. Simulating a millisecond-scale event ($T_{real} = 1\ ms$) of a 1 million-atom system ($N = 10^6$) with a femtosecond timestep MD ($\Delta = 10^{-15}\ s$) on a petaflop architecture ($\nu = 10^{17}$ flops) will take 159 days, assuming perfect scalability. This duration increases to $\sim 10^9$ years, when $N = 3 \times 10^9$, representing the number of atoms in a simple bacterial cell, and the T_{real} scales up to hours. More dramatic are estimates for simulating firing neurons in a human brain with $\sim 10^{26}$ atoms, which takes up to an impossible 10^{23} years. Besides this practical limit, one also raises philosophical arguments about the value of simulation systems in which the majority of the atoms are bulk water molecules that often contribute only parametrically to the biophysical properties.

Limited-timescale larger system simulations are also plagued by a lack of convergence in capturing diffusive

Figure 1



History of large-scale simulations. Searches on the Scopus database showing an overall increase in the number of documents per year containing (a) MD and (b) multiscale simulations of large biological systems. (c) Pie-chart showing the broad range of application areas of MD simulations of biomolecules (data accumulated over the same period as in panel a). Interestingly, over the pandemic the number of studies with large-system MD simulations seems to have decreased, while those with the multiscale simulations have plateaued.

events. Appropriate tests of convergence are required for all simulations, but more so as the simulation-sizes and timescales grow for monitoring the emergent properties. Studies have looked into the convergence for specific properties, but distributions capturing physical properties of macromolecules do not converge on timescales that can be assessed currently; even the timescales for converging the distributions of smaller molecules such as lipids around membrane proteins are tens of microseconds or more [12]. In practice, this means the interpretation of large-scale simulations requires a careful assessment of which properties are likely to be converged (Summarized in Table 1) *vs.* which ones are affected by an initial-model bias. In some cases, many copies of a particular protein in a system or simulating many replicas of a system increases the statistical information, but this is not helpful for properties that are only apparent over longer time scales. For example, simulating slow interface dynamics or modeling of viscous properties arising from the collective interactions of many system components offer exemplary challenges to MD approaches. Thus, how many repeats are required to trust a large-scale simulation result remains contentious. Rarely have repeat simulations been reported to examine and establish reproducibility. Also, the initial models are often an eclectic mixture of molecular structures determined over a range of resolutions, reflecting different levels of disorder on different components of the model. How this uncertainty in inputs shows up as unphysical interactions in the force fields, and manifests in the quality of the final outcomes is seldom tracked.

Identifying important features within a large-scale simulation creates its own challenges. Unlike single-molecule simulations, where reproducible visualization over multiple replica often picks up interesting conformational changes, such visual hints are lost in crowded large-scale models. Machine learning in the context of dimensionality-reduction approaches can potentially identify key observations. But searching hundred million-dimensional datasets requires a deep network architecture with a huge input layer, which greatly increases the number of weights, often making the training process highly memory-intensive and practically infeasible. Also, the choice of a library of descriptors for training makes the observation of a key property susceptible to the users' bias.

Unlike relatively simple single-protein systems, cell-scale systems are highly complex, thereby complicating direct comparison of theory with experiments. Cell-level results are often not useful to validate

Table 1

Examples of large-scale AA and CG simulations. Table summarizes some recent simulations, scientific discoveries, the physical property used by the authors to ascertain convergence, and validation of simulated results.

Publication (software)	System size/dimensions	Simulation length	Discovery	Converged property	Expt. Validation
AA SIMULATION					
Perilla, 2017 [13] NAMD	HIV-1 capsid 64 M atoms	1.2 μs	Electrostatics, and acoustic properties of empty capsid	RMSD, cross-section area	Ion-binding affinity
Hadden, 2018 [14] NAMD	HBV capsid 6 M atoms	1 μs	Viral mechanism for displaying of cellular signals	Volume and sphericity	Local resolution of cryo-EM density
Bulow, 2019 [15] (GROMACS)	3.6 M atoms	1 μs	Soluble proteins in concentrated solutions diffuse as transient clusters	RMSD	Viscosity, diffusivity
Singharoy, 2019 [10] (NAMD)	Purple bacteria's chromatophore 136 M atoms	500 ns MD, 40 μs CG, 30 ms BD	Electrostatic environment of an organelle supports low light-adaptation	Radial distribution of charge carriers, radius of gyration, PMF profiles	Cell doubling times
Choudhary, 2020 [16] (NAMD)	Cadherin repeats 1.1–2.3 M atoms	3–500 ns	Plasticity and structural determinants of sensory perception	Force distributions	X-ray crystallography SAXS
Durrant, 2020 [17] (NAMD)	H1N1 virus capsid 160 M atoms	121.04 ns	Secondary substrate binding site on capsid surface	RMSD, rate matrices	Mutational assays
Farr, 2021 [18] (LAMMPS)	Nucleosome assembly (30 M atoms/50K beads)<	1 μs REMD 40 μs CG	Plasticity is critical for liquid–liquid phase separation	Persistence length of DNA	Force spectroscopy
Jung, 2021 [7] (GENESIS)	GATA4 gene 1 billion atoms	Only reported benchmarks			
CG SIMULATIONS					
Chavent, 2018 [19] (GROMACS)	Outer membrane proteins (OMP)-containing membrane 480 × 480 nm^2	120 μs	Restricted diffusion of OMPs	OMP cluster size	Single-molecule TIRFM
Vogele, 2018 [20] (NAMD)	Lipids, proteins, and carbon nanotubes in membranes 152 M particles	0.5–2 μs	Infinite-system diffusion coefficients and membrane surface viscosities	Diffusion coefficient using MSD	N/A
Pezeshkian, 2020 [9] (GROMACS)	Back mapping mitochondrial membrane 80 M particles	0.2 μs	Membrane bud formation upon Shiga toxin binding	Total area of the monolayer triangulated surfaces	N/A
Jefferies, 2020 [21] (GROMACS)	Outer membrane vesicles (OMVs) 20 nm diameter	2 μs	Membrane composition can be manipulated to suppress OMV encapsulation	No. of lipid-lipopolysaccharide contacts	Confocal microscopy
Duncan, 2020 [22] (GROMACS)	Inward rectifier potassium channel (Kir2) 3.5 M particles	80 μs	Cooperation between PIP_2 and PS lipids for activating Kir2 channel	Radial distribution function of lipids	N/A
Maity, 2020 [23] (GROMACS)	Macrocyclic polymer fiber 10.8 × 11.7 × 14.5 nm^3	1 μs	Diffusion-driven growth of self-replicating fiber	Distribution of proper and improper dihedrals	Diffusion seen by AFM
Yu, 2021 [24] (NAMD, LAMMPS)	SARS-CoV2 virion 100 × 140 nm^2	10 μs	Collective surface modes of virion	Radial distribution function of proteins	N/A

simulations because they tend to have a small number of data points with a very complex origin. Consider simulating ATP activity of an enzyme as function of environmental conditions. For this case, it is difficult to pinpoint whether, within a chain of computational results specific issues has emerged from the convergence the ATP-hydrolysis computations, force field discrepencies between the protein and the ligand, scaling mismatch between the rate of ATP turnover and that of the environmental changes or compensation of errors, which ultimately leads to a disagreement or agreement between the computations and experiments. Given the sparse amount of experimental data that can be used to either constrain or test the simulation, overfitting can be expected at the level of atomistic models. Due to such over-interpretation, the transferability of cellular models is compromised. An n-fold cross validation analysis is utilized to examine and avoid the over-fitting artifacts, but such procedures only work for data-rich models.

Large simulation systems are still well-defined. When using classical force fields, all chemically equivalent units are treated similarly, allowing the derived parameters to be generalized to other molecules where the same units are used as building blocks. Yet, it is all but impossible to define an experimental system at an equivalent level, due to fluctuations in concentration of components, arising out of metabolism. Molecular modeling a cell-scale system that further includes biological functions such as protein synthesis, degradation, and osmosis, remains unfeasible. Such complications make it even more difficult to interpret simulation results in the context of experimental data.

Early surprises in the enhanced sampling of confined subsystems

Even in the early days of biomolecular computations of multi-million or larger sized systems, and despite the aforementioned outstanding questions of simulation convergence and quality, a number of studies performed over the past five years have paved the way to billion-atoms MD simulations e.g., to model the structure of a complete GATA4 gene (Figure 2). There is a remarkable similarity in the properties derived from these studies of very different biological systems. Three common themes emerge: insights in (i) the mobility and binding of water and ionic species in confined protein environments; (ii) rare substrate binding mechanisms; and (iii) stability and diffusion of proteins in crowded environments. We summarize some of these milestone discoveries.

Solution dynamics in confined environments

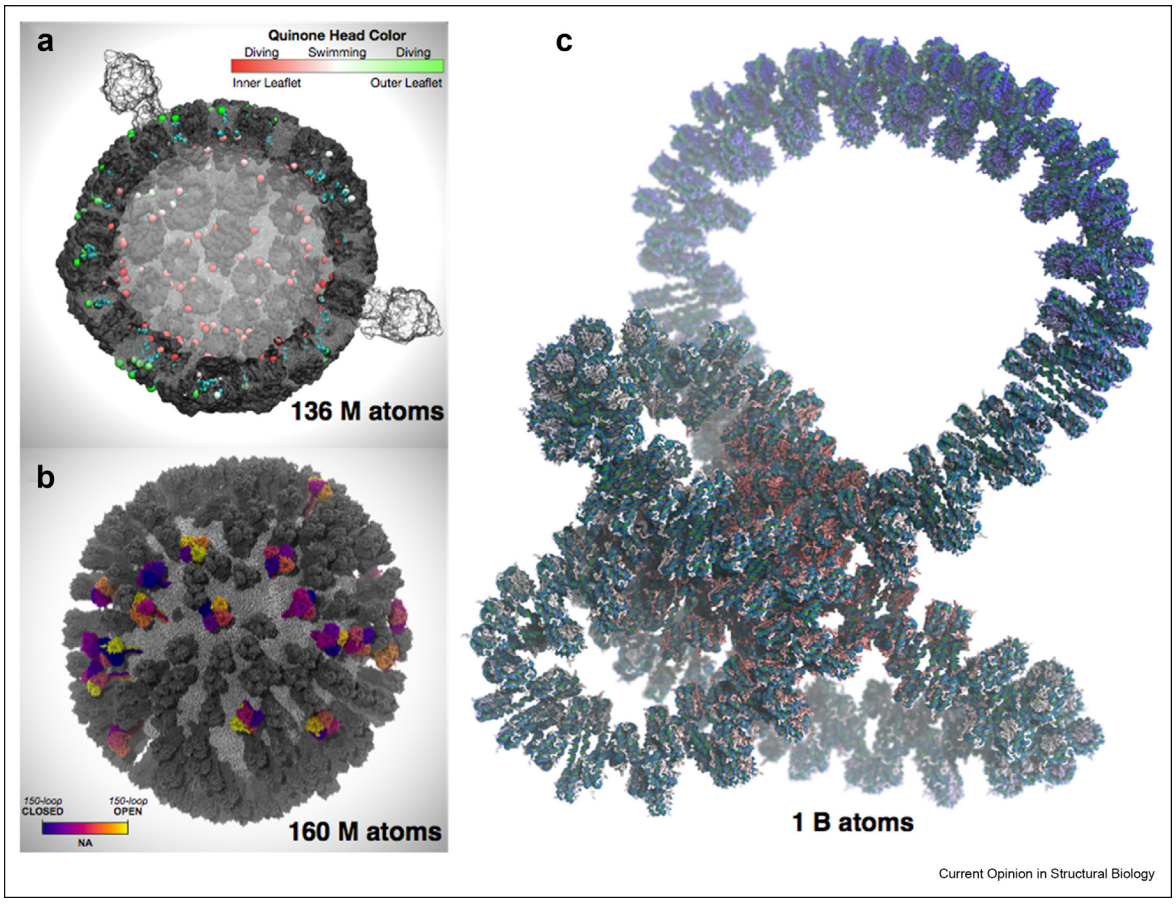
Recognizing that the majority of simulation systems are composed of ions and solvent molecules, the dynamics of which are characterized by rapid picosecond-scale relaxation times, which may scale up to nanoseconds on the protein surfaces, the statistics from microsecond-scale MD simulations was clustered to find ion localization sites in the trimer—of—dimer interfaces of brome mosaic virus capsids [14]. In the same spirit, a rapid exchange of water molecules across an empty poliovirus was determined [25]. This exchange rate is so high that all water molecules inside the capsid can be replaced by new ones from the outside in $\sim$25 μs, explaining the capsid's tolerance to high pressures. Protein—protein interactions mediate these solvent transport properties within large yet finite system boundaries [13]. Such finite-size effects of the nanocontainer cannot be reproduced with periodic representations of the isolated smaller subunits, where the spatial correlations and the impact of confinement on transport rates are lost. Thus, AA or CG simulations of conformationally-correlated transport of solvent molecules and ions across symmetric or asymmetric porous protein nanocontainers have gained popularity over the past few years [6].

Protein–ligand binding mechanisms

Determination of accurate thermodynamic or kinetic properties remains challenging in any large-system simulation. Nonetheless, a key advantage of large simulations is derived by synthesizing ideas from kinetic theories, which could have been easily overlooked had a handful of teams not pursued cell-scale modeling [10,15,17,26]. The multimeric architecture of large systems, which encompasses recurrent instances of key structural features, offers a "multi-copy" representation of the protein or associated cofactors. The so-called "multi-replica" representation of a single protein, seen in free energy simulations, can be recovered in this multi-copy instance of the same protein within the large-system architecture.

For example, within the first 0.5 μs long AA simulation of an entire cell organelle, we simulated 900 ubiquinone molecules embedded within a bacterial photosynthetic membrane composed of 101 protein complexes [10]. On this timescale, which is much shorter than the diffusion of membrane proteins or even lipids, a single quinone is displaced minimally (by < 5 Å). However, a cumulative sampling of 800×0.5 μs $= 400$ μs led to the surprising discovery of two distinct modes of quinone movement in the bacterial membrane—the "swimming" mode with the quinone tail parallel to those of the lipids in lipid-rich environments, and the "diving" mode with the tail perpendicular to the membrane within protein-rich environments. Arguably, 400 replicas of 0.5 μs of single quinone simulations would have captured the swimming motion adequately; however, in the absence of the heterogeneous protein environments in smaller protein-membrane patches (Figure 2A), the rare diving mode would have been missed or underestimated, clearly establishing the need for a large-system simulation.

Figure 2



Examples of large-scale AA models of heterogeneous macromolecular assemblies. (a) a photosynthetic organelle of purple bacteria, showing conformational diversity of "swimming" vs "diving" charge-carrier quinone molecules (colored by their penetration depth in the membrane) [10]; (b) capsid of the H1N1 influenza virus showing an ensemble of open vs closed conformations of salicylic acid binding pockets [17]; and finally (c) a typical architecture of an entire gene, the GATA4 gene, colored by position along the DNA [7].

Following this general notion and combining Markov state models with AA simulation of an entire capsid, two rare conformations of soluble protein loops were discovered on the surface of the influenza virus [17], prompting a "bind and transfer" mechanism for binding sialic acid residues (Figure 2B). Taken together, the realistic environments created by large system simulations offer unique binding conformations that are seldom seen while modeling soluble or membrane-bound proteins and cofactors in isolation.

Protein crowding

Modeling of crowded environments requires quintessential big-system simulations, wherein a majority of the simulation box is filled with macromolecules, taking up to 70% of the volume. Such set-ups are effective in capturing how the heterogeneity in protein environments induces unexpected structural changes. For example, simple volume exclusion would suggest that with crowding the proteins should resort to contracted conformations. However, atomistic simulations reveal that a few proteins actually assume extended forms. Akin to molecular flooding simulations, crowded environments enhance the feasibility of rare events, making them more entropically favorable and reducing their energy barriers over simulations of single proteins in infinite dilution. Thus, large scale crowding simulations offer a tangible way of looking into *in vivo*-like structures that are unachievable within single or multi-replica MD simulations of the same protein in isolation. Crowding studies show that proteins are able to "hitchhike" within the crowded periplasm by binding to lipoprotein carriers and remain rarely un-complexed when in the periplasm, forming both transient and long-lived interactions with proteins, osmolytes, the outer membrane and the cell wall [26]. MD simulations also revealed an increase in the viscosity of protein solutions at higher protein fractions than expected from colloidal models [15]. This increase in viscosity emerges from accessing explicit atomistic interactions and macromolecular clustering within the MD simulations that are missing from the colloidal models. Altogether, starting from well-mixed

models, million atom-MD simulations reveal a plethora of physically correct and experimentally verifiable results, even with sub-microsecond simulations. The challenge, however, lies in generating well-mixed starting states, the methods for which we describe next.

New tricks for old dogs

The core methodologies used in large-system applications have been developed little over the last decade. Despite being established methods, their implementations to model assemblies of $>$ 50 nm sizes pose new practical problems. While all-atom MD simulations clearly gained efficiency by porting most of the underlying computations to Graphical Processor Units (or GPUs), parallel scalability over thousands of nodes is achieved in programs such as NAMD [27] by dynamic load-balancing. These algorithms vary the number of atoms or interactions per processor to preserve the nearest-neighbor computations within the same GPU as more processors are added for handling multi-million atom systems, even going up to 2×10^9 atoms [10].

While equilibration of soluble small proteins has become computationally tractable, though sometimes taking up to tens of microseconds, the same cannot be said for larger constructs. As reported in the study by Perilla et al. [13], it takes over 1200 ns for an empty HIV capsid model to reach the early stages of equilibration. This equilibration problem is further complicated with folding–unfolding transitions, and yet using population correlation functions it has been shown that equilibration time for unfolded proteins is $\sim$100 ns, which is computationally tractable with all-atom MD [28]. In contrast, heterogeneous membrane systems have much higher relaxation or mixing times, given the two orders of magnitude slower diffusion coefficients in the viscous lipid environments. Here, brute-force MD fails to achieve equilibrated models within finite simulation resources. One solution to the problem is to use pre-equilibrated membranes in the simulations, as is provided by platforms such as CHARMM-GUI. However, this platform does not guarantee equilibration of a newly-built system and this approach is challenging for cell-scale systems with exotic shapes. Here, coarse-grained simulations offer a practical alternative to construct large-scale models.

Coarse-grained methodologies

Coarse-grained or CG simulations are a useful tool in large-scale modeling. The Martini model is widely used for simulations of systems that would be very challenging to simulate atomistically (Table 1). At a computational cost of 2–3 orders of magnitude less than AA models, it enables both larger system sizes and, importantly, longer time scales, which are harder to achieve even by parallelization of atomistic simulations. In addition to direct insight from such simulations, CG models are useful to test scaling laws, finite-size effects, and collective phenomena that may not be worth the computational effort on national supercomputers. Two recent examples include simulations of toxin-induced budding of vesicles from membranes [9] and of hydrodynamic modes that explain finite-size effects on diffusion of proteins in membranes [20]. CG simulations also show promise towards building complex atomistic models, as it is in practice considerably easier to build, minimize, and equilibrate large coarse-grained systems than atomistic ones due to softer potentials, much lower particle numbers, and larger integration time steps. One concern is how accurate the resulting distributions of molecules are compared to reality and to AA simulations, but that is a fundamental question for all computational methods, and it is not *a priori* obvious that atomistic simulations are more accurate in this respect. The entropy and temperature dependence of CG models can be inaccurate because of a reduced number of degrees of freedom; free energies often are accurate but this is because enthalpy overcompensates (by parameterization), and the accuracy of rates is variable. Also, for disordered systems the parametrization of CG models is non-trivial. Addressing the issue of distributions, methods such as multiscale coarse-graining (MSCG) or Langevin dynamics of order parameters [29] tend to coarse-grain and yet retain multiple levels of details simultaneously. A recent example of the MSCG approach is the reconstruction of an entire SARS-CoV2 virion showing long–range correlations on the virus surface [24].

Ultra-CG methodologies

While CG simulations are successful in accessing events at the microseconds timescale, further assumptions are needed to study cellular processes at the millisecond timescale. In Brownian dynamics (BD) simulations, entire protein or nucleic acid subunits are approximated as atom-resolved rigid bodies diffusing under the influence of a mean electrostatic and van der Waals field created by the surroundings. Given integration time steps of 10–100 fs and a greatly reduced number of explicit particles, molecular recognition events are routinely captured at the millisecond timescale. For example, we have employed BD simulations to model the membrane-wide transport of charge carriers across a crowded chromatophore, revealing how salinity plays a role controlling the rate of energy conversions [10].

A two-layered coarse-graining was recently introduced to introduce flexibility in ultra-CG models [18]. These simulations show that nucleosome breathing favors stochastic folding of chromatin and promotes compartmentalization of the nucleus by simultaneously boosting the transient nature and heterogeneity of nucleosome–nucleosome contacts. Thus, by combining two coarse-grained models, namely a 3-sites/nucleotide DNA model and an energy model for proteins, nucleosome folding, and organization is

captured. Analogous to the development of all-atom force fields, multiple CG models are now becoming available for tracking distinct physical properties of the same cellular system, exemplified here by application to nucleosomes. Bayesian inferencing offers a first step in exchanging information between multiple models. Also, AA simulations still offer an initial "parts-list" of key interactions that is essential for subsequent CG and ultra-CG endeavors.

Multi-resolution methodologies

A choice made in CG and ultra-CG schemes is that the associated energy surfaces are smooth and easy to sample. Therefore, ruggedness of the "true" free energy surface that underpins the molecular dynamics of the system will not be captured. Reverse coarse-graining as a post-facto refinement is a popular solution to recover the more detailed underlying molecular physics. Nonetheless, by leveraging adiabatic scale separations, a number of multi-resolution MD methodologies have been developed to model large biological systems [30]. These methods are especially complicated to parallelize because (i) boundaries and coupling between the different levels of resolutions are monitored on-the-fly, (ii) adaptive decision-making is needed to treat the same moieties at different resolutions (i.e. on energy surfaces of different ruggedness) at different timepoints and (iii) despite larger timesteps, compute nodes engaged at the CG level wait for completion of the slower high-level tasks, thus affecting parallel efficiency of the overall computation. We highlight some recent examples of multi-resolution MD simulations that will be of relevance to modeling cell-scale systems in the future.

Multi-resolution simulations have been applied to phase separation between saturated lipids, unsaturated lipids, and cholesterol [31]. In these simulations, inhomogeneous populations of disordered chains with higher chain configurational entropy are found to have strong interdigitation [32]. Membrane dynamics revealed specifically by the multi-resolution methods show the multicomponent complexity of bacterial membranes, central to antimicrobial resistance [33]. The multi-resolution methods have also started gaining traction in the structure prediction area, particularly for modeling DNA/RNA aptamers [34] and for rapid protein folding [35]. Recovery of near-atomic resolution in these simulations allows estimates of realistic fluctuations, and consequently experimentally verifiable rates of association/dissociation events [34]. Thus, the precise kinetic information that is accessible in multi-resolution simulations is clearly an advantage over brute-force CG models.

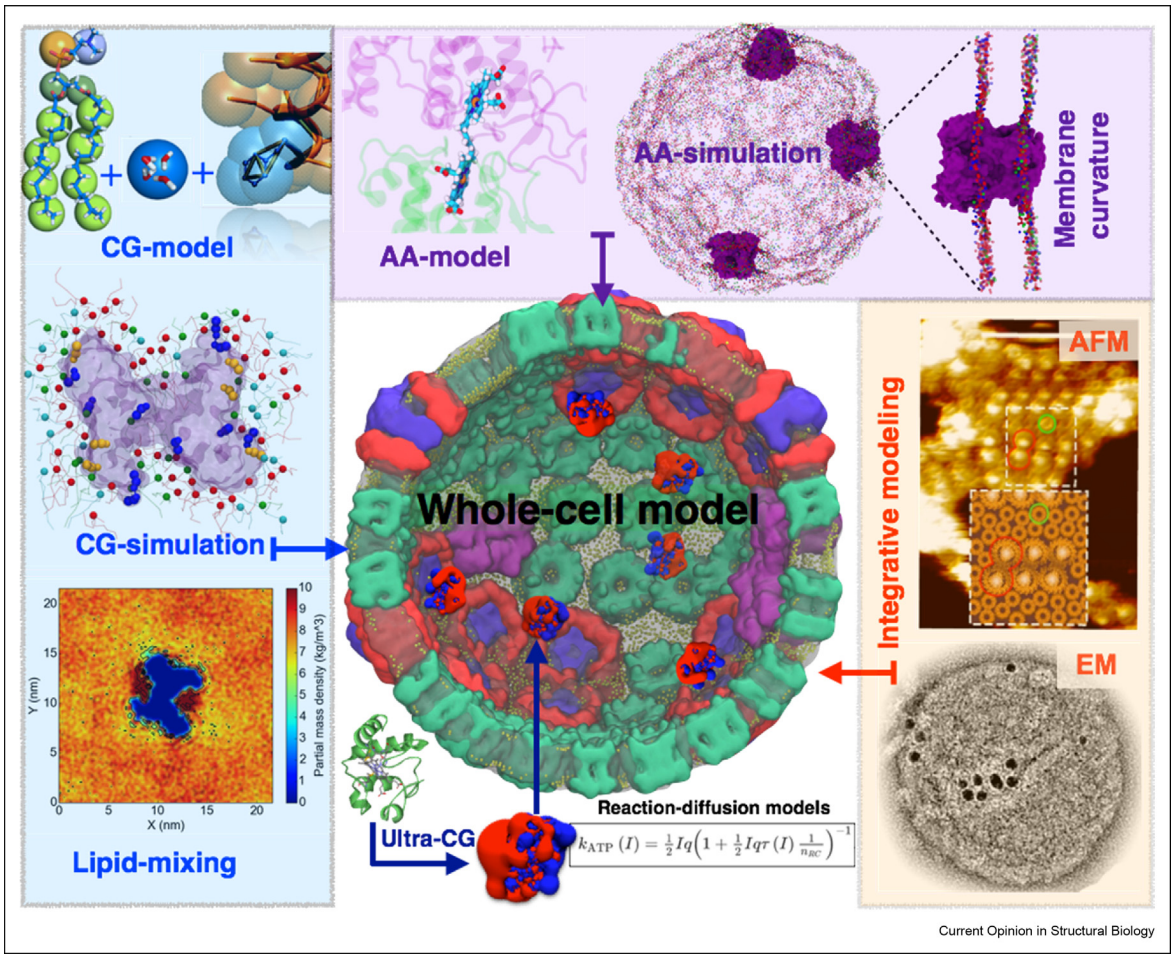
Practical means of cell-scale molecular modeling

Intuitively speaking, one would assume that performing AA simulations would be a first step prior to subsequent coarse-graining, as has been done historically (Review in the study by Saunders et al. [30]). However, this conventional wisdom seems to be less useful for simulating large systems as equilibration of the model at the atomic level remains contentious. Presented in Figure 3, an alternate modeling approach starts at an intermediate resolution. If the CG force fields are available, these simulations are useful to overcome diffusive bottlenecks to obtain well-mixed subsystems. Thereafter, the equilibrated CG subsystems are stitched together and back-mapped to complete AA models, using methods such as dynamically triangulated surface reconstruction [9]. A practical issue arises during this step. Re-solvation of closed membrane systems and backmapping of the CG models to AA representations causes density and pressure imbalances in MD simulations, creating unphysical cavities on the membrane—protein interfaces [36]. The most widely used Martini-level script [37] can go almost seamlessly from Cooke-type lipids (3 beads/lipid) to Martini to atomistic. But in practice the resulting atomistic systems still have overlapping problems and are difficult to minimize without manual intervention. Schemes such as LipidWrapper [38] are available to simultaneously identify holes across the system and fill them with membrane fragments to restore water density and overall pressure of the simulation box. Successful recovery of the whole-system AA models brings to light the molecular origins of the diffusive bottlenecks seen at the CG level. Since the AA models refined from the CG simulations are nearly-equilibrated, short nanosecond-scale MD simulations are adequate to construct ultra-CG models (e.g., BD trajectories) using mean field approximations. Thus, capitalizing on the CG-guided AA models, the ultra-CG results can now be scaled up from atomistic details to spatio-temporal ranges amenable to cell-biology measurements [39].

Dawn of new technologies

The upcoming exascale computers will be able to perform 10^{18} floating point operations per second, easily increasing the timescale of the simulations in Table 1 by at least an order of magnitude, already in 2022. Using distributed computing on folding@home, exascale performance is already achieved for smaller systems, such as the SARS-CoV-2 spike protein [40]. By optimizing memory usage of the nonbonded interactions, billion-atom simulations of fully connected biological systems are now feasible [7], and exascale computing will only boost the timescale of such simulations. Machine learning tools such as Deep drive MD [41] are well posited to enrich the statistics gained from the large-scale simulations. The dimensionality problem of learning dynamics at the multi-million to billion dimensions will be a challenge. However, large matrix manipulations within *ab-initio* MD can now be accelerated on tensor cores without compromising floating point precision [42]. Thus, armed with the graphical

Figure 3



Recipe for large system modeling. A workflow of the creation of CG and AA parts-list to be integrated into a cell-scale model using an organizational blue-prints of (e.g. location, shape, stoichiometry) from experimental datasets.

and tensor processing units that already challenge the limits set by Moore's law, and the disruptive simulation methodologies that have started capturing the cellular mechanics, we note that at least the million-atom models will shortly become a norm in the area of molecular modeling.

Integrative modeling is the future

The atomic models and their equilibrated forms contribute to the parts list on which the cellular models are based. Construction of such larger architectures necessitates additional knowledge of the "blueprint" of the assembly and conveying this information as part of the AA or CG simulation field. These blueprints are now experimentally tractable in varying sparsity: while cryo-electron tomography and microscopy, and also X-ray scattering and high-speed atomic force imaging experiments offer information on anywhere between near-atomic positions to inter-domain conformations up to shapes, cross-linking experiments offer a distribution of intra- and inter-macromolecular distances in dynamic assemblies; optical spectroscopy has been key in determining protein location and even concentration close to the cell surface together with mass spectrometry. Bayesian inferencing tools have emerged that can build on these meso-to-macroscale experimental data, while still maintaining detailed balance with the chemical details derived from AA and CG levels of description [43]. A key strength of integrative modeling lies in the consensus interpretation it offers, resolving both uncertainties in the experimental data as well as sampling inadequacies of the molecular simulations. Thus, the outcome of such modeling are data-guided ensembles that visit multiple metastable states of even large multimeric systems, where each state is captured by one or more experimental datasets. Unsurprisingly, methods such as Molecular Dynamics Flexible Fitting, or maximum entropy-based conformational sampling approaches such as Modeling Employing Limited Data [44] and CryoFold [45] are available across almost all popular MD engines. Another

class of inferencing scheme, namely Meta-inference [46], allows constructing such data-guided ensemble models by also accounting for experimental errors. However, when using the maximum entropy principle to enforce agreement between simulation and experiment, one free parameter is used for each data point. As a consequence, different chemically equivalent units might be treated differently. This does not allow the corrections to be transferred to other molecules, for which new experimental data would be required. Some recent examples include our determination of molecular dynamics of an active Ryanodine receptor of size 2.2 MDa [47], molecular modeling of the desmosome complexes to study bacterial cell adhesion [39], and flagellar motors to study chemotaxis and microbial locomotion [48]. Recognizing that cooperative interactions within all these complexes could have only been determined using data-guided simulations, we posit that molecular simulations of even larger cellular systems will heavily depend on today's developments in integrative modeling.

Conflict of interest statement

Nothing declared.

Acknowledgement

Work in DPT's group is supported by the Natural Sciences and Engineering Research Council (Canada) and the Canadian Institutes for Health Research. Further support came from the Canada Research Chairs Program. AS acknowledges start-up funds from the SMS and CASD at Arizona State University, CAREER award by NSF-MCB 1942763. Many thanks to *Lorenzo Casalino* and *Rommie Amaro* (UCSD), *Karissa Sanbonmatsu* (LANL) and *John Vant* (ASU) for contributing unpublished images to Figure 1. We also thank the Biophysics community on Twitter, particularly *Syma Khalid*, *Mathieu Chavent* and *Viola Vogele* for valuable suggestions.

References

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

* of special interest
** of outstanding interest

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589.

2. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, *et al.*: **Highly accurate protein structure prediction for the human proteome**. *Nature* 2021, https://doi.org/10.1038/s41586-021-03828-1.

3. Bock LV, Kolář MH, Grubmüller H: **Molecular simulations of the ribosome and associated translation factors**. *Curr Opin Struct Biol* 2018, **49**:27–35.

4. Wehmer M, Rudack T, Beck F, Aufderheide A, Pfeifer G, Plitzko JM, Förster F, Schulten K, Baumeister W, Sakata E: **Structural insights into the functional cycle of the ATPase module of the 26S proteasome**. *Proc Natl Acad Sci Unit States Am* 2017, **114**:1305–1310.

5. Fragasso A, de Vries HW, Andersson J, van der Sluis EO, van der Giessen E, Dahlin A, Onck PR, Dekker C: **A designer FG-Nup that reconstitutes the selective transport barrier of the nuclear pore complex**. *Nat Commun* 2021, **12**:1–15. *121* 2021.

6. Goh BC, Hadden JA, Bernardi RC, Singharoy A, McGreevy R, Rudack T, Cassidy CK, Schulten K. *Computational methodologies for real-space structural refinement of large macromolecular complexes*, vol. 45; 2016:253–278. *101146/annurev-biophys-062215-011113*.

7. * 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.
This article is the first instance of a billion-atom simulation with 1ns/day performance.

8. Feig M, Sugita Y: **Whole-cell models and simulations in molecular detail**. *Annu Rev Cell Dev Biol* 2019, **35**:191–211.

9. Pezeshkian W, König M, Wassenaar TA, Marrink SJ: **Backmapping triangulated surfaces to coarse-grained membrane models**. *Nat Commun* 2020, **11**:1–9. *111* 2020.

10. ** Singharoy A, Maffeo C, Delgado-Magnero KH, Swainsbury DJK, Sener M, Kleinekathöfer U, Vant JW, Nguyen J, Hitchcock A, Isralewitz B, *et al.*: **Atoms to phenotypes: molecular design principles of cellular energy metabolism**. *Cell* 2019, **179**:1098–1111. e23.
This article is the first integrative approach to first-principles modeling of a whole living cell. The authors present a 100 M all-atom molecular dynamics simulation of an entire bacterial photosynthetic organelle connecting phenotypic outcomes to molecular interactions.

11. Netz RR, Eaton WA: **Estimating computational limits on theoretical descriptions of biological cells**. *Proc Natl Acad Sci U S A* 2021, **118**. 2022753118.

12. Corradi V, Mendez-Villuendas E, Ingólfsson HI, Gu RX, Siuda I, Melo MN, Moussatova A, Degagné LJ, Sejdiu BI, Singh G, *et al.*: **Lipid-protein interactions are unique fingerprints for membrane proteins**. *ACS Cent Sci* 2018, **4**:709–717.

13. Perilla JR, Schulten K: **Physical properties of the HIV-1 capsid from all-atom molecular dynamics simulations**. *Nat Commun* 2017, **8**:1–10.

14. Hadden JA, Perilla JR, Schlicksup CJ, Venkatakrishnan B, Zlotnick A, Schulten K: **All-atom molecular dynamics of the HBV capsid reveals insights into biological function and cryo-EM resolution limits**. *Elife* 2018, **7**.

15. von Bülow S, Siggel M, Linke M, Hummer G: **Dynamic cluster formation determines viscosity and diffusion in dense protein solutions**. *Proc Natl Acad Sci U S A* 2019, **116**:9843–9852.

16. Choudhary D, Narui Y, Neel BL, Wimalasena LN, Klanseck CF, De-la-Torre P, Chen C, Araya-Secchi R, Tamilselvan E, Sotomayor M: **Structural determinants of protocadherin-15 mechanics and function in hearing and balance perception**. *Proc Natl Acad Sci Unit States Am* 2020, **117**:24837–24848.

17. ** Durrant JD, Kochanek SE, Casalino L, Ieong PU, Dommer AC, Amaro RE: **Mesoscale All-atom influenza virus simulations suggest new substrate binding mechanism**. *ACS Cent Sci* 2020, **6**:189–196.
This article reports a Markov State Model of an entire virus capsid to seek cryptic pockets on the capsid surface.

18. 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**:1–17.

19. Chavent M, Duncan AL, Rassam P, Birkholz O, Hélie J, Reddy T, Beliaev D, Hambly B, Piehler J, Kleanthous C, *et al.*: **How nanoscale protein interactions determine the mesoscale dynamic organisation of bacterial outer membrane proteins**. *Nat Commun* 2018, **9**:1–12. *91* 2018.

20. Vögele M, Köfinger J, Hummer G: **Hydrodynamics of diffusion in lipid membrane simulations**. *Phys Rev Lett* 2018, **120**:268104.

21. Jefferies D, Khalid S: **To infect or not to infect: molecular determinants of bacterial outer membrane vesicle internalization by host membranes**. *J Mol Biol* 2020, **432**:1251–1264.

22. Duncan AL, Corey RA, Sansom MSP: **Defining how multiple lipid species interact with inward rectifier potassium (Kir2) channels**. *Proc Natl Acad Sci U S A* 2020, **117**:7803–7813.

23. Maity S, Ottelé J, Santiago GM, Frederix PWJM, Kroon P, Markovitch O, Stuart MCA, Marrink SJ, Otto S, Roos WH: **Caught in the act: mechanistic insight into supramolecular polymerization-driven self-replication from real-time visualization**. *J Am Chem Soc* 2020, **142**:13709–13717.

24. 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.

25. Andoh Y, Yoshii N, Yamada A, Fujimoto K, Kojima H, Mizutani K, Nakagawa A, Nomoto A, Okazaki S: **All-atom molecular dynamics calculation study of entire poliovirus empty capsids in solution**. *J Chem Phys* 2014:141.

26. * Pedebos C, Smith IPS, Boags A, Khalid S: **The hitchhiker's guide to the periplasm: unexpected molecular interactions of polymyxin B1 in E. coli**. *Structure* 2021, **29**:444–456. e2.
The authors here report a very important biophysical problem on protein–protein interaction in the periplasm of Gram-negative bacteria as the molecular system is a highly crowded molecular environment. Diffusion of antibiotics across the periplasm of this large molercular system was studied using all-atom molecular dynamics simulations.

27. Phillips JC, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, Bernardi RC, Buch R, Fiorin G, Hénin J, Jiang W, *et al.*: **Scalable molecular dynamics on CPU and GPU architectures with NAMD**. *J Chem Phys* 2020, **153**, 044130.

28. Levy RM, Dai W, Deng N-J, Makarov DE: **How long does it take to equilibrate the unfolded state of a protein?** *Protein Sci* 2013, **22**:1459.

29. Singharoy A, Cheluvaraja S, Ortoleva P: **Order parameters for macromolecules: application to multiscale simulation**. *J Chem Phys* 2011:134.

30. Saunders MG, Voth GA: **Coarse-graining methods for computational biology**. *Annu Rev Biophys* 2013, **42**:73–93.

31. Liu Y, Vries AH de, Pezeshkian W, Marrink SJ: **Capturing membrane phase separation by dual resolution molecular dynamics simulations**. *J Chem Theor Comput* 2021, https://doi.org/10.1021/ACS.JCTC.1C00151.

32. Srivastava A, Debnath A: **Asymmetry and rippling in mixed surfactant bilayers from all-atom and coarse-grained simulations: interdigitation and per chain entropy†**. *J Phys Chem B* 2020, **124**:6420–6436.

33. Matamoros-Recio A, Franco-Gonzalez JF, Forgione RE, Torres-Mozas A, Silipo A, Martín-Santamaría S: **Understanding the antibacterial resistance: computational explorations in bacterial membranes**. *ACS Omega* 2021, **6**:6041–6054.

34. Shang X, Guan Z, Zhang S, Shi L, You H: **Predicting the aptamer SYL3C–EpCAM complex's structure with the Martini-based simulation protocol**. *Phys Chem Chem Phys* 2021, **23**:7066–7079.

35. Zhong Q, Li G: **Adaptively iterative multiscale switching simulation strategy and applications to protein folding and structure prediction**. *J Phys Chem Lett* 2021, **12**:3151–3162.

36. Wilson E, Vant J, Layton J, Boyd R, Lee H, Turilli M, Hernández B, Wilkinson S, Jha S, Gupta C, *et al.*: **Large-scale molecular dynamics simulations of cellular compartments**. *Methods Mol Biol* 2021, **2302**:335–356.

37. Wassenaar TA, Pluhackova K, Böckmann RA, Marrink SJ, Tieleman DP: **Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models**. *J Chem Theor Comput* 2014, **10**: 676–690.

38. Durrant JD, Amaro RE, LipidWrapper: **An algorithm for generating large-scale membrane models of arbitrary geometry**. *PLoS Comput Biol* 2014, **10**, e1003720.

39. Sikora M, Ermel UH, Seybold A, Kunz M, Calloni G, Reitz J, Martin Vabulas R, Hummer G, Frangakis AS: **Desmosome architecture derived from molecular dynamics simulations and cryo-electron tomography**. *Proc Natl Acad Sci U S A* 2020, **117**: 27132–27140.

40. Zimmerman MI, Porter JR, Ward MD, Singh S, Vithani N, Meller A, Mallimadugula UL, Kuhn CE, Borowsky JH, Wiewiora RP, *et al.*: **SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome**. *Nat Chem* 2021, **13**:651–659.

41. Lee H, Turilli M, Jha S, Bhowmik D, Ma H, Ramanathan A: **DeepDriveMD: deep-learning driven adaptive molecular simulations for protein folding**. *Proc DLS* 2019, https://doi.org/10.1109/DLS49591.2019.00007. *Deep Learn Supercomput - Held conjunction with SC 2019 Int Conf High Perform Comput Networking*, Storage Anal 2019.

42. Finkelstein J, Smith JS, Mniszewski SM, Barros K, Negre CFA, Rubensson EH, Niklasson AMN: **Quantum-based molecular dynamics simulations using tensor cores**. *210702737v1 arXiv* 2021.

43. Perez A, Morrone JA, Dill KA: **Accelerating physical simulations of proteins by leveraging external knowledge**. *Wiley Interdiscip Rev Comput Mol Sci* 2017, **7**.

44. MacCallum JL, Perez A, Dill KA: **Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference**. *Proc Natl Acad Sci Unit States Am* 2015, **112**:6985–6990.

45. ** Shekhar M, Terashi G, Gupta C, Sarkar D, Debussche G, Sisco N, Nguyen J, Mondal A, Zook J, Vant J, *et al.*: **Cryofold: determining protein structures and data- guided ensembles from cryo-em density maps**. *SSRN Electron J* 2021, https://doi.org/10.2139/SSRN.3866834.
The authors introduce Bayesian inferencing for ab initio determination of molecular ensembles of soluble, membrane and complex proteins simultaniously from cryo-EM, X-ray and NMR datasets.

46. Bonomi M, Camilloni C, Cavalli A, Vendruscolo M, Metainference: **A Bayesian inference method for heterogeneous systems**. *Sci Adv* 2016, **2**, e1501177.

47. Dashti A, Mashayekhi G, Shekhar M, Ben Hail D, Salah S, Schwander P, des Georges A, Singharoy A, Frank J, Ourmazd A: **Retrieving functional pathways of biomolecules from single-particle snapshots**. *Nat Commun* 2020, **11**:1–14. *111* 2020.

48. Cassidy CK, Himes BA, Sun D, Ma J, Zhao G, Parkinson JS, Stansfeld PJ, Luthey-Schulten Z, Zhang P: **Structure and dynamics of the E. coli chemotaxis core signaling complex by cryo-electron tomography and molecular simulations**. *Commun Biol* 2020, **3**:1–10.