International Journal of
HIGH PERFORMANCE
COMPUTING APPLICATIONS

Special Issue Paper

The International Journal of High Performance Computing Applications
2021, Vol. 35(5) 432–451
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10943420211006452
journals.sagepub.com/home/hpc
SAGE

AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics

Lorenzo Casalino[1,†], Abigail C Dommer[1,†], Zied Gaieb[1,†], Emilia P Barros[1], Terra Sztain[1], Surl-Hee Ahn[1], Anda Trifan[2,3], Alexander Brace[2], Anthony T Bogetti[4], Austin Clyde[2,5], Heng Ma[2], Hyungro Lee[6], Matteo Turilli[6], Syma Khalid[8], Lillian T Chong[4], Carlos Simmerling[9], David J Hardy[3], Julio DC Maia[3], James C Phillips[3], Thorsten Kurth[10], Abraham C Stern[10], Lei Huang[11], John D McCalpin[11], Mahidhar Tatineni[12], Tom Gibbs[10], John E Stone[3], Shantenu Jha[6,7], Arvind Ramanathan[2] and Rommie E Amaro[1]

Abstract

We develop a generalizable AI-driven workflow that leverages heterogeneous HPC resources to explore the time-dependent dynamics of molecular systems. We use this workflow to investigate the mechanisms of infectivity of the SARS-CoV-2 spike protein, the main viral infection machinery. Our workflow enables more efficient investigation of spike dynamics in a variety of complex environments, including within a complete SARS-CoV-2 viral envelope simulation, which contains 305 million atoms and shows strong scaling on ORNL Summit using NAMD. We present several novel scientific discoveries, including the elucidation of the spike's full glycan shield, the role of spike glycans in modulating the infectivity of the virus, and the characterization of the flexible interactions between the spike and the human ACE2 receptor. We also demonstrate how AI can accelerate conformational sampling across different systems and pave the way for the future application of such methods to additional studies in SARS-CoV-2 and other molecular systems.

Keywords

Molecular dynamics, deep learning, multiscale simulation, weighted ensemble, computational virology, SARS-CoV-2, COVID19, HPC, GPU, AI

1. Justification

We:

- develop an AI-driven multiscale simulation framework to interrogate SARS-CoV-2 spike dynamics,
- reveal the spike's full glycan shield and discover that glycans play an active role in infection, and
- achieve new high watermarks for classical MD simulation of viruses (305 million atoms) and the weighted ensemble method (~500,000 atoms).

2. Overview of the problem

The SARS-CoV-2 virus is the causative agent of COVID19, a worldwide pandemic that has infected over 35 million people and killed over one million. As such it is

[1] University of California San Diego, La Jolla, CA, USA
[2] Argonne National Lab, Lemont, IL, USA
[3] University of Illinois at Urbana-Champaign, Urbana, IL, USA
[4] University of Pittsburgh, Pittsburgh, PA, USA
[5] University of Chicago, Chicago, IL, USA
[6] Rutgers University, Piscataway, NJ, USA
[7] Brookhaven National Lab, Upton, NY, USA
[8] University of Southampton, Southampton, UK
[9] Stony Brook University, Stony Brook, NY, USA
[10] NVIDIA Corporation, Santa Clara, CA, USA
[11] Texas Advanced Computing Center, Austin, TX, USA
[12] San Diego Supercomputing Center, La Jolla, CA, USA
[†] Authors with symbol indicate equal contribution.

Corresponding authors:
Arvind Ramanathan, Argonne National Lab, Lemont, IL, USA.
Email: ramanathana@anl.gov

Rommie E Amaro, University of California San Diego, 3234 Urey Hall, MC-0340, La Jolla, CA 92093-0340, USA.
Email: ramaro@ucsd.edu

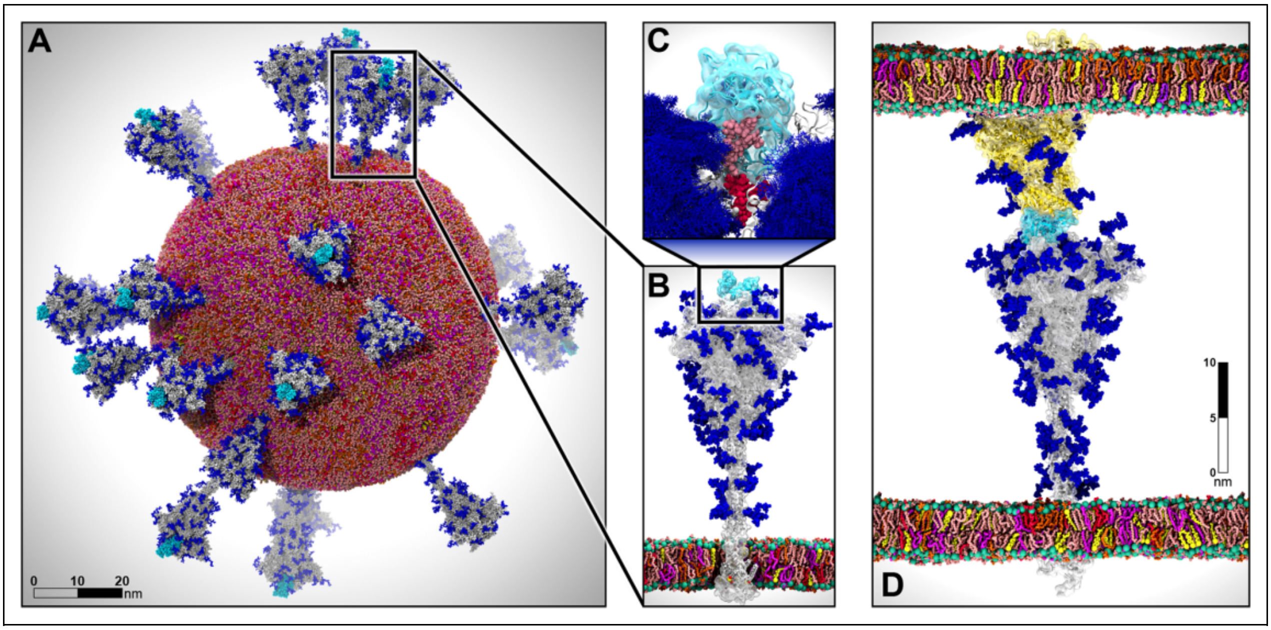

Figure 1. Multiscale modeling of SARS-CoV-2. A) All-atom model of the SARS-CoV-2 viral envelope (305 M atoms), including 24 spike proteins (colored in gray) in both the open (16) and closed states (8). The RBDs in the "up" state are highlighted in cyan) N-/O-Glycans are shown in blue. Water molecules and ions have been omitted for clarity. B) Full-length model of the glycosylated SARS-CoV-2 spike protein (gray surface) embedded into an ERGIC-like lipid bilayer (1.7 M atoms). RBD in the "up" state is highlighted in cyan. C) The glycan shield is shown by overlaying multiple conformations for each glycan collected at subsequent timesteps along the dynamics (blue bushlike representation). Highlighted in pink and red are two N-glycans (linked to N165 and N234, respectively) responsible for the modulation of the RBD dynamics, thus priming the virus for infection. The RBD "up" is depicted with a cyan surface. D) Two-parallel-membrane system of the spike-ACE2 complex (8.5 M atoms). The spike protein, embedded into an ERGIC-like membrane, is depicted with a gray transparent surface, whereas ACE2 is shown with a yellow transparent surface and it is embedded into a lipid bilayer mimicking the composition of mammalian cell membranes. Glycans are shown in blue, whereas water has been omitted for clarity. Visualizations were created in VMD using its custom GPU-accelerated ray tracing engine (Humphrey et al., 1996; Stone et al., 2013a, 2013b, 2016a, 2016b).

the subject of intense scientific investigations. Researchers are interested in understanding the structure and function of the proteins that constitute the virus, as this knowledge aids in the understanding of transmission, infectivity, and potential therapeutics.

A number of experimental methods, including x-ray crystallography, cryoelectron (cryo-EM) microscopy, and cryo-EM tomography are able to inform on the structure of viral proteins and the other (e.g., host cell) proteins with which the virus interacts. Such structural information is vital to our understanding of these molecular machines, however, there are limits to what experiments can tell us. For example, achieving high resolution structures typically comes at the expense of dynamics: flexible parts of the proteins (e.g., loops) are often not resolved, or frequently not even included in the experimental construct. Glycans, the sugar-like structures that decorate viral surface proteins, are particularly flexible and thus experimental techniques are currently unable to provide detailed views into their structure and function beyond a few basic units. Additionally, these experiments can resolve static snapshots, perhaps catching different states of the protein, but they are unable to elucidate the thermodynamic and kinetic relationships between such states.

In addition to the rich structural datasets, researchers have used a variety of proteomic, glycomic, and other methods to determine detailed information about particular aspects of the virus. In one example, deep sequencing methods have informed on the functional implications of mutations in a key part of the viral spike protein (Starr et al., 2020). In others, mass spectrometry approaches have provided information about the particular composition of the glycans at particular sites on the viral protein (Shajahan et al., 2020; Watanabe et al., 2020). These data are each valuable in their own right but exist as disparate islands of knowledge. Thus there is a need to integrate these datasets into cohesive models, such that the fluctuations of the viral particle and its components that cause its infectivity can be understood.

In this work, we used all-atom molecular dynamics (MD) simulations to combine, augment, and extend available experimental datasets in order to interrogate the structure, dynamics, and function of the SARS-CoV-2 spike protein (Figure 1). The spike protein is considered the main infection machinery of the virus because it is the only glycoprotein on the surface of the virus and it is the molecular machine that interacts with the human host cell receptor, ACE2, at the initial step of infection. We have developed

MD simulations of the spike protein at three distinct scales, where each system (and scale) is informative, extensive, and scientifically valuable in its own right (as will be discussed). This includes the construction and simulation of the SARS-CoV-2 viral envelope that contains 305 million atoms, and is thus among one of the largest and most complex biological systems ever simulated (Figure 1A). We employ both conventional MD as well as the weighted ensemble enhanced sampling approach (which again breaks new ground in terms of applicable system size). We then collectively couple these breakthrough simulations with artificial intelligence (AI) based methods as part of an integrated workflow that transfers knowledge gained at one scale to “drive” (enhance) sampling at another.

An additional significant challenge faced in bringing this work to fruition is that it pushes the boundaries of several fields simultaneously, including biology, physics, chemistry, mathematics, and computer science. It is intersectional in nature, and requires the collective work of and effective communication among experts in each of these fields to construct, simulate, and analyze such systems—all while optimizing code performance to accelerate scientific discovery against SARS-CoV-2.

Our work has brought HPC to bear to provide unprecedented detail and atomic-level understanding of virus particles and how they infect human cells. Our efforts shed light on many aspects of the spike dynamics and function that are currently inaccessible with experiment, and have provided a number of experimentally testable hypotheses—some of which have already been experimentally validated. By doing so, we provide new understandings for vaccine and therapeutic development, inform on basic mechanisms of viral infection, push technological and methodological limits for molecular simulation, and bring supercomputing to the forefront in the fight against COVID19.

2.1. Methods

2.1.1. Full-length, fully-glycosylated spike protein. In this work, we built two full-length glycosylated all-atom models of the SARS-CoV-2 S protein in both closed and open states, fully detailed in Casalino et al (Casalino et al., 2020b). The two all-atom models were built starting from the cryo-EM structures of the spike in the open state (PDB ID: 6VSB (Wrapp et al., 2020a)), where one receptor binding domain (RBD) is in the “up” conformation, and in the closed state, bearing instead three RBDs in the “down” conformation (PDB ID: 6VXX (Walls et al., 2020)). Given that the experimental cryo-EM structures were incomplete, the remaining parts, namely (i) the missing loops within the head (residues 16–1141), (ii) the stalk (residues 1141–1234) and (iii) the cytosolic tail (residues 1235–1273), were modelled using MODELLER (Šali and Blundell, 1993) and I-TASSER (Zhang, 2008). The resulting full-length all-atom constructs were subsequently N-/O-glycosylated using the Glycan Reader & Modeler tool (Jo et al., 2008) integrated into Glycan Reader (Jo et al., 2011) in CHARMM-GUI (Park et al., 2019). Importantly, an asymmetric glycoprofile was generated (e.g., not specular across monomers) taking into account the N-/O-glycans heterogeneity as described in the available glycoanalytic data (Shajahan et al., 2020; Watanabe et al., 2020). The two glycosylated systems were embedded into their physiological environment composed of an ERGIC-like lipid bilayer (Casares et al., 2019; Van Meer et al., 2008) built using CHARMM-GUI (Jo et al., 2008; Wu et al., 2014), explicit TIP3P water molecules (Jorgensen et al., 1983a), and neutralizing chloride and sodium ions at 150 mM concentration, generating two final systems each tallying ~1.7 million atoms. Using CHARMM36 all-atom additive force fields (Guvench et al., 2009; Huang and Mackerell, 2013) and NAMD 2.14 (Phillips et al., 2020), the systems were initially relaxed through a series of minimization, melting (for the membrane), and equilibration cycles. The equilibrated systems were then subjected to multiple replicas of all-atom MD simulation production runs of the open (6x) and closed (3x) systems on the NSF Frontera computing system at the Texas Advanced Computing Center (TACC). A cumulative extensive sampling of ~4.2 and ~1.7 μs was attained for the open and closed systems, respectively. Additionally, a third, mutant system bearing N165A and N234A mutations was built from the open system in order to delete the N-linked glycans and delineate their structural role in the RBD dynamics. This system was also simulated for ~4.2 μs in 6 replicas (Casalino et al., 2020b).

2.1.2. ACE2-RBD complex MD simulations. The model of the ACE2-RBD complex was based on cryo-EM structure trapping ACE2 as a homo-dimer co-complexed with two RBDs and B0AT1 transporter (PDB ID 6M17 (Yan et al., 2020)). Upon removal of B0AT1, ACE2 missing residues at the C terminal end were modeled using I-TASSER (Zhang, 2008), whereas those missing at the N terminal end were taken from 6M0J and properly positioned upon alignment of the N terminal helix. Zinc sites including the ions and the coordinating residues were copied from 1R42. The construct was fully N-/O-glycosylated using CHARMM-GUI tools (Jo et al., 2008, 2011; Park et al., 2019) for glycan modeling, reproducing the glycan heterogeneity for ACE2 and RBD reported in the available glycoanalytic data (Shajahan et al., 2020; Sun et al., 2020; Zhao et al., 2020). Similarly, the apo ACE2 homo-dimer was also built upon removal of the RBDs from the holo construct. The glycosylated models were embedded into separate lipid patches with a composition mimicking that of mammalian cellular membranes (Casares et al., 2019; Van Meer et al., 2008) and simulated in explicit water molecules at 150 mM ion concentration, affording two final systems of ~800,000 atoms each. MD simulations were performed using CHARMM36 all-atom additive force fields (Guvench et al., 2009; Huang and Mackerell, 2013) along with NAMD 2.14 (Phillips et al., 2020). The MD protocol was identical to that adopted for the simulation of the full-length spike and it is fully described in Casalino et al (Casalino et al., 2020b). This work is fully detailed in Barros et al (Barros et al., 2020).

2.1.3. Weighted ensemble simulations of spike opening. The spike must undergo a large conformational change for activation and binding to ACE2 receptors, where the receptor binding domain transitions from the "down," or closed state to the "up," or open state (Wrapp et al., 2020b). Such conformational changes occur on biological timescales generally not accessible by classical molecular dynamics simulations (Onuchic et al., 1997). To simulate the full unbiased path at atomic resolution, we used the weighted ensemble (WE) enhanced sampling method (Huber and Kim, 1996; Zuckerman and Chong, 2017). Instead of running one single long simulation, the WE method runs many short simulations in parallel along the chosen reaction coordinates. The trajectories that rarely sample high energy regions are replicated, while the trajectories that frequently sample low energy regions are merged, which makes sampling rare events computationally tractable and gives enhanced sampling. The trajectories also carry probabilities or weights, which are continuously updated, and there is no statistical bias added to the system. Hence, we are able to directly obtain both thermodynamic and kinetic properties from the WE simulations (Zhang et al., 2010).

For this study, the closed model of the glycosylated spike from Casalino et al. (2020b), was used as the initial structure by only keeping the head domain. The WE simulations were run using the highly scalable WESTPA software (Zwier et al., 2015), with the Amber GPU accelerated molecular dynamics engine (Götz et al., 2012; Salomon-Ferrer et al., 2013), version 18. Chamber (Crowley et al., 2009) was used to convert CHARMM36 (Guvench et al., 2009; Huang and Mackerell, 2013) force fields and parameters from the system developed by Casalino et al. (2020b) into an Amber readable format. A TIP3P (Jorgensen et al., 1983b) water box with at least 10 Å between protein and box edges was used with 150 mM NaCl, leading the total number of atoms to 490,621. Amber minimization was carried out in two stages. First the solvent was minimized for 10,000 cycles with sugars and proteins restrained with a weight of 100 kcal/mol Å^2, followed by unrestrained minimization for 100,000 cycles. Next the system was incrementally heated to 300 K over 300 ps. Equilibration and production were carried out in 2 fs timesteps with SHAKE (Ryckaert et al., 1977) constraints on non-polar hydrogens and NPT ensemble. Pressure and temperature were controlled with Monte Carlo barostat and Langevin thermostat with 1 ps^{-1} collision frequency. The particle-mesh Ewald (PME) method was used with 10 Å cutoff for non-bonded interactions. The system was first equilibrated for 21 ns of conventional MD. The RMSD of the alpha carbons began to level off around 16 ns, and 24 structures were taken at regular intervals between 16 and 21 ns to use as equally weighted basis states for the WE simulation.

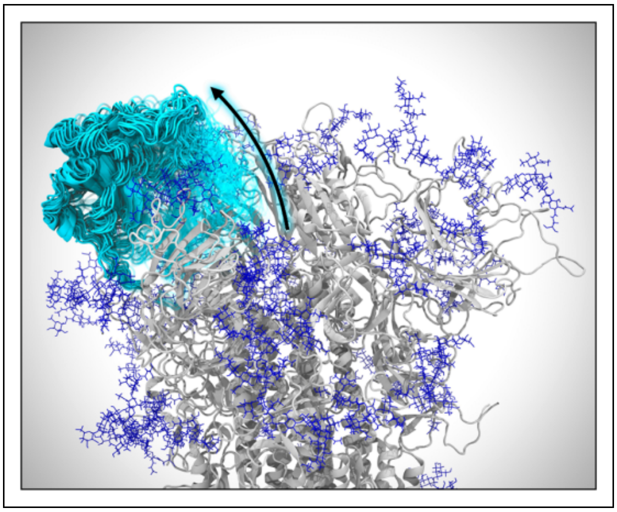

Figure 2. Opening of the spike protein. VMD visualization of weighted ensemble simulations shows the transition of the spike's RBD from the closed state to the open state. Many conformations of the RBD along its opening pathway are represented at the same time using cyan cartoons and a transparency gradient. Glycans appear as dark blue.

For each WE simulation, the fixed time interval for resampling was set to 100 ps followed by progress coordinate evaluation, splitting/merging of trajectories and updating trajectory weights, with a target of 8 trajectories per bin. A two dimensional progress coordinate was defined by (*i*): the distance between the center of mass (COM) of the alpha carbons in the structured region of the spike helical core, and the alpha carbons in the four main beta sheets of the RBD (refers to RBD from chain A unless otherwise specified) and (*ii*): the RMSD of the alpha carbons in the four main beta sheets of the RBD to the initial structure (obtained from 1 ns equilibration). This simulation was run for 8.77 days on 80 P100 GPUs on Comet at SDSC collecting a comprehensive sampling of ~7.5 µs, with bin spacing continuously monitored and adjusted to maximize sampling.

After extensive sampling of the RBD closed state, the second progress coordinate was changed to the RMSD of the alpha carbons in the four main beta sheets of the RBD compared to the final open structure, obtained from system 1, after 1 ns of equilibration carried out with identical methods as the closed structure described above, which was initially calculated as 11.5 Å. This allowed more efficient sampling of the transition to the open state by focusing sampling on states which are closer in rotational or translational space to the final state, rather than sampling all conformations that are distinctly different from the closed state. Bin spacing was continuously monitored and adjusted to maximize traversing the RMSD coordinate. The full transition was confirmed when the RMSD coordinate reached below 6 Å and the RBD COM coordinate reached above 8.5 Å (Figure 2). The simulation was stopped for analysis after 1099 iterations, upon running for 26.74 days on 100 V100 GPUs on Longhorn at TACC and harvesting ~70.0 μs.

A second, independent WE simulation was conducted to determine if the findings of the initial simulation were reproducible, and to use the information on the free energy landscape of the successful transition in the first WE to inform bin spacing and target state definition to run an unsupervised simulation. After 19.64 days on 100 V100 GPUs on TACC Longhorn and ~51.5 μs of comprehensive sampling, successful transitions to the open state were observed, as well as further open states, in which the RBD was observed to be peeling off of the spike core.

2.1.4. Two-parallel-membrane system of the spike-ACE2 complex. The SARS-CoV-2 virus gains entry into the host cell through a membrane fusion process taking place upon the recognition of the ACE2 receptors exposed on the host cell. This binding event triggers several, dramatic conformational changes within the spike protein, which becomes primed to pull the two membranes together for fusion, allowing the virus to pour the viral RNA into the host cell. In order to disentangle the mechanistic intricacies underlying this key process, we exploited the wealth of information obtained from the individual simulations described above to assemble an all-atom complex between the full-length spike and the ACE2 dimer. As a first step, equilibrated structures of the spike in the open state and of the ACE2-RBD complex were extracted from their respective individual simulations (Barros et al., 2020; Casalino et al., 2020a). Subsequently, the spike protein was superimposed onto the ACE2-RBD complex by aligning the spikes's RBD "up" with the RBD of the ACE2-RBD complex, allowing for a fairly vertical arrangement of the new construct. In order to preserve the best possible binding interface, the RBD of the spike was discarded, whereas the RBD from the ACE2-RBD complex was retained and linked to the rest of the spike. The spike-ACE2 complex was embedded into a double membrane system: the spike's transmembrane domain was inserted into a 330 Å × 330 Å ERGIC-like lipid bilayer, whereas for ACE2 a mammalian cellular membrane of the same dimension was used (Casares et al., 2019; Van Meer et al., 2008). The two membranes were kept parallel to each other, allowing the use of an orthorhombic box. In order to facilitate the water and ion exchange between the internal and external compartment, an outer-membrane-protein-G (OmpG) porin folded into a beta barrel was embedded into each membrane. The OmpG equilibrated model was obtained from Chen et al (Chen et al., 2008). The generated two-membrane construct was solvated with explicit TIP3P water molecules, with the total height of the external water compartment matching the internal one exhibiting a value of 380 Å. Sodium and chloride ions were added at a concentration of 150 mM to neutralize the charge and reshuffled to balance the charge between the two compartments.

The composite system, counting 8,562,698 atoms with an orthorhombic box of 330 Å × 330 Å × 850 Å, was subjected to all-atom MD simulation on the Summit computing system at ORNL using NAMD 2.14 (Phillips et al., 2020) and CHARMM36 all-atom additive force fields (Guvench et al., 2009; Huang and Mackerell, 2013). Two cycles of conjugate gradient energy minimization and NPT pre-equilibration were conducted using a 2 fs timestep for a total of ~3 ns. During this phase, the ACE2 and spike proteins and the glycans were harmonically restrained at 5 kcal/mol, allowing for the relaxation of the two lipid bilayers, the OmpG porins, water molecules and ions within the context of the double membrane system. We remark that the two lipid patches were previously equilibrated, therefore not requiring a melting phase at this stage. The dimension of the cell in the xy plane was maintained constant while allowing fluctuation along the z axis. Upon this initial pre-equilibration phase, a ~17 ns NPT equilibration was performed by releasing all the restraints, preparing the system for production run. From this point, three replicas were run or a total of ~522 ns comprehensive simulation time. By using the trained AI learning model, three conformations were extracted from this set of simulations, each of them representing a starting point of a new replica with re-initialized velocities. A total of three additional simulations were therefore performed, collecting ~180 ns and bringing the total simulation time to ~702 ns.

2.1.5. SARS-CoV-2 viral envelope. The full-scale viral envelope was constructed using the LipidWrapper program (v1.2) previously developed and described by Durrant and Amaro (2014). A 350 Å × 350 Å lipid bilayer patch used as the pdb input was generated using CHARMM-GUI with an ERGIC-like lipid composition and an estimated area per lipid of 63 Å. An icospherical mesh with a 42.5 nm radius, in accordance with experimentally-observed CoV-2 radii, was exported as a collada file from Blender (v2.79b) and used as the surface file (Ke et al., 2020).[1] LipidWrapper was run in a Python 2.7 conda environment with lipid headgroup parameters "_P, CHL1_O3," a lipid clash cut-off of 1.0 Å, and filling holes enabled.[2] The final bilayer pdb was solvated in a 110 nm cubic box using explicit TIP3P water molecules and neutralized with sodium and chloride ions to a concentration of 150 mM. The final system contained 76,134,149 atoms.

Since the LipidWrapper program operates via tessellation, lipid clash removal, and a subsequent lipid patching algorithm, the bilayer output attains a lower surface pressure than that of a bilayer of the same lipid composition at equilibrium (Casalino et al., 2020a). Due to this artifact, as the bilayer equilibrates, the lipids undergo lateral compression resulting in the unwanted formation of pores. Thus, the envelope was subjected to multiple rounds of minimization, heating, equilibration, and patching until the appropriate equilibrium surface pressure was reached.

All-atom MD simulations were performed using NAMD 2.14 and CHARMM36 all-atom additive force fields. The conjugate-gradient energy minimization procedure included two phases in which the lipid headgroups were restrained with 100 and 10 kcal/mol weights, respectively, at 310 K for 15,000 cycles each. The membrane was then melted by incremental heating from 25 K to 310 K over 300 ps prior to NPT equilibration. The equilibration sequentially released the harmonic restraints on the lipid headgroups from 100 to 0 kcal/mol over 0.5 ns. Following this sequence, the structure was visually evaluated to determine whether to continue equilibration or to proceed with pore patching. Most structures continued with unrestrained equilibration for 4–26 ns prior to patching, with longer unrestrained equilibrations attributed to later, more stable envelopes.

Patching of the envelope was done by overlapping the initial LipidWrapper bilayer output with the newly-equilibrated envelope. All superimposed lipids within 2.0 Å of the equilibrated lipids were removed to eliminate clashes. Superimposed lipids within 4.0 Å of an equilibrated cholesterol molecule were also removed to eliminate ring penetrations. The patched system, with new lipids occupying the pores, was then re-solvated, neutralized, and subjected to the next round of minimization, heating, and equilibration.

After ten rounds of equilibration and patching, 24 spike proteins with glycans, 8 in the closed and 16 in the open state, were inserted randomly on the envelope using a house tcl script. A random placement algorithm was used in accordance with experimental microscopy imaging which has suggested that there is no obvious clustering of the spikes and no correlation between RBD state and location on the spike surface (Ke et al., 2020). The number of spikes was selected based on experimental evidence reporting a concentration of 1000 spikes/nm ^{2} on the envelope (Ke et al., 2020). The new structure containing spikes was re-solvated, neutralized, and processed to remove clashing lipids prior to further simulation. The resulting cubic solvent box was 146 nm per side and contained 304,780,149 atoms. The spike-inclusive envelope was then subjected to three more equilibration and patching sequences. The final virion used for all-atom MD production runs had a lipid envelope of 75 nm in diameter with a full virion diameter of 120 nm. The complete equilibration of the viral envelope totaled 41 ns on the TACC Frontera system and 75 ns on ORNL Summit. Full-scale viral envelope production simulations were performed on Summit for a total of 84 ns in an NPT ensemble at 310 K, with a PME cutoff of 12 Å for non-bonded interactions.

3. Performance attributes

Performance Attribute	Our Submission
Category of achievement	Scalability, Time-to-solution
Type of method used	Explicit, Deep Learning
Results reported on the basis of	Whole application including I/O
Precision reported	Mixed Precision
System scale	Measured on full system
Measurement mechanism	Hardware performance counters, Application timers, Performance Modeling

4. Current state of the art

4.1. Parallel molecular dynamics

NAMD (Phillips et al., 2005) has been developed over more than two decades, with the goal of harnessing parallel computing to create a computational microscope (Lee et al., 2009; Shaw et al., 2007) enabling scientists to study the structure and function of large biomolecular complexes relevant to human health. NAMD uses adaptive, asynchronous, message-driven execution based on Charm++ (Kalé et al., 2019; Kalé and Zheng, 2013). It was one of the first scientific applications to make use of heterogeneous computing with GPUs (Phillips et al., 2008), and it implements a wide variety of advanced features supporting state-of-the-art simulation methodologies. Continuing NAMD and Charm++ developments have brought improved work decomposition and distribution approaches and support for low overhead hardware-specific messaging layers, enabling NAMD to achieve greater scalability on larger parallel systems (Kumar et al., 2012; Phillips et al., 2014). NAMD incorporates a collective variables module supporting advanced biasing methods and a variety of in-situ analytical operations (Fiorin et al., 2013). Simulation preparation, visualization, and post-hoc analysis are performed using both interactive and offline parallel VMD jobs (Humphrey et al., 1996; Stone et al., 2013a, 2013b, 2016b). NAMD has previously been used to study viruses and large photosynthetic complexes on large capability-oriented and leadership class supercomputing platforms, enabling the high-fidelity determination of the HIV-1 capsid structure (Zhao et al., 2013), the characterization of substrate binding in influenza (Durrant et al., 2020), and the structure and kinetics of light harvesting bacterial organelles (Singharoy et al., 2019).

4.2. Weighted ensemble MD simulations

The weighted ensemble (WE) method is an enhanced sampling method for MD simulations that can be orders of magnitude more efficient than standard simulations in generating pathways and rate constants for rare-event processes. WE runs many short simulations in parallel, instead of one long simulation, and directly gives both thermodynamic and kinetic properties. The simulations go through "resampling" where simulations are replicated in less-visited regions and merged in well-visited regions. The simulations also carry probabilities or "weights" that are continuously updated to ensure that no statistical bias is added to the system. In addition, the WE method is one of the few methods that can obtain continuous unbiased pathways between states, so this was the most suitable method for us to obtain and observe the closed to open transition for the spike system. Before the WE method was applied to the spike system under investigation here (about ~500,000 atoms), the largest system used for the WE method was the barnase-barstar complex (100,000 atoms) (Saglam and Chong, 2019).

4.3. AI-driven multiscale MD simulations

A number of approaches, including deep learning methods, have been developed for analysis of long timescale MD simulations (Noé, 2020). These linear, non-linear, and hybrid ML approaches cluster the simulation data along a small number of latent dimensions to identify conformational transitions between states (Bernetti et al., 2020; Ramanathan et al., 2012). Our group developed a deep learning approach, namely the variational autoencoder that uses convolutional filters on contact maps (from MD simulations) to analyze long time-scale simulation datasets and organize them into a small number of conformational states along biophysically relevant reaction coordinates (Bhowmik et al., 2018). We have used this approach to characterize protein conformational landscapes (Romero et al., 2019). However, with the spike protein, the intrinsic size of the simulation posed a tremendous challenge in scaling our deep learning approaches to elucidate conformational states relevant to its function.

Recently, we extended our approach to adaptively run MD simulation ensembles to fold small proteins. This approach, called DeepDriveMD (Lee et al., 2019), successively learns which parts of the conformational landscape have been sampled sufficiently and initiates simulations from undersampled regions of the conformational landscape (that also constitute "interesting" features from a structural perspective of the protein). While a number of adaptive sampling techniques exist (Allison, 2020; Bonati et al., 2019; Kasson and Jha, 2018; Lamim Ribeiro and Tiwary, 2019; Ribeiro et al., 2018; Wang et al., 2019, 2020), including based on reinforcement learning methods (Pérez et al., 2020), these techniques have been demonstrated on prototypical systems. In this paper, we utilize the deep learning framework to suggest additional points for sampling and do not necessarily use it in an adaptive manner to run MD simulations (mainly due to the limitations posed by the size of the system). However, extensions to our framework for enabling support of such large-scale systems are straightforward and further work will examine such large-scale simulations.

5. Innovations realized

5.1. Parallel molecular dynamics

Significant algorithmic improvements and performance optimizations have been required for NAMD to achieve high performance on the GPU-dense Summit architecture (Acun et al., 2019; Phillips et al., 2020; Stone et al., 2016a). New CUDA kernels for computing the short-range non-bonded forces were developed that implement a "tile list" algorithm for decomposing the workload into lists of finer grained tiles that more fully and equitably distribute work across the larger SM (streaming multiprocessor) counts in modern NVIDIA GPUs. This new decomposition uses the symmetry in Newton's Third Law to eliminate redundant calculation without incurring additional warp-level synchronization (Stone et al., 2016a). CUDA kernels also were added to offload the calculation of the bonded force terms and non-bonded exclusions (Acun et al., 2019). Although these terms account for a much smaller percentage of the work per step than that of the short-range non-bonded forces, NAMD performance on Summit benefits from further reduction of CPU workload. NAMD also benefits from the portable high-performance communication layer in Charm++ that communicates using the IBM PAMI (Parallel Active Messaging Interface) library, which improves performance by up to 20% over an MPI-based implementation (Acun et al., 2019; Kumar et al., 2012).

Additional improvements have benefited NAMD performance on Frontera. Recent developments in Charm++ now include support for the UCX (Unified Communication X) library which improves performance and scaling for Infiniband-based networks. Following the release of NAMD 2.14, a port of the CUDA tile list algorithm to Intel AVX-512 intrinsics was introduced, providing a 1.8× performance gain over the "Sky Lake" (SKX) builds of NAMD.

A significant innovation in NAMD and VMD has been the development of support for simulation of much larger system sizes, up to two billion atoms. Support for larger systems was developed and tested through all-atom modeling and simulation of the protocell as part of the ORNL CAAR (Center for Accelerated Application Readiness) program that provided early science access to the Summit system (Phillips et al., 2020). This work has greatly improved the performance and scalability of internal algorithms and data structures of NAMD and VMD to allow modeling of biomolecular systems beyond the previous practical limitation on the order of 250 million atoms. This work has redefined the practical simulation size limits in both NAMD and VMD and their associated file formats, added new analysis methods specifically oriented toward virology (Gonzalez-Arias et al., 2020), and facilitates modeling of cell-scaled billion-atom assemblies, while making smaller modeling projects significantly more performant and streamlined than before (Acun et al., 2019; Gonzalez-Arias et al., 2020; Phillips et al., 2020; Stone et al., 2016a, 2016b).

5.2. Multiscale molecular dynamics simulations

Often referred to as "computational microscopy," MD simulations are a powerful class of methods that enable the exploration of complex biological systems, and their time-dependent dynamics, at the atomic level. The systems studied here push state of the art in both their size and complexity. The system containing a full-length, fully-glycosylated spike protein, embedded in a realistic viral membrane (with composition that mimics the endoplasmic reticulum) contains essentially all of the biological complexity known about the SARS-CoV-2 spike protein. The composite system contains ~ 1.7 million atoms and combines data from multiple cryoEM, glycomics, and lipidomics datasets. The system was simulated with conventional MD out to microseconds in length, and several mutant systems were simulated and validated with independent experiments.

A related set of experiments utilizing the weighted ensemble method, an enhanced sampling technique, explored a truncated version of the spike protein ($\sim 600{,}000$ atoms with explicit solvent) in order to simulate an unbiased spike protein conformational transition from the closed to open state. This is the largest system, by an order of magnitude, that has been simulated using the WE method (biggest system until now was $\sim 60{,}000$ atoms). Using calculations optimized to efficiently make use of extensive GPU resources, we obtained several full, unbiased paths of the glycosylated spike receptor binding domain activation mechanism.

The second system increases the complexity by an order of magnitude by combining the spike system described above with a full-length, fully-glycosylated model of the ACE2 receptor bound into a host cell plasma membrane. This system represents the encounter complex between the spike and the ACE2 receptor, contains two parallel membranes of differing composition, has both the spike and ACE2 fully glycosylated, and forming a productive binding event at their interface. The composite system contains ~ 8.5 Million atoms with explicit water molecules and provides unseen views into the critical handshake that must occur between the spike protein and the ACE2 receptor to begin the infection cascade.

Our final system is of the SARS-CoV-2 viral envelope. This system incorporates 24 full-length, fully-glycosylated spike proteins into a viral membrane envelope of realistic (ER-like) composition, where the diameter of the viral membrane is ~ 80 nm and the diameter of the virion, inclusive of spikes, is 146 nm. Until now, the largest system disclosed in a scientific publication was the influenza virus, which contained ~ 160 million atoms. The SARS-CoV-2 viral envelope simulation developed here contains a composite 305 million atoms, and thus breaks new ground for MD simulations of viruses in terms of particle count, size, and complexity.

Moreover, typical state of the art simulations are run in isolation, presenting each as a self-contained story. While we also do that for each of the systems presented here, we advance on state of the art by using an AI-driven workflow that drives simulation at one scale, with knowledge gained from a disparate scale. In this way, we are able to explore relevant phase space of the spike protein more efficiently and in environments of increasing complexity.

5.3. Using AI for driving multiscale simulations

5.3.1. Using deep learning to characterize conformational states sampled in the SARS-CoV-2 spike simulations. MD simulations such as the ones described above generate tremendous amounts of data. For e.g. the simulations of the WE sampling of the spike protein's closed-to-open state generated over 100 terabytes of data. This imposes a heavy burden in terms of understanding the intrinsic latent dimensions along which large-scale conformational transitions can be characterized. A key challenge then is to use the raw simulation datasets (either coordinates, contact matrices, or other data collected as part of a standard MD runs) to cluster conformational states that have been currently sampled, to identify biologically relevant transitions between such states (e.g., open/closed states of spike), and suggest conformational states that may not be fully sampled to characterize these transitions (Ramanathan et al., 2012).

To deal with the size and complexity of these simulation datasets, approaches that analyze 3D point clouds are more appropriate. Indeed, such approaches are becoming more commonly utilized for characterizing protein binding pockets and protein-ligand interactions. We posited that such representations based on the C $^{\alpha}$ representation of protein structures could be viable to characterize large-scale conformational changes within MD simulation trajectories. We leverage the 3D PointNet based (Qi et al., 2017) adversarial autoencoder (3D-AAE) developed by Zamorski and colleagues (Zamorski et al., 2020) to analyze the spike protein trajectories. In this work, we employ the chamfer distance based reconstruction loss and a Wasserstein (Arjovsky et al., 2017) adversarial loss with gradient penalty (Gulrajani et al., 2017) to stabilize training. The original PointNet backbone treats the point cloud as unordered, which is true for general point clouds. In our case however, the protein is essentially a 1D embedding into a 3D space. This allows us to define a canonical order of points, i.e. the order in which they appear in the chain of atoms. For that reason, we increase the size-1 1D convolutional encoder kernels from the original PointNet approach to larger kernels up to size 5. This allows the network to not only learn features solely based on distance, but also based on *neighborhood* in terms of position of each atom in the chain. We found that a 4-layer encoder network with kernel sizes $[5, 3, 3, 1, 1]$ and filter sizes $[64, 128, 256, 256, 512]$ performs well for most tasks. A final dense layer maps the vectors into latent space with dimensionality 64. For the generator, we only use unit size kernels with filter dimensions $[64, 128, 512, 1024, 3]$ respectively (the output filter size is always the dimensionality of the problem). The discriminator is a 5 layer fully connected network with layer widths $[512, 512, 128, 64, 1]$.

The trajectories from the WE simulations were used to build a combined data set consisting of 130,880 examples. The point cloud data, representing the coordinates of the 3,375 backbone C $^{\alpha}$ atoms of the protein, was randomly split into training (80%) and validation input (20%) and was used to train the 3D-AAE model for 100 epochs using a batch size of 32. The data was projected onto a latent space of 64 dimensions constrained by a gaussian prior distribution with a standard deviation of 0.2. The loss optimization was performed with the Adam optimizer, a variant of stochastic gradient descent, using a learning rate of 0.0001. We also added hyperparameters to scale individual components of the loss. The reconstruction loss was scaled by 0.5 and the gradient penalty by a factor of 10.

The embedding learned from the 3D-AAE model summarizes a latent space that is similar to variational autoencoders, except that 3D-AAEs tend to be more robust to outliers within the simulation data. The embeddings learned from the simulations allow us to cluster the conformations (in an unsupervised manner) based on their similarity in overall structure, which can be typically measured using quantities such as root-mean squared deviations (RMSD).

We trained the model using several combinations of hyperparameters, mainly varying learning rate, batch size and latent dimension (Figure 3A). For visualizing and assessing the quality of the model in terms latent space structure, we computed t-SNE (van der Maaten and Hinton, 2008) dimensionality reductions on the high-dimensional embeddings from the validation set. A good model should generate clusters with respect to relevant biophysical observables not used in the training process. Therefore, we painted the t-SNE plot with the root mean squared deviation (RMSD) of each structure to the starting conformation and observed intelligible clustering of RMSD values (Figure 3B). We tested this model on a set of trajectories from the full scale spike-ACE2 system, using the same atom selection (3,375 C $^{\alpha}$ atoms) as the corresponding WE spike protein. We subsequently performed outlier detection using the local outlier factor (LOF) algorithm, which uses distance from neighboring points to identify anomalous data. The goal of the outlier detection step is to identify conformations of the protein that are most distinct from the starting structure, in order to story board important events in the transition of the protein from an open to closed conformation. Although the number of outlier conformations detected can be a parameter that the end-user can specify, we selected 20 outlier conformations, based on the extreme LOF scores. These conformations were visualized in VMD (Humphrey et al., 1996; Stone et al., 2016a), and further analyzed using tilt angles of the stalk and the RBD. The final selection included 3 structures which were used as the starting conformations for the next set of simulations (Figure 3C). These "outlier" conformers are cycled through additional MD simulations that are driven by the ML-methods.

6. How performance was measured

6.1. 3D-AAE

Since this application dominantly utilizes the GPU, we do not need to profile CPU FLOPs. Instead, we measure FLOPs for all precisions using the methodology explained in (Yang, 2020) with the NVIDIA NSight Compute 2020 GPU profiling tool. We collect floating point instructions of relevant flavors (i.e. adds, mults, fmas (fused multiply adds) and tensor core operations for FP16, FP32 and FP64) and multiply those with weighting factors of $\{1, 1, 2, 512\}$ respectively in order to transform those into FLOP counts. The sum of all these values for all precisions will yield our overall mixed precision FLOP count. To exclude FLOPs occuring during initialization and shutdown, we wrap the training iteration loop into start/stop profiler hooks provided by the NVIDIA CuPy Python package.[3]

6.2. NAMD

NAMD performance metrics were collected on TACC Frontera, using the Intel msr-tools utilities, with NAMD 2.14 with added Intel AVX-512 support. FLOP counts were measured for each NAMD simulation with runs of two different step counts. The results of the two simulation lengths were subtracted to eliminate NAMD startup operations, yielding an accurate estimate of the marginal FLOPs per step for a continuing simulation (Phillips et al., 2002).

FLOP counts were obtained by reading the hardware performance counters on all CPU cores on all nodes, using the rdmsr utility from msr-tools.[4] At the beginning of each job, the "TACC stats" system programs the core performance counters to count the 8 sub-events of the Intel FP_ARITH_INST_RETIRED.[5] Counter values are summed among the 56 cores in each node, and ultimately among each node. Each node-summed counter value is scaled by the nominal SIMD-width of the floating point instruction being counted and the 8 classes are added together to provide the total FLOP count per node. The hardware counters do not take masked SIMD instructions into account. SIMD lanes that are masked-out still contribute to the total FLOPs, however static analysis of the AVX-512-enabled NAMD executable showed that only 3.7% of FMA instructions were masked.

A breakdown of floating point instruction execution frequency for the AVX-512 build of NAMD across 2048 nodes is shown in Table 1. For CPU versions of NAMD, arithmetic is performed in double precision, except for single-precision PME long-range electrostatics calculations and associated FFTs. In the GPU-accelerated NAMD on Summit, single-precision arithmetic is used for both PME and also for short-range non-bonded force calculations, significantly increasing the fraction of single-precision instructions, at the cost of requiring a mixed-precision patch-center-based atomic coordinate representation to maintain full force calculation precision (Phillips et al., 2020; Stone et al., 2016a).

7. Performance results

7.1. 3D-AAE training performance

We used the aforementioned recipe for GPU profiling to determine the performance for the 3D-AAE training. We measure the FLOP counts individually for 2 training and 1 validation steps for a batch size of 32. The latent dimension of the model is a free hyperparameter and affects the FLOP count. We trained three models with latent dimensions $[32, 64, 128]$ in order to determine an optimal model for the task and thus we profile and report numbers for all of those. All models were trained for 100 epochs with batch size 32 on a single V100 GPU each. As mentioned above, the train/valdiation dataset split is 80%/20% and we do one

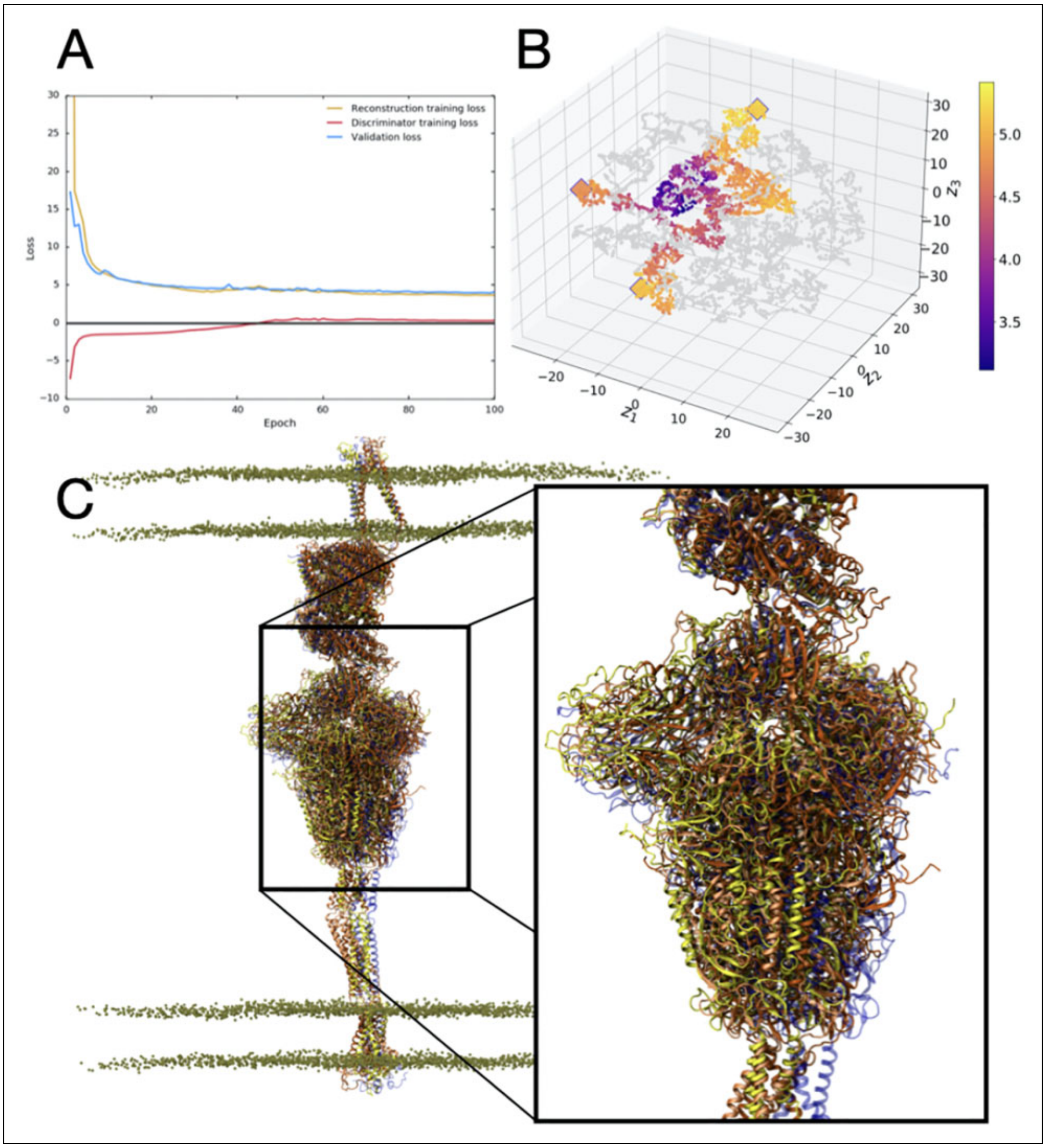

Figure 3. 3D-AAE training and test results. A) The loss progression for reconstruction, discriminator and validation loss over 100 epochs. B) The t-SNE plot visualization of the reduced latent space, with training embeddings represented in grey and test examples represented in color over the range of RMSD values. Outliers identified in the outlier detection stage are represented with an outlined diamond. C) VMD visualization of outlier structures (yellow, orange, dark orange) aligned and compared to the starting structure (blue). The inset highlights the relative motions between the Spike head region based on the outliers selected. Note the significant displacement of the head region from the initial structure (blue).

Table 1. NAMD AVX-512 FP operation breakdown.

FP Instr.	Ops	% total	FP Instr.	Ops	% total [t]
DblScalar	4.99e16	26.9%	SglScalar	2.09e15	1.1% [t]
Dbl128b	6.86e15	3.7%	Sgl128b	3.61e15	1.9%
Dbl256b	1.06e17	57.1%	Sgl256b	1.18e16	6.3%
Dbl512b	4.96e15	2.7%	Sgl512b	3.43e14	0.2%

validation pass after each training epoch. Thus, we can assume that this fraction translates directly into the FLOP counts for these alternating two stages. Our sustained performance numbers are computed using this weighted FLOP count average and the total run time. In order to determine peak performance, we compute the instantaneous FLOP rate for the fastest batch during training. Note that the 3D-AAE does exclusively use float (FP32) precision. The performance results are summarized in Table 2. Although the model is dense linear algebra heavy, it is also rather lightweight so it cannot utilize the full GPU and thus only delivering 20% of theoretical peak performance.

As expected, the peak performance is very consistent between the runs. The big difference in sustained performance between latent dim 64 and the other two models is that the frequency for computing the t-SNE was significantly

Table 2. 3D-AAE training performance on one V100 GPU.

Latent Dimensions	Peak TFLOP/s	Sustained TFLOP/s
32	2.96	0.97
64	3.16	2.28
128	3.13	0.91

Table 3. NAMD simulation floating point ops per timestep.

NAMD Simulation	Atoms	FLOPS/step [t]
ACE2-RBD complex	800 k	21.57 GFLOPS/step [t]
Single Spike	1.7 M	47.96 GFLOPS/step
Spike-ACE2 complex	8.5 M	243.7 GFLOPS/step
SARS-CoV-2 virion	305 M	8.3511 TFLOPS/step

reduced, i.e. from every epoch to every 5th. The t-SNE computation and plotting happens after each validation in a background thread on the CPU, but the training epochs can be much shorter than the t-SNE time. In that case, the training will stall until the previous t-SNE has completed. Evidently, decreasing the t-SNE frequency reduces that overhead significantly. We expect that the other models would perform similarly if we would have enabled this optimization for those runs as well. The remaining difference in peak vs. sustained performance can be explained by other overhead, e.g. storing embedding vectors, model checkpoints and the initial scaffolding phase. Furthermore, it includes the less FLOP-intensive validation phase whereas the peak estimate is obtained from the FLOP-heavy training phase.

7.2. NAMD simulation performance

Low-level NAMD performance measurements were made on the TACC Frontera system, to establish baseline counts of FLOPs per timestep for the four different biomolecular systems simulated as part of this work, summarized in Table 3, with the breakdown of CPU FLOPs described in Table 1. Sustained NAMD performance measurements were obtained using built-in application timers over long production science runs of several hours, including all I/O, and reported in units of nanoseconds per day of simulation time. NAMD sustained simulation performance for the spike-ACE2 complex is summarized for the TACC Frontera and ORNL Summit systems in Table 4 and Figure 4. NAMD sustained simulation performance, parallel speedup, and scaling efficiency are reported for the full SARS-CoV-2 virion in Table 5. Peak NAMD mixed-precision FLOP rates on ORNL Summit are estimated in Table 6 by combining sustained performance measurements with FLOPs/timestep measurements.

Table 4. NAMD performance: 8.5 M-atom Spike-ACE2.

Nodes	Frontera	Summit	Summit [t]
	CPU-only	CPU-only	CPU + GPU
64	7.52 ns/day	6.67 ns/day	52.15 ns/day [t]
128	13.00 ns/day	12.59 ns/day	79.68 ns/day
256	22.09 ns/day	24.19 ns/day	105.54 ns/day
512	34.32 ns/day	41.31 ns/day	135.31 ns/day
1024	41.88 ns/day	66.31 ns/day	162.22 ns/day

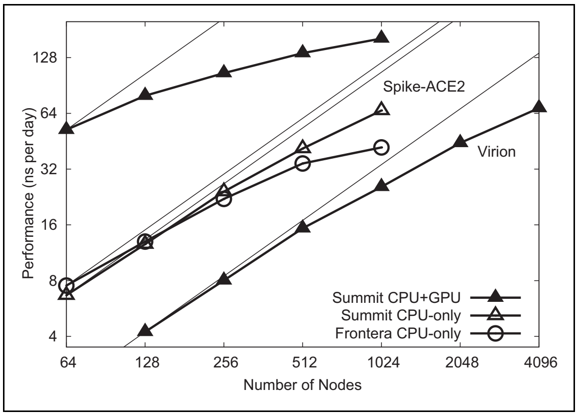

Figure 4. NAMD scaling on Summit and Frontera for 8.5 M-atom spike-ACE2 complex (upper lines) and 305 M-atom virion (lower line). Thin lines indicate linear scaling.

Table 5. NAMD performance: 305 M-atom virion.

Nodes	Summit	Speedup	Efficiency [t]
	CPU + GPU		
128	4.23 ns/day	~ 1.0×	~ 100% [t]
256	8.02 ns/day	1.9×	95%
512	15.32 ns/day	3.6×	91%
1024	25.66 ns/day	6.1×	75%
2048	44.27 ns/day	10.5×	65%
4096	68.36 ns/day	16.2×	51%

Table 6. Peak NAMD FLOP rates, ORNL Summit.

NAMD Simulation	Atoms	Nodes	Sim rate	Performance [t]
Spike-ACE2 complex	8.5 M	1024	162 ns/day	229 TFLOP/s [t]
SARS-CoV-2 virion	305 M	4096	68 ns/day	3.06 PFLOP/s

8. Implications

Our major scientific achievements are:

1. We characterize for the first time the glycan shield of the full-length SARS-CoV-2 spike protein (including the stalk), and find that two N-glycans linked to N165 and N234 have a functional role in modulating the dynamics of the spike's RBD. This unprecedented finding establishes a major new role of glycans in this system as playing an active role in infection, beyond shielding (Figure 1C) (Casalino et al., 2020b).
2. We discover that the human ACE2 receptor has a flexible hinge in the linker region near the membrane that

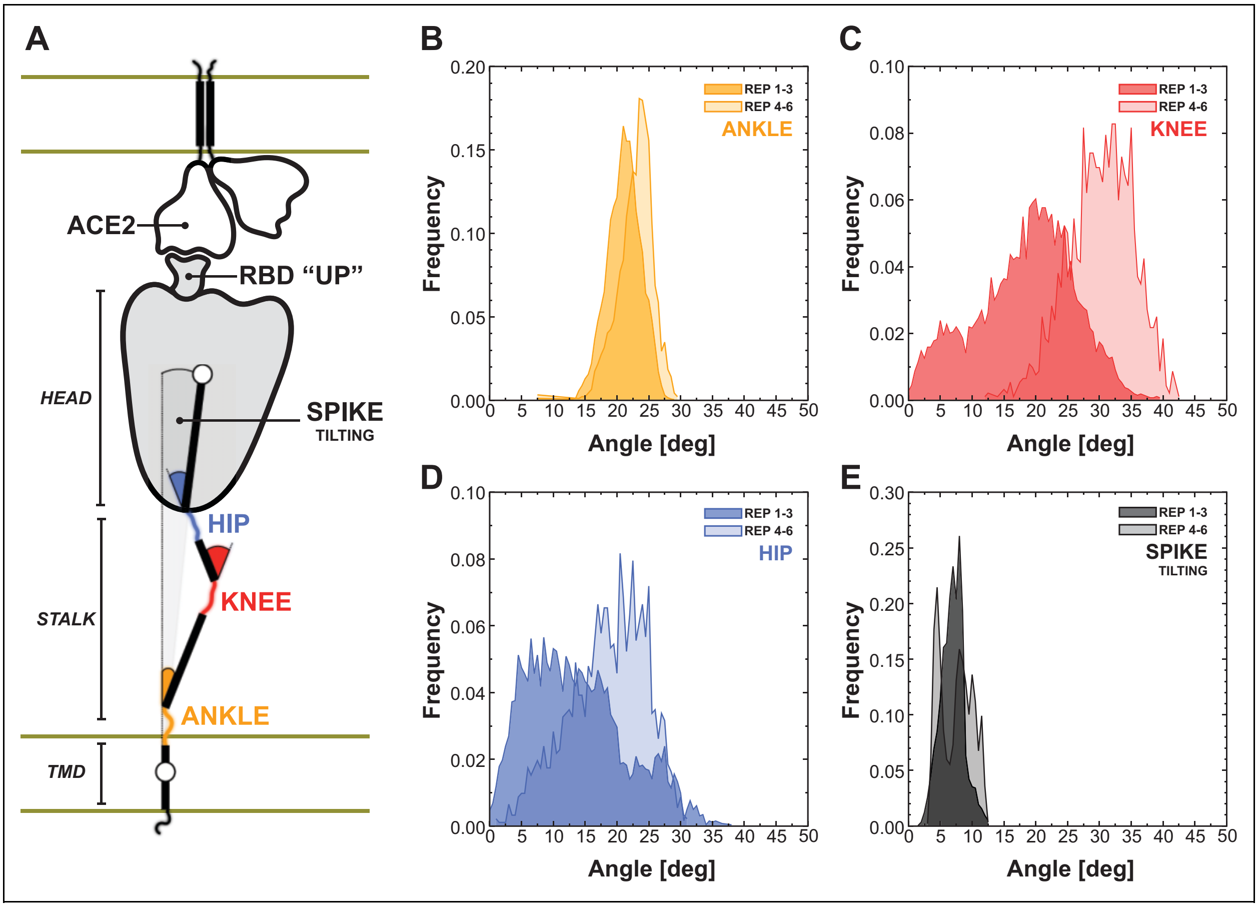

Figure 5. Flexibility of the spike bound to the ACE2 receptor. A) Schematic representation of the two-parallel-membrane system of the spike-ACE2 complex. (B–E) Distributions of the ankle, knee, hip and spike-tilting angles resulting from MD replicas 1–3 (darker color) and 4–6 (lighter color). Starting points for replicas 4–6 have been selected using DeepDriveMD.

enables it to undergo exceptionally large angular motions relative to the plane of the membrane. We predict this flexibility will aid forming productive complexes with the spike protein and may serve as a mechanical energy source during the cell fusion process (Barros et al., 2020).

3. We openly share our models, methods, and data, making them freely available to the scientific community. We are committed to the shared set of principles outlined in Ref. (Amaro and Mulholland, 2020): depositing findings as preprints in advance of formal peer review, making available our models at the time of deposition into a preprint server (Barros et al., 2020), and releasing the full datasets upon peer review (Casalino et al., 2020b). By doing so, the reproducibility and robustness of our findings and methods are enhanced, and the scientific findings from our simulations are amplified through reuse by others.
4. We describe for the first time unbiased pathways for the full closed-to-open transition of the spike's RBD (Figure **2**), where knowledge of this pathway has the potential to inform on mechanisms of viral infection as well as potentially aid in the discovery of novel druggable pockets within the spike. Our work set a new milestone for the use of the weighted ensemble method in biomolecular simulation, increasing applicable system size by an order or magnitude over current state of the art.
5. We characterize the spike's flexibility in the context of ACE2 binding. One of the most important properties of the spike protein is its intrinsic flexibility, a key feature that facilitates the interaction with the ACE2 receptors exposed on the host cell. CryoEM and cryoET structural data revealing the architecture of the SARS-CoV-2 viral particle showed that the spike can tilt up to 60° with respect to the perpendicular to the membrane (Ke et al., 2020; Yao et al., 2020). Behind this flexibility is the structural organization of the extra-virion portion of the spike, composed of two major domains, the stalk and the head, that are connected through a flexible junction that has been referred to as “hip” (Figure 5A) (Casalino et al., 2020b; Turoňová et al., 2020). Moreover, the stalk can be further divided into an upper and a lower leg, which correspond to the extra-virion alpha-helices of the coil-coiled trimeric bundle, and the transmembrane domain, which can be intended as the foot of this organizational scaffold. The stalk's upper leg, lower leg and the foot are interspersed by highly flexible loops

defined as “knee” and “ankle” junctions (Figure 5A) (Turoňová et al., 2020).

We then harnessed DeepDriveMD to perform adaptive MD on the Spike-ACE2 8.5 million atoms system. Following this workflow, we extracted three conformations from the first set of Spike-ACE2 MD simulations (replicas 1–3) and subsequently used them as starting points for a new round of MD (replicas 4–6). We then calculated the distribution of the overall spike tilting with respect to the perpendicular to the membrane (Figure 5E) and of other three angles involving the stalk, namely the “hip” angle between the stalk’s upper leg and the head (Figure 5B), the “knee” angle between the stalk’s lower and upper legs (Figure 5C), and the “ankle” angle between the perpendicular to the membrane and the stalk’s lower leg (Figure 5D).

The AI-driven adaptive MD approach expanded the conformational space explored, especially for the knee and hip angles, showing average values of 18.5° ± 7.7° and 13.8° ± 7.6° for replicas 1–3, shifted to 30.4° ± 5.1° and 18.8° ± 6.0° for the subsequent set of MD (replicas 4–6), respectively. The population shift is less pronounced for the ankle, exhibiting an average angle of 21.8° ± 2.7°. These results, in agreement with the data from Turoňova et al. (2020) that however did not consider the spike in complex with ACE2, reveal large hinge motions throughout the stalk and between the stalk and the head that accommodate the interaction between the spike’s RBD and the ACE2 receptor, preventing the disruption of the binding interface. This is further highlighted by the overall tilting of the spike that remains well defined around 7.3° ± 2.0° (Figure 5E), showing that the stalk’s inner hinge motions prevent a larger scale bending that could potentially disrupt the RBD-ACE2 interaction.

(6) **Our approach points to the very near term ability to accelerate the sampling of dynamical configurations of the complicated viral infection machinery within in the context of its full biological complexity using AI**. The enormous amount of data arising from MD and WE simulations of the single spike served to build and train an AI model using the variational autoencoder deep learning approach, which we demonstrate to accelerate dynamical sampling of the spike in a larger, more complex system (i.e., the two parallel membrane spike-ACE2 complex). Thus, the combination of the AI-driven workflows together with the groundbreaking simulations opens the possibility to overcome a current major bottleneck in the development and use of such ultra-large scale MD simulations, which relates to the efficient and effective sampling of the conformational dynamics of a system with so many degrees of freedom. The scientific implications of such a technological advance, in terms of understanding of the basic science of molecular mechanisms of infection as well as the development of novel therapeutics, are vast.

(7) **We establish a new high watermark for the atomic-level simulation of viruses with the simulation of the SARS-CoV-2 viral envelope, tallying 305 million atoms including explicit water molecules, and exhibiting a strong scaling on Summit**. The virion has a realistic ERGIC-like membrane, contains 24 fully glycosylated full-length spikes (in both the open and closed states) and replicates the spatial patterning and density of viral proteins as determined from cryoelectron tomography experiments (Ke et al., 2020). These groundbreaking simulations, just now in the process of being fully analyzed, set the stage for future work on SARS-CoV-2 that will be unprecedented in terms of their ability to more closely mimic realistic biological conditions. This includes, for example, the ability to explore the interactions of the virus with multiple receptors on the host cell, or multiple antibodies. It will allow researchers to explore the correlated dynamics of the molecular pieceparts on the surface of the virus and the host cell, and the effects of curvature on such behavior. It will be used as the ground-truth in the development of other simulation approaches, including coarse grained simulation methods, which are under development (Yu et al., 2020). It will aid in the development of methods related to the construction of complicated biological membranes (Gonzalez-Arias et al., 2020). And the list goes on.

(8) **We developed an AI-driven workflow as a generalizable framework for multiscale simulation**. Though we focus here on advances made relevant to COVID19, the methods and workflow established here will be broadly applicable to the multiscale simulation of molecular systems.

Authors’ note

This work used resources, services, and support from the COVID-19 HPC Consortium (https://covid19-hpc-consortium.org/), a private-public effort uniting government, industry, and academic leaders who are volunteering free compute time and resources in support of COVID-19 research. We dedicate this contribution to the memory of Klaus Schulten.

Acknowledgments

The authors thank D. Maxwell, B. Messer, J. Vermaas, and the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory supported by the DOE under Contract DE-AC05-00OR22725. We also thank the Texas Advanced Computing Center Frontera team, especially D. Stanzione and T. Cockerill, and for compute time made available through a Director’s Discretionary Allocation (NSF OAC-1818253). We thank the Argonne Leadership Computing Facility supported by the DOE under DE-AC02-06CH11357. NAMD and VMD are funded by NIH P41-GM104601. The NAMD team thanks Intel and M.

Brown for contributing the AVX-512 tile list kernels. Anda Trifan acknowledges support from a DOE CSGF (DE-SC0019323).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by NIH GM132826, NSF RAPID MCB-2032054, an award from the RCSA Research Corp., a UC San Diego Moore's Cancer Center 2020 SARS-COV-2 seed grant, to R.E.A. This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the US DOE Office of Science and the National Nuclear Security Administration. Research was supported by the DOE through the National Virtual Biotechnology Laboratory, a consortium of DOE national laboratories focused on response to COVID-19, with funding from the Coronavirus CARES Act. This work was supported by the NIH (1R01GM115805-01) to L.T.C.

ORCID iD

Emilia P Barros https://orcid.org/0000-0001-9755-260X
Surl-Hee Ahn https://orcid.org/0000-0002-3422-805X
Alexander Brace https://orcid.org/0000-0001-9873-9177
Rommie E Amaro https://orcid.org/0000-0002-9275-9553

Notes

1. https://www.blender.org/
2. https://docs.anaconda.com/
3. https://cupy.dev/
4. https://github.com/intel/msr-tools
5. https://github.com/TACC/tacc_stats

References

Acun B, Hardy DJ, Kale L, et al. (2019) Scalable molecular dynamics with NAMD on the Summit system. *IBM Journal of Research and Development* 62: 4:1–4:9.

Allison JR (2020) Computational methods for exploring protein conformations. *Biochemical Society Transactions* 48(4): 1707–1724.

Amaro RE and Mulholland AJ (2020) A community letter regarding sharing biomolecular simulation data for COVID-19. *Journal of Chemical Information and Modeling* 60(6): 2653–2656.

Arjovsky M, Chintala S and Bottou L (2017) Wasserstein gan. Available at : https://arxiv.org/abs/1701.07875

Barros EP, Casalino L, Gaieb Z, et al (2020) The flexibility of ACE2 in the context of SARS-CoV-2 infection. DOI: 10.1101/2020.09.16.300459.

Bernetti M, Bertazzo M and Masetti M (2020) Data-driven molecular dynamics: a multifaceted challenge. *Pharmaceuticals* 13(9): 253.

Bhowmik D, Gao S, Young MT, et al. (2018) Deep clustering of protein folding simulations. *BMC Bioinformatics* 19(18): 484.

Bonati L, Zhang YY and Parrinello M (2019) Neural networks-based variationally enhanced sampling. *Proceedings of the National Academy of Sciences* 116(36): 17641–17647.

Casalino L, Gaieb Z, Dommer AC, et al (2020a) Shielding and beyond: the roles of glycans in SARS-CoV-2 spike protein. DOI: 10.1101/2020.06.11.146522.

Casalino L, Gaieb Z, Goldsmith JA, et al (2020b) Beyond shielding: the roles of glycans in the SARS-CoV-2 spike protein. *ACS Central Science* 6(10): 1722–1734.

Casares D, Escribá PV and Rosselló CA (2019) Membrane lipid composition: effect on membrane and organelle structure, function and compartmentalization and therapeutic avenues. *International Journal of Molecular Sciences* 20(9): 2167.

Chen M, Khalid S, Sansom MSP, et al. (2008) Outer membrane protein G: engineering a quiet pore for biosensing. *Proceedings of the National Academy of Sciences* 105(17): 6272–6277.

Crowley MF, Williamson MJ and Walker RC (2009) CHAMBER: Comprehensive support for CHARMM force fields within the AMBER software. *International Journal of Quantum Chemistry* 109(15): 3767–3772.

Durrant JD and Amaro RE (2014) LipidWrapper: an algorithm for generating large-scale membrane models of arbitrary geometry. *PLoS Computational Biology* 10(7): e1003720.

Durrant JD, Kochanek SE, Casalino L, et al. (2020) Mesoscale all-atom influenza virus simulations suggest new substrate binding mechanism. *ACS Central Science* 6(2): 189–196.

Fiorin G, Klein ML and Hénin J (2013) Using collective variables to drive molecular dynamics simulations. *Molecular Physics* 111(22-23): 3345–3362.

Gonzalez-Arias F, Reddy T, Stone J, et al. (2020) Scalable analysis of authentic viral envelopes on FRONTERA. *Computing in Science and Engineering* 22: 11–20.

Götz AW, Williamson MJ, Xu D, et al. (2012) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation* 8(5): 1542–1555.

Gulrajani I, Ahmed F, Arjovsky M, et al. (2017) Improved training of wasserstein gans. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, December 2017, pp. 5769–5779.

Guvench O, Hatcher E, Venable RM, et al. (2009) CHARMM additive all-atom force field for glycosidic linkages between hexopyranoses. *Journal of Chemical Theory and Computation* 5(9): 2353–2370.

Huang J and Mackerell AD (2013) CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of Computational Chemistry* 34(25): 2135–2145.

Huber GA and Kim S (1996) Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophysical Journal* 70(1): 97–110.

Humphrey W, Dalke A and Schulten K (1996) VMD—visual molecular dynamics. *Journal of Molecular Graphics and Modelling* 14(1): 33–38.

Jo S, Kim T, Iyer VG, et al. (2008) CHARMM-GUI: a web-based graphical user interface for CHARMM. *Journal of Computational Chemistry* 29(11): 1859–1865.

Jo S, Song KC, Desaire H, et al. (2011) Glycan reader: automated sugar identification and simulation preparation for carbohydrates and glycoproteins. *Journal of Computational Chemistry* 32(14): 3135–3141.

Jorgensen WL, Chandrasekhar J, Madura JD, et al. (1983a) Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* 79(2): 926–935.

Jorgensen WL, Chandrasekhar J, Madura JD, et al. (1983b) Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* 79(2): 926–935.

Kalé L, Acun B, Bak S, et al. (2019) The Charm++ Parallel Programming System. DOI: 10.5281/zenodo.3370873.

Kalé LV and Zheng G (2013) Chapter 1: the Charm++ programming model. In: LV Kale, and A Bhatele (eds) *Parallel Science and Engineering Applications: The Charm++ Approach*, 1st ed, chapter 1. Boca Raton, FL: CRC Press, Inc, pp. 1–16.

Kasson PM and Jha S (2018) Adaptive ensemble simulations of biomolecules. *Current Opinion in Structural Biology* 52: 87–94. Cryo electron microscopy: the impact of the cryo-EM revolution in biology • Biophysical and computational methods—Part A.

Ke Z, Oton J, Qu K, et al (2020) Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature* 588: 498–502.

Kumar S, Mamidala AR, Faraj DA, et al. (2012) PAMI: a parallel active message interface for the Blue Gene/Q supercomputer. In: *2012 IEEE 26th International Parallel and Distributed Processing Symposium*, Shanghai, China, 21–25 May 2012, pp. 763–773. DOI: 10.1109/IPDPS.2012.73.

Lamim Ribeiro JM and Tiwary P (2019) Toward achieving efficient and accurate ligand-protein unbinding with deep learning and molecular dynamics through rave. *Journal of Chemical Theory and Computation* 15(1): 708–719.

Lee EH, Hsin J, Sotomayor M, et al. (2009) Discovery through the computational microscope. *Structure* 17: 1295–1306.

Lee H, Turilli M, Jha S, et al. (2019) DeepDriveMD: deep-learning driven adaptive molecular simulations for protein folding. In: *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*. New York, NY: IEEE, pp. 12–19.

Noé F (2020) *Machine Learning for Molecular Dynamics on Long Timescales*. Cham: Springer International Publishing, pp. 331–372.

Onuchic JN, Luthey-Schulten Z and Wolynes PG (1997) Theory of protein folding: the energy landscape perspective. *Annual Review of Physical Chemistry* 48(1): 545–600.

Park SJ, Lee J, Qi Y, et al. (2019) CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates. *Glycobiology* 29(4): 320–331.

Pérez A, Herrera-Nieto P, Doerr S, et al. (2020) AdaptiveBandit: a multi-armed bandit framework for adaptive sampling in molecular simulations. *Journal of Chemical Theory and Computation* 16(7): 4685–4693.

Phillips J, Zheng G, Kumar S, et al. (2002) NAMD: biomolecular simulation on thousands of processors. In: *Proceedings of the IEEE/ACM SC2002 Conference*, *Technical Paper 277*, November 2002. Baltimore, MD: IEEE Press, pp. 1–18.

Phillips JC, Braun R, Wang W, et al. (2005) Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* 26: 1781–1802.

Phillips JC, Hardy DJ, Maia JDC, et al (2020) Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics* 153: 044130.

Phillips JC, Stone JE and Schulten K (2008) Adapting a message-driven parallel application to GPU-accelerated clusters. In: *SC '08: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing*, November 2008. Piscataway, NJ: IEEE Press, pp. 1–9. ISBN 978-1-4244-2835-9.

Phillips JC, Sun Y, Jain N, et al. (2014) Mapping to irregular torus topologies and other techniques for petascale biomolecular simulation. In: *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, SC '14. New York, NY: IEEE Press, pp. 81–91. DOI: 10.1109/SC.2014.12.

Qi CR, Su H, Mo K, et al. (2017) PointNet: deep learning on point sets for 3D classification and segmentation. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017.

Ramanathan A, Savol AJ, Burger VM, et al. (2012, January) Statistical inference for big data problems in molecular biophysics. In *Neural Information Processing Systems: Workshop on Big Learning*.

Ribeiro JML, Bravo P, Wang Y, et al. (2018) Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *The Journal of Chemical Physics* 149(7): 072301.

Romero R, Ramanathan A, Yuen T, et al (2019) Mechanism of glucocerebrosidase activation and dysfunction in Gaucher disease unraveled by molecular dynamics and deep learning. *Proceedings of the National Academy of Sciences* 116(11): 5086–5095.

Ryckaert JP, Ciccotti G and Berendsen HJC (1977) Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics* 23(3): 327–341.

Saglam AS and Chong LT (2019) Protein–protein binding pathways and calculations of rate constants using fully-continuous, explicit-solvent simulations. *Chemical Science* 10(8): 2360–2372.

Šali A and Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology* 234(3): 779–815.

Salomon-Ferrer R, Götz AW, Poole D, et al. (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *Journal of Chemical Theory and Computation* 9(9): 3878–3888.

Shajahan A, Supekar NT, Gleinich AS, et al. (2020) Deducing the N- and O-glycosylation profile of the spike protein of novel coronavirus SARS-CoV-2. *Glycobiology* 2020: 1–8. DOI: 10.1093/glycob/cwaa042/5826952.

Shaw D, Deneroff M, Dror R, et al (2007) Anton, a special-purpose machine for molecular dynamics simulation. *ACM SIGARCH Computer Architecture News* 35: 1–12.

Singharoy A, Maffeo C, Delgardo K, et al (2019) Atoms to phenotypes: molecular design principles of cellular energy metabolism. *Cell* 179: 1098–1111.

Starr TN, Greaney AJ, Hilton SK, et al (2020) Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182(5): 1295–1310.e20.

Stone JE, Hynninen AP, Phillips JC, et al. (2016a) Early experiences porting the NAMD and VMD molecular simulation and analysis software to GPU-accelerated OpenPOWER platforms. *International Workshop on OpenPOWER for HPC (IWOPH'16)* 9945: 188–206.

Stone JE, Isralewitz B and Schulten K (2013a) Early experiences scaling VMD molecular visualization and analysis jobs on Blue Waters. In: *Extreme Scaling Workshop (XSW), 2013*. New York, NY: IEEE, pp. 43–50.

Stone JE, Sener M, Vandivort KL, et al. (2016b) Atomic detail visualization of photosynthetic membranes with GPU-accelerated ray tracing. *Parallel Computing* 55: 17–27.

Stone JE, Vandivort KL and Schulten K (2013b) GPU-accelerated molecular visualization on petascale supercomputing platforms. In: *Proceedings of the 8th International Workshop on Ultrascale Visualization*, UltraVis '13. New York, NY: ACM, pp. 6:1–6:8.

Sun Z, Ren K, Zhang X, et al. (2020) Mass spectrometry analysis of newly emerging coronavirus HCoV-19 spike protein and human ACE2 reveals camouflaging glycans and unique post-translational modifications. *Engineering*. DOI: 10.1016/j.eng.2020.07.014.

Turoňová B, Sikora M, Schürmann C, et al. (2020) In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges. *Science*: eabd5223. DOI: 10.1126/science.abd5223.

van der Maaten L and Hinton G (2008) Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research* 9: 2579–2605.

Van Meer G, Voelker DR and Feigenson GW (2008) Membrane lipids: where they are and how they behave. *Nature Reviews Molecular Cell Biology* 9: 112–124.

Walls AC, Park YJ, Tortorici MA, et al. (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* 181(2): 281–292.e6.

Wang Y, Ribeiro JML and Tiwary P (2019) Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature Communications* 10(1): 3573.

Wang Y, Ribeiro JML and Tiwary P (2020) Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current Opinion in Structural Biology* 61: 139–145.

Watanabe Y, Allen JD, Wrapp D, et al. (2020) Site-specific glycan analysis of the SARS-CoV-2 spike. *Science (New York, N. Y.)* 369: 330–333.

Wrapp D, Wang N, Corbett KS, et al (2020a) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science (New York, N.Y.)* 1263: 1260–1263.

Wrapp D, Wang N, Corbett KS, et al. (2020b) Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367(6483): 1260–1263.

Wu EL, Cheng X, Jo S, et al. (2014) CHARMM-GUI membrane builder toward realistic biological membrane simulations. *Journal of Computational Chemistry* 35(27): 1997–2004.

Yan R, Zhang Y, Li Y, et al. (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367(6485): 1444–1448.

Yang C (2020) Hierarchical roofline analysis: How to collect data using performance tools on intel cpus and nvidia gpus.

Yao H, Song Y, Chen Y, et al (2020) Molecular architecture of the SARS-CoV-2 Virus. *Cell* 183(3): 730–738.e13.

Yu A, Pak AJ, He P, et al. (2020) A multiscale coarse-grained model of the SARS-CoV-2 virion. DOI: 10.1101/2020.10.02.323915.

Zamorski M, Zięba M, Klukowski P, et al (2020) Adversarial autoencoders for compact representations of 3D point clouds. *Computer Vision and Image Understanding* 193: 102921.

Zhang BW, Jasnow D and Zuckerman DM (2010) The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures. *The Journal of Chemical Physics* 132(5): 054107.

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9(1): 1–8.

Zhao G, Perilla JR, Yufenyuy EL, et al (2013) Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 497: 643–646.

Zhao P, Praissman JL, Grant OC, et al (2020) Virus-receptor interactions of glycosylated SARS-CoV-2 spike and human ACE2 receptor. *Cell Host and Microbe* 28(4): 586–601.e6.

Zuckerman DM and Chong LT (2017) Weighted ensemble simulation: review of methodology, applications, and software. *Annual Review of Biophysics* 46: 43–57.

Zwier MC, Adelman JL, Kaus JW, et al (2015) WESTPA: an interoperable, highly scalable software package for weighted ensemble simulation and analysis. *Journal of Chemical Theory and Computation* 11(2): 800–809.

Author biographies

Lorenzo Casalino is a postdoctoral research scholar under the supervision of Prof. Rommie Amaro at University of California San Diego, where he is currently working on SARS-CoV-2 and influenza viruses using computer simulations. Lorenzo received his M.Sc. in 2013 in

Pharmaceutical Chemistry and Technologies at University of Milan (Italy), and he earned his Ph.D. in 2017 in Computational Biophysics at SISSA, Trieste (Italy), under the guidance of Dr. Alessandra Magistrato. During his Ph.D. he spent six months in the group of Prof. Ursula Rothlisberger at EPFL (Switzerland). Lorenzo is author of more than 15 publications and more than 10 covers on international scientific journals, with his recent work being featured in the New York Times, Forbes and Scientific American.

Abigail C Dommer is a graduate student working toward a computational chemistry PhD in the laboratory of Dr. Rommie Amaro at the University of California, San Diego. She earned her A.B. in Chemistry from Washington University in St. Louis in 2014 and her M.S. in Physical Chemistry at UCSD in 2016. Abigail is a researcher in the NSF Center for Aerosol Impacts on Chemistry of the Environment, where she uses molecular dynamics to probe the impacts of ocean biology on the climate-relevant properties of marine aerosols. Her work focuses primarily on understanding the behavior of organic acids on the sea spray aerosol surface and in organic-aqueous phase separation. Through the Amaro lab, Abigail has also worked on various large-scale molecular dynamics projects to understand the role of lipids in viruses.

Zied Gaieb, as a postdoctoral research scholar at University of California San Diego, worked under the supervision of Professors Rommie E. Amaro and Michael K. Gilson. There, he focused on the structural characterization of the SARS-CoV-2 Spike protein; and pioneered state-of-the-art solutions for automating multi-target drug design and analysis. He authored more than 15 publications including his latest work on the Coronavirus that was featured in numerous prestigious media outlets including WIRED, Forbes, KBPS, Scientific American, and The New York Times. Zied was awarded his Ph.D. in Bioengineering at the University of California Riverside under Professor Dimitrios Morikis, and is currently the co-founder and Senior VP of Computation at Athae Bio Inc.

Emilia P Barros recently started a position as a postdoctoral scholar in the lab of Prof. Sereina Riniker at ETH Zurich. Previous to that, she earned her PhD in Computational Chemistry at the lab of Prof. Rommie Amaro at the University of California San Diego, where she focused on using molecular dynamics and other computational methods to understand and predict the effect of sequence mutations in protein dynamics and function. She also conducted research looking into the intrinsic dynamic of the ACE2 receptor and how that affects interactions with SARS-CoV-2 Spike protein. She received her M. Sc. in Physical Chemistry and B.S. in Chemistry from the University of Campinas, Brazil.

Terra Sztain is an NSF Graduate Research Fellow at the University of California, San Diego. She is a PhD candidate in the areas of Biochemistry and Biophysics, under the advisement of Prof. Michael Burkart and Prof. J. Andrew McCammon. Her research focuses on the integration of experimental and computational techniques to understand proteins structure and dynamics to manipulate their functions. She primarily iterates between NMR spectroscopy and conventional and enhanced sampling molecular dynamics simulations to make predictions and validations. Terra received her B.S. in Biochemistry from the University of California, Los Angeles in 2016, graduating *cum laude.*

Surl-Hee Ahn is a postdoctoral scholar in the Department of Chemistry and Biochemistry at the University of California San Diego, advised under Prof. J. Andrew McCammon and Prof. Rommie Amaro. She has received her B.A. in Biochemistry and Mathematics, M.A. in Mathematics, and M.S. in Chemistry at the University of Pennsylvania, and her Ph.D. in Chemistry (Chemical Physics) at Stanford University, advised under Prof. Eric Darve. Surl-Hee is interested in developing enhanced sampling methods for molecular dynamics simulations and applying those methods to study important biophysical phenomena. She was selected to participate in the 2018 MIT Rising Stars in Mechanical Engineering Workshop.

Anda Trifan is a Biophysics and Quantitative Biology graduate student in the Theoretical and Computational Biophysics Group at the University of Illinois, Urbana-Champaign. She has been working with Argonne National Laboratory since January 2020 for her practicum and currently serves as a Research Assistant under the supervision of Prof. Arvind Ramanathan. She holds the Computational Science Graduate Student Fellowship from the Department of Energy. She is interested in using machine learning and artificial intelligence techniques to study complex biophysical phenomena. She obtained her B.S. in Chemistry from DePaul University in 2012.

Alexander Brace is a Research Associate in the Data Science and Learning Division at Argonne National Laboratory. He received his B.S.E in Computer Science (2020) from the University of Michigan, Ann Arbor. Alexander is interested in multidisciplinary research applying artificial intelligence to problems in biology through automated science.

Anthony T Bogetti is a graduate student at the University of Pittsburgh working toward a Ph.D. in Chemistry with Prof. Lillian T Chong. He received his B.S. in Chemistry (2017) from Messiah University. Anthony is interested in applying the weighted ensemble path sampling strategy to the simulation of chemical reactions with QM/MM potentials. He also works to improve the underlying methodology and usability of the weighted ensemble strategy.

Austin Clyde is a Ph.D. student in computer science at the University of Chicago and a research assistant in the Data Science and Learning Division at Argonne National Laboratory. He obtained his M.S. in computer science from the University of Chicago. His research interests involve the combination of drug discovery, high-performance computing, and mathematics.

Heng Ma is a postdoctoral appointee at Data Science and Learning Division in Argonne National Laboratory. He received B.S. from Beijing Institute of Technology, PhD from Lamar University, Texas, and later joined Oak Ridge National Laboratory for postdoctoral training. Heng is interested in Machine Learning/Artificial Intelligence applications to molecular biophysics through workflow automation and high performance computing.

Hyungro Lee is a Research Associate at Rutgers, the State University of New Jersey, under supervision of Prof. Shantenu Jha, where he has worked on performance characterization for ML based applications by profiling and resource analysis. His research focuses on improving resource efficiency of large-scale workflow on HPC systems. He received his PhD in Computer Science from Indiana University in 2019.

Syma Khalid is graduated with a first class degree in Chemistry from the University of Warwick in 2000. She remained at Warwick to read for a PhD under the supervision of Prof. P. Mark Rodger. After obtaining her PhD in 2003, she moved to the University of Oxford as a postdoc in Prof Mark Sansom's lab, to study the structure-function relationship of bacterial outer membrane proteins. In 2007, she was appointed as RCUK fellow in chemical biology at the University of Southampton. In 2010, she was appointed to a full lectureship at Southampton. In 2012, she was promoted to Senior lecturer, and to Professor in 2016. She specializes in membrane simulations that mirror complexity found in real-world systems.

Matteo Turilli is Assistant Research Professor at Rutgers University, Electrical & Computer Engineering Department. His main research interest is the design and development of distributed systems to support the computing requirements of diverse science drivers at scale. He focuses on merging high-throughput and high-performance computing, enabling the execution of heterogeneous workflows on among the largest computing platforms in the world. He holds a DPhil in Computer Science from the University of Oxford, UK.

Lillian T Chong is an Associate Professor in the Department of Chemistry at the University of Pittsburgh. She earned a B.S. in Chemistry at the Massachusetts Institute of Technology, a Ph.D. in Biophysics with Prof. Peter Kollman from the University of California in San Francisco, and completed her postdoctoral work with Prof. Vijay Pande at Stanford University and Dr. William Swope at the IBM Almaden Research Center. She is the recipient of a Silicon Therapeutics Fellowship, National Science Foundation CAREER Award, Carnegie Science Emerging Female Scientist Award, and Hewlett-Packard Outstanding Junior Faculty Award.

Carlos Simmerling obtained his BS (1991) and PhD (1994) in Chemistry at the University of Illinois at Chicago, developing new methods for computer modeling of biomolecules. He did a post-doctoral fellowship in Pharmaceutical Chemistry at UCSF, where he became a lead developer of the Amber biomolecular simulation software, used in thousands of research labs worldwide. In 1998 he joined the Chemistry department at Stony Brook University, where he is currently a Professor and Associate Director of the Laufer Center for Physical & Quantitative Biology. His research, funded by the National Institutes of Health, National Science Foundation, and Department of Energy, focuses on development of improved molecular simulation methods and using these tools to study biomolecular recognition mechanisms. His articles on improving the physics underlying biomolecular modeling have been cited nearly 10,000 times. Prof. Simmerling is currently the Marsha Laufer Chair of Physical & Quantitative Biology at Stony Brook University.

David J Hardy is a Senior Research Programmer in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign, where he has since 2018 been the lead developer of the parallel molecular dynamics application NAMD. He received his M.S. in Computer Science from the Missouri University of Science and Technology and his Ph.D. in Computer Science from the University of Illinois at Urbana-Champaign. His research interests include fast methods for calculating electrostatics, numerical methods for time integration, GPU computing, and parallel computing.

Julio DC Maia is a Senior Research Programmer at the NIH Center for Macromolecular Modeling and Bioinformatics. Julio is currently maintaining the highly-scalable molecular dynamics package NAMD by developing high-performance algorithms for graphics processing units, both on small machines as well as on large supercomputers. He has also worked on optimizing quantum chemistry packages on GPUs, as well as developing linear scaling electronic structure methods for large molecular systems.

James C Phillips is a Senior Research Programmer leading quality assurance and testing in the NCSA Blue Waters Project Office at the University of Illinois at Urbana-Champaign. He has a Ph.D. in Physics from the University of Illinois. From 1999 to 2017 James was the lead developer of the highly scalable parallel molecular dynamics program NAMD, for which he received a Gordon Bell Award in 2002. His research interests include improving the performance and accuracy of biomolecular and other simulations through parallelization,optimization, hardware acceleration, better algorithms, and new methods.

Thorsten Kurth works at NVIDIA on optimizing scientific codes for GPU based supercomputers. His main focus is on providing optimized deep learning applications for HPC systems. These include end-to-end optimizations such as input pipeline including IO tuning, distributed training and data visualization. Before he joined NVIDIA, Thorsten worked at NERSC with the application readiness team to deliver optimized codes for the NERSC HPC infrastructure. He was leading the Learning application category of the NERSC Exascale Science Application Program (NESAP), targeting at improving experimental and observational data analysis or simulation codes using machine learning and artificial intelligence methods. In 2018 he was awarded the Gordon Bell Prize for the first Deep Learning application which achieved more than 1 ExaOp peak performance on the OLCF Summit HPC system.

Abraham C Stern is a senior data scientist on the solution architecture & engineering team with NVIDIA. His interests lie at the intersection of scientific computing and machine learning, especially as applied to problems in the chemistry and materials science domain. Abe obtained his Ph.D. in computational chemistry from the University of South Florida and was previously a postdoctoral scholar at the University of California, Irvine.

Lei Huang is a research member in Texas Advanced Computing Center. Huang obtained his Ph.D. in Chemistry in 2008 from the University of Texas at Austin, then have worked as a postdoctoral fellow at Harvard University and a research assistant professor at University of Chicago. He has worked on diverse projects spanning from material science, biophysics to high performance computing and machine learning. He is a developer for widely used modeling/simulation software including CHARMM and NAMD. He also developed a scalable distributed high performance file system, FanStore, and many other tools for debugging and profiling.

John D McCalpin is a research scientist at the Texas Advanced Computing Center of the University of Texas at Austin, where he runs the Advanced Computing Evaluation Laboratory—evaluating performance and system architecture for current and future high performance computing systems. Before joining the university, McCalpin worked in performance analysis in the processor and system design teams at SGI, IBM, and AMD. He received his B.S. in Physics and M.S. in Physical Oceanography from Texas A&M University and his Ph.D. in Physical Oceanography from the Florida State University. He has developed and maintained the STREAM benchmark since 1991.

Mahidhar Tatineni received his M.S. & Ph.D. in Aerospace Engineering from UCLA and his B. Tech degree in Aerospace Engineering from IIT Madras, India. He currently leads the User Services group at SDSC. He has worked on many NSF funded optimization and parallelization research projects such as petascale computing for magnetosphere simulations, MPI performance tuning frameworks, hybrid programming models, topology aware communication and scheduling, big data middleware, and application performance evaluation using next generation communication mechanisms for emerging HPC systems. He is Co-PI on multiple NSF grants including the SDSC Comet and Expanse HPC systems at SDSC.

Tom Gibbs is currently responsible for strategy and implementation of programs to enable and promote developers and researchers to take full advantage of NVIDIA technology to pursue Grand Challenge Science Problems. Tom brings over 40 years of experience in HPC, and has applications expertise in industries ranging from Physics, Aerospace, Healthcare, Life Sciences, Energy and Financial Services. Prior to NVIDIA Tom held senior management positions for early-stage cloud startup companies in the healthcare market segment. He spent 15 years with the Intel Corporation, where he managed a global team responsible for leading innovation programs at CERN, NCSA, British Petroleum and Morgan Stanley as Director of Strategy and Architecture in the Solutions Group. During his time at Intel Tom was part of the HPC Business Unit responsible for ASCI RED and other large scale computing systems.

Tom was a past Chairman of the Open Grid Forum and a member of the Center for Excellence in Supply Chain Management at MIT.

John E Stone is a Senior Research Programmer and lead developer of VMD, a high performance biomolecular visualization and analysis tool used by over 100,000 researchers worldwide. His research interests include molecular and immersive visualization, GPU and parallel computing, and ray tracing. He provides software architecture and GPU-accelerated algorithm design leadership for the NAMD and Lattice Microbes simulation tools. He is a member of the Khronos ANARI and Vulkan visualization and rendering standards efforts. Mr. Stone has been honored as an NVIDIA CUDA Fellow (2010), and an IBM Champion for POWER (2017), for innovative thought leadership in the technical community. He holds an M.Sc. in Computer Science, from the Missouri University of Science and Technology. He is a consultant for projects involving computer graphics, GPU computing, and HPC.

Shantenu Jha is a Professor of Computer Engineering at Rutgers-New Brunswick, and also leads the Computation and Data Driven Discovery (C3D) department at Brookhaven National Laboratory. His research interests are at the intersection of high-performance distributed computing and computational & data science. More details can be found at http://radical.rutgers.edu.

Arvind Ramanathan is a computational biologist in the Data Science and Learning Division at Argonne National Laboratory and a senior scientist within the University of Chicago Consortium for Advanced Science and Technology. He obtained his Ph.D. from Carnegie Mellon University in computational biology and completed his postdoctoral training at Oak Ridge National Laboratory. He is the recipient of the UT-Battelle Early Career Researcher Award and the Outstanding Mentor Award at Oak Ridge National Laboratory. His research interests are at the intersection of data science, high performance computing and biological/biomedical sciences.

Rommie E Amaro holds the Distinguished Professorship in Theoretical and Computational Chemistry at the Department of Chemistry and Biochemistry at the University of California, San Diego. She received her B.S. in Chemical Engineering (1999) and her Ph.D. in Chemistry (2005) from the University of Illinois at Urbana-Champaign. Rommie was a NIH postdoctoral fellow with Prof. J. Andrew McCammon at UC San Diego from 2005-2009, and started her independent lab in 2009. She is the recipient of an NIH New Innovator Award, the Presidential Early Career Award for Scientists and Engineers, the ACS COMP OpenEye Outstanding Junior Faculty Award, the ACS Kavli Foundation Emerging Leader in Chemistry, and the Corwin Hansch Award.