Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action

Anda Trifan¹,²†, Defne Gorgun¹,²†, Zongyi Li³†, Alexander Brace¹,⁴†, Maxim Zvyagin¹,⁴†, Heng Ma¹, Austin Clyde¹,⁴, David Clark⁵, Michael Salim¹, David J. Hardy², Tom Burnley⁶, Lei Huang⁷, John McCalpin⁷, Murali Emani¹, Hyenseung Yoo¹, Junqi Yin⁸, Aristeidis Tsaris⁸, Vishal Subbiah⁹, Tanveer Raza⁹, Jessica Liu⁹, Noah Trebesch², Geoffrey Wells¹⁰, Venkatesh Mysore⁵, Thomas Gibbs⁵, James Phillips¹, S. Chakra Chennubhotla¹¹, Ian Foster¹,⁴, Rick Stevens¹,⁴, Anima Anandkumar³,⁵†, Venkatram Vishwanath¹, John E. Stone²‡, Emad Tajkhorshid²‡, Sarah A. Harris¹²‡, Arvind Ramanathan¹⁴

¹Argonne National Laboratory, ²University of Illinois Urbana-Champaign, ³California Institute of Technology, ⁴University of Chicago, ⁵NVIDIA, ⁶Science and Technology Facilities Council, ⁷Texas Advanced Computing Center, ⁸Oak Ridge National Laboratory, ⁹Cerebras Inc., ¹⁰University College of London, ¹¹University of Pittsburgh, ¹²University of Leeds
†Joint first authors, ‡Contact authors: s.a.harris@leeds.ac.uk, anima@caltech.edu, venkat@anl.gov, johns@ks.uiuc.edu, emad@illinois.edu, ramanathana@anl.gov

ABSTRACT

The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) replication transcription complex (RTC) is a multi-domain protein responsible for replicating and transcribing the viral mRNA inside a human cell. Attacking RTC function with pharmaceutical compounds is a pathway to treating COVID-19. Conventional tools, e.g., cryo-electron microscopy and all-atom molecular dynamics (AAMD), do not provide sufficiently high resolution or timescale to capture important dynamics of this molecular machine. Consequently, we develop an innovative workflow that bridges the gap between these resolutions, using mesoscale fluctuating finite element analysis (FFEA) continuum simulations and a hierarchy of AI-methods that continually learn and infer features for maintaining consistency between AAMD and FFEA simulations. We leverage a multi-site distributed workflow manager to orchestrate AI, FFEA, and AAMD jobs, providing optimal resource utilization across HPC centers. Our study provides unprecedented access to study the SARS-CoV-2 RTC machinery, while providing general capability for AI-enabled multi-resolution simulations at scale.

KEYWORDS

multi-resolution simulations, SARS-CoV-2, COVID19, HPC, AI

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

ACM Reference Format:

Anda Trifan¹,²†, Defne Gorgun¹,²†, Zongyi Li³†, Alexander Brace¹,⁴†, Maxim Zvyagin¹,⁴†, Heng Ma¹, Austin Clyde¹,⁴, David Clark⁵, Michael Salim¹, David J. Hardy², Tom Burnley⁶, Lei Huang⁷, John McCalpin⁷, Murali Emani¹, Hyenseung Yoo¹, Junqi Yin⁸, Aristeidis Tsaris⁸, Vishal Subbiah⁹, Tanveer Raza⁹, Jessica Liu⁹, Noah Trebesch², Geoffrey Wells¹⁰, Venkatesh Mysore⁵, Thomas Gibbs⁵, James Phillips¹, S. Chakra Chennubhotla¹¹, Ian Foster¹,⁴, Rick Stevens¹,⁴, Anima Anandkumar³,⁵†, Venkatram Vishwanath¹, John E. Stone²‡, Emad Tajkhorshid²‡, Sarah A. Harris¹²‡, Arvind Ramanathan¹⁴. 2020. Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action. In Supercomputing ’21: International Conference for High Performance Computing, Networking, Storage, and Analysis. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3441828.3448605

1 JUSTIFICATION

We developed an AI-enabled multi-resolution simulation framework for studying complex biomolecular machines by directly integrating experimental data. Our framework sets high-water marks for AI-driven multi-resolution simulations and achieving high utilization of resources across diverse supercomputing platforms at multiple sites.

2 PERFORMANCE ATTRIBUTES

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

 Chattanooga, Tennessee, USA, 14 November 14–19, 2021.
 ACM ISBN ISBN. . . $15.00
 https://doi.org/10.1145/3441828.3448605

Supercomputing ’21, November 14–19, 2021, St. Louis, MO
© 2020 Association for Computing Machinery.
3 OVERVIEW OF THE PROBLEM

The novel coronavirus 2019 (COVID-19) pandemic has led to a massive acceleration in the pace of development in experimental structural biology. Multiple research teams have invested their best resources towards the common goal of characterizing viral structural biology. Multiple research teams have invested their best resources towards the common goal of characterizing viral structural biology. The details of motions within flexible domains can be enriched using complimentary tools such as molecular dynamics (MD) simulations and Bayesian inference techniques (Bowerman et al., 2017, Bratholm et al., 2015, Cavalii et al., 2007, Grishaev and Linéas, 2005, Scheres, 2012); however, the timescales accessible to these atomistic simulations can be a limiting factor. In addition, advances in 4D imaging modalities (Earnest et al., 2017, Engel et al., 2015, Mahamid et al., 2016, Villa and Lasker, 2014) and the volume of data generated from such experimental datasets can be overwhelming.

Therefore, in this paper we address the urgent, yet unmet need to develop scalable computational tools that can aid the improvement of resolution within cryo-EM datasets through multi-resolution simulations. In an effort to bridge the gap between experimental and purely all-atom molecular dynamics (AAMD), we leverage a complementary mesoscale method of representing biophysical systems, treating biomolecules as visco-elastic continuum solids using fluctuating finite element analysis (FFEA) (Oliver et al., 2013) technique. These continuum-scale lower resolution FFEA simulations provide a generative model for the cryo-EM data. However, implementing an approach that directly models electron density information from cryo-EM data requires a radically different way to model conformational ensembles, one that moves away from atomistic-resolution towards a continuum-representation, where by the intrinsic resolution of the data can be captured with nodes
Intelligent Resolution: Integrating Cryo-EM with AI-driven Multi-resolution Simulations to Observe the SARS-CoV-2 Replication-Transcription Machinery in Action

Figure 2: Hybrid structure of the FFEA mesh superimposed with the all-atom representation. The all-atom structure of the RTC dimer is shown as a cartoon (blue) and the FFEA tetrahedral mesh structure determined from the experimental cryo-EM map is shown as a wireframe. The top inset represents a 90° rotation of the RTC dimer capturing the range of protein-protein interfaces in the machine. A close-up view of the mesh at the interface between the RdRp (nsp12) and nsp13 reveals the surface interface potentials inferred from AAMD simulations. Each mesh point is painted with the surface interface potentials, with darker shades of red indicating higher and lighter shades indicating lower interaction energy in the protein-protein interface.

and meshes, allowing experimentally obtained cryo-EM data to be used directly in simulations, at time and length-scale orders of magnitude increased over traditional all-atom methodologies.

We present a radically innovative workflow, providing the opportunity to accelerate conformational sampling (both AAMD and FFEA) with AI, thus holding much promise in providing mechanistic insights into the dynamics of large biomolecular complexes (Fig. 1). We leverage FFEA and AAMD simulations to iteratively fill in the gaps of cryo-EM data by refining the pairwise atomic/molecular interactions while being automatically constrained by the degrees of freedom embodied in the cryo-EM data. This iterative coarse-graining approach maintains the fidelity between experimentally observed densities with the detailed potentials from rigorous physics-based AAMD simulations to learn informative priors and effectively guide the conformational search and can result in better fits to the experimental observations. We demonstrate our iterative approach in modeling the complex conformational transitions within the SARS-CoV-2 RTC to observe how the various nsps coordinate the proofreading process of the viral-RNA. These conformational changes can aid in understanding and guiding the design of novel therapeutics such as molnupiravir, which aim to target specific nsps. (Agostini et al., 2019, Sheahan et al., 2020)

4 CURRENT STATE OF THE ART

4.1 Parallel molecular dynamics

NAMD (Phillips et al., 2005, 2020) has been one of the most utilized parallel molecular dynamics engines for over two decades, being cited once every ~70 minutes 1. NAMD uses adaptive, asynchronous, message-driven execution based on Charm++ (Kalé et al., 2019, Kalé and Zheng, 2013), while efficiently utilizing GPUs (Phillips et al., 2008). It contains advanced features giving scientists an extensive set of tools to observe and/or bias their systems using state-of-the-art simulation methodologies such as collective variables (Fiorin et al., 2013) and molecular dynamics flexible fitting (MDFF) (Trabuco et al., 2009, Vant et al., 2020) modules.

VMD and its psfgen plugin were used to build and analyze the molecular complexes studied herein. Simulation preparation, visualization, and conventional analysis approaches were performed using VMD, with extensive use of GPU-accelerated remote visualization resources (Humphrey et al., 1996a, Stone et al., 2013a,b, 2016). VMD incorporates features for processing cryo-EM density maps to prepare and analyze mdf hybrid fitting simulations as required in this project (Stone et al., 2014). VMD incorporates a custom GPU-accelerated ray tracing engine that exploits hardware accelerated construction of ray tracing BVH acceleration structures, BVH traversal, and ray-triangle intersections (Sener et al., 2021).

4.2 FFEA: continuum molecular simulations

Recent advances in cryo-EM, and -ET aided scientists to generate structural and dynamic information of large biomolecules in 3D volumetric data (Kühlbrandt, 2014). Expanding on this volumetric shape, FFEA uses a 3D tetrahedral finite element mesh providing dynamic insight of the structure alone and in response to interactions with other molecules (Solernou et al., 2018). FFEA is a new

1https://www.ks.uiuc.edu/Highlights/?section=2021&highlight=2021-09
Figure 3: The atomistic representation of a functional RTC monomer (A) with the RNA (red), extended to the active site of the other nsp14-10 unit of the opposite monomer (nsp14B - orange), and the exit site on the helicase of the same monomer (nsp13-2A). The mechanism of RNA unwinding using non-equilibrium molecular dynamics simulations starting from the initial structure as captured in 7EGQ (B-left), to the final structure (B-right). The RNA backbone is restrained with an mdff grid, represented in a grey mesh, while using colvars to pull the strands to the active sites on the protein. This is an unprecedented simulation of the mechanism of RNA unwinding which can give significant insight into the RTC function.

physics algorithm for simulation of mesoscale biological structures obtained from cryo-EM and cryo-ET (Solernou et al., 2018). FFEA treats biomolecules as continuum visco-elastic solids subject to thermal noise, chosen so as to satisfy the fluctuation-dissipation theorem (Oliver et al., 2013). To describe interacting proteins, FFEA includes short-range van der Waals attractive forces and steric repulsion. Therefore, FFEA uses a unique "top-down" rather than the more conventional "bottom-up" coarse-graining strategy, by introducing nanoscale thermal fluctuations into macroscopic continuum mechanics equations.

FFEPA also includes functionality to exert external forces on proteins, to connect proteins together with harmonic springs and to represent conformational changes between distinct protein conformational states, for example between the pre- and post-powerstroke states of molecular motors (Richardson et al., 2020). FFEA has been used to successfully model diverse biological systems, including the rotary ATPase motor (Richardson et al., 2014), axonemal (Richardson et al., 2020) and cytoplasmic (Hanson et al., 2021) dynein motors, and protein antibodies subjected to external forces (van der Heijden et al., 2020).

4.3 Multiscale biological simulations

Integrating data across diverse spatial, temporal, and functional scales has played an important role in understanding the role of molecular interactions in various diseases (Pak and Voth, 2018, Tozzini, 2010, Walpole et al., 2013, Zhou, 2014). In order to bridge information across multiple scales, a number of coarse-graining approaches have been widely employed. These methods have been used to study individual proteins from the SARS-CoV-2 proteome (De Sancho et al., 2020, Garay et al., 2021), as well as more complex systems such as the entire SARS-CoV-2 virion (Yu et al., 2021). We note that compared to the state-of-the-art approaches, we choose a continuum representation of molecules, one that is more conducive for accessing much larger length- and time-scales using FFEA. While finite element methods are used to model biological systems (e.g., growth models, bone growth, etc.) (Dong and Skalak, 1992, Kennaway and Coen, 2019, Panagiotopoulou et al., 2017), we note that our collective approach enables tight integration between all-atom and continuum representations for large biomolecular complexes such as the SARS-CoV-2 RTC.

While physics-based approaches for coarse-graining are usually employed to overcome limitations of the space- and time-scales accessed by traditional molecular simulations, machine learning techniques are now being regularly employed to adaptive coarse-grained simulations from all-atom simulations (Durumeric and Voth, 2019, Husic et al., 2020, Noé, 2020).

4.4 AI enabled adaptive MD simulations

MD simulation provides atomistic details of molecular interactions and ensued dynamics. It’s however computationally expensive, and the sampling efficiency can be severely hindered, when trapped by local minima in the free energy landscape. Our group has developed AI-enabled adaptive MD simulations, namely DeepDriveMD (Lee et al., 2019), which records MD propagation in latent space via convolutional variational autoencoder (CVAE) (Bhowmik et al., 2018), and drives adaptive sampling in the high dimensional conformational landscape. The AI inference constantly observes the simulation runs, prunes stagnant ones that are trapped in local energy minima, and spawns new ones from less sampled conformations. We have shown that this approach can accelerate sampling of rare events (for example, in the SARS-CoV-2 Spike opening simulations (Casalino et al., 2021), protein folding (Lee et al., 2019))
by at least an order of magnitude. When integrated with specialized AI-hardware to accelerate the learning, it can provide nearly 4 orders of magnitude speedup (Brace et al., 2021).

### 4.5 Workflow infrastructure

Distributed science workflows must invariably handle data and control flow across networks spanning administrative domains. Challenging barriers to workflow management frameworks’ adoption in multi-user HPC environments lie in deploying distributed client/server infrastructures and establishing connectivity among remote systems. For instance, Fireworks (Jain et al., 2015) and the RADICAL-Pilot / Ensemble Toolkit (EnTK) (Balasubramanian et al., 2016, Merzky et al., 2018) are three widely-used WMFs at supercomputing facilities that expose Python APIs to define and submit directed acyclic graphs (DAGs) of stateful tasks to a database. These WMFs possess various implementations of a common *pilot job* design, whereby tasks are efficiently executed on HPC resources by a process that synchronizes state with the workflow database. Because the database is directly written by user or pilot job clients, users of these WMF typically deploy and manage their own database servers such as MongoDB servers. These methods are non-portable and depend on factors such as whether the compute facility mandates multi-factor authentication (MFA).

In the context of REST interfaces to HPC resources, efforts such as the *Superfacility* concept (Enders et al., 2020) envisions a future of automated experimental and observational science workflows linked to national computing resources through Web APIs. Implementations such as the NERSC Superfacility API and the Swiss National Supercomputing Centre’s FirecREST API (Cruz and Martinasso, 2019) expose methods to submit and monitor batch jobs, move data between remote filesystems, and check system availability. However, these facility services alone do not address workflow management or high-throughput execution; instead, they provide web-interfaced abstractions of the underlying facility, analogous to modern cloud storage and infrastructure services.

funcX (Chard et al., 2020) is an HPC-oriented instantiation of the *function-as-a-service* (FaaS) paradigm, where users invoke Python functions in remote containers via an API web service. funcX endpoints run on the login nodes of target HPC resources where Globus Auth is used for authentication and user-endpoint association. The funcX model is tightly focused on Python functions and therefore advocates decomposing workflows into functional building blocks. Each function executes in a containerized worker process running on one of the HPC compute nodes. This fails to support a generalized model of applications such as simulations with executables that may or may not leverage containers, per-task remote data dependencies, programmable error- and timeout-handling, and flexible per-task resource requirements (e.g. tasks may occupy a single core or specify a multi-node, distributed memory job with some number of GPU accelerators per MPI rank).

Balsam (Salim et al., 2021) provides a unified API for describing distributed workflows, and the user site agents manage the full lifecycle of data and control flow. The Balsam service-oriented architecture shifts administrative burdens away from individual researchers by routing all user, user agent, and pilot job client interactions through a hosted, multi-tenant web service. As Balsam execution sites communicate with the central service only as HTTP clients, deployment involves a simple user-space *pip* package installation on any platform with a modern Python and outbound internet access. Due to the ubiquity of HTTP data transport, Balsam works “out of the box” on supercomputers spanning DOE and NSF facilities against a cloud-hosted Balsam service.

### 5 INNOVATIONS REALIZED

Our innovations can be briefly summarized as follows. (1) We implement an automated workflow to seamlessly transition from AAMD to FFEA via an AI-based approach. By optimizing the performance of AAMD simulations (Sec. 5.1) we achieve optimal performance on modern multi-GPU-based computing systems. (2) By clustering the conformations from AAMD simulations using unsupervised AI methods (Sec. 5.3), we learn the parameters for constraining the FFEA simulations based on the implicit fluctuations between protein-protein interfaces inferred from AAMD simulations. We develop a novel AI-approach, namely Graph neural operators (GNO) (Sec. 5.3.3) to summarize time-dependent conformational changes observed from ensemble AAMD simulations. (3) To overcome the bottleneck of training deep learning models on large volumes of data generated by AAMD simulations, we scaled their performance on emerging AI-hardware such as the Cerebras CS-2 accelerator as well as on the Summit supercomputer. Finally, to our knowledge, this is the first attempt at executing a single coordinated workflow across multiple supercomputing facilities. We achieved this through the Balsam multisite workflow manager (Sec. 5.4). The performance gains enabled by these innovations provide a generalized, multi-scale computational toolkit for exploring dynamic biomolecular machines.

#### 5.1 NAMD all-atom MD simulations

##### 5.1.1 GPU-resident NAMD 3 for GPU-accelerated MD simulations

To achieve previously unrealized performance levels for MD simulations on GPU-accelerated HPC platforms, we have developed a major new “GPU-resident” computing approach, implemented in NAMD 3. The GPU-resident computing mode in NAMD 3 employs GPU acceleration for the most common molecular dynamics force and energy calculations together with time integration and rigid bond constraints, allowing sequences of timesteps to be simulated entirely on the GPU without any significant host–GPU data transfers, thereby exploiting a significantly higher fraction of theoretical peak GPU performance (Phillips et al., 2020). Most recently, we have further extended the GPU-resident capabilities of NAMD 3 to enable strong-scaling of a simulation across multiple GPUs, with particular emphasis on hardware platforms that incorporate NVLink connectivity among peer GPUs. NVLink facilitates direct load/store memory access among peered GPUs with low latency and high bandwidth data exchange performance levels far beyond what is currently possible with a conventional HPC network interconnect, or even host-internal PCIe transfers among GPUs.

The multi-GPU simulation algorithm generally mirrors the techniques used in the conventional distributed-memory builds of NAMD,
except that it is designed to leverage the completely unified virtual memory address space of the GPUs to exchange data directly through fine-grained memory load/store operations. NVLink support for direct fine-grained memory loads/stores in concert with atomic increment and atomic compare-swap operations permit efficient implementation of performance-critical multi-GPU reductions and synchronization of GPUs within algorithm phases. Platforms providing GPUs with a fully connected topology permit strong scaling of molecular systems ~1M atoms on up to 8 GPUs with acceptable efficiency, limited primarily by poor scalability of the PME full electrostatics algorithm.

While running a simulation per GPU avoids these scaling challenges, multi-GPU scaling enables simulation over longer timescales. For systems of ~1M atoms, the timescale is typically a limiting factor in understanding how a complex function. Performance considerations dictate that any advanced sampling method implemented for GPU-resident must avoid frequent data transfers to and from the host. Many of these features are contributed by external methodological contributors and have not yet been implemented to run on GPUs. As such, NAMD 2 and CPU-based resources still play an important role in the workflow.

5.1.2 VMD full-time interactive ray tracing. To facilitate creation of high-fidelity visualizations with high interactivity, VMD has been extended with a new real-time ray tracing (RTRT) rendering mode that uses full-time interactive progressive refinement ray tracing for rendering of the molecular scenes instead of OpenGL rasterization and includes ambient occlusion lighting, shadows, high-quality transparency, and depth of field focal blur. The new RTRT rendering mode provides substantially more visually informative visualizations of complex molecular scenes as compared with conventional OpenGL rasterization, by bringing offline-quality rendering features to the interactive VMD display window for the first time. When the RTRT rendering mode is active, advanced visualization features that were formerly only usable for offline rendering, or for static structures, can now be applied to simulation trajectories while maintaining full interactivity.

5.1.3 RTC System Preparation for all-atom simulations with NAMD. The SARS-CoV-2 RTC is a multi-subunit structure composed of several nsps, including the RNA-dependent RNA polymerase (RdRp, nsp12), Zinc-bound helicase (HEL, nsp13), the RNA-capping enzymes such as the nsp14 (N7-methyltransferase/MTase) and nsp16 (O2-MTase), the proofreading enzyme, namely nsp14, and uridylate-specific endoribonuclease activity (NendoU, nsp15)(Fig. 3A) (Chen et al., 2020, Perry et al., 2021, Yan et al., 2020, 2021a). Together they participate in the process of reproducing the viral genome (Wu et al., 2020). This complex plays an important role in the viral life cycle, and is therefore an important drug target (Gao et al., 2020, Wang et al., 2020). Revealing the atomic details of this complex is essential for understanding how the viral RNA is processed.

We prepared the structure of the Cryo-EM dimeric form of the complex 7EGQ.pdb (Yan et al., 2021b). Cations have functional significance for the RTC, therefore the linking HIS and CYS residues were mutated to HSE/HSD and CYM respectively, to have the correct ionization state to coordinate the present 3 Mg$^{2+}$ ions and 26 Zn$^{2+}$ ions. A complementary CYT base was added to the end of two
incomplete RNA strands. Missing loop residues ARG ARG ARG corresponding to residue numbers 314-316 in the nsp13 subunits were modeled using pdbfixer tool in OpenMM (Eastman et al., 2017). The protein subunits were capped with an NTER and CTER patches. The PFGSFIN package from VMD (Visual Molecular Dynamics) (Humphrey et al., 1996b) was used to assemble the system, which was then solvated using the SOLVATE plugin and neutralized with K⁺ and Cl⁻ ions for a concentration of 150 mM KCl using AUTOIONIZE. The resulting systems consist of ~1.1 M atoms and box size of ~168 x 220 x 312 Å³. The systems were energy minimized using steepest descent method for 500 steps, followed by a 3 ns equilibration protocol restraining all but the water molecules using a constant k = 3 kcal mol⁻¹ Å⁻². The system was further equilibrated restraining only the protein backbone alpha carbon atoms and nucleic backbone phosphate atoms with a constant of k = 0.05 kcal mol⁻¹ Å⁻² with NAMD2.14 (Phillips et al., 2005) NVT ensemble using CHARMM36 forcefields for proteins and nucleic acids (Vannommeslaeghe et al., 2010).

5.1.4 Grid-steered Molecular Dynamics Simulation of RNA Unwinding. One of the crucial functions of the RTC complex is to proofread the RNA, which is a process that requires a major conformational change of the protein complex and unzipping of the RNA. To mimic this operation, we performed extensive preliminary studies using an isolated RNA double-helical structure from the aforementioned prepared structure in order to optimize the combination of grid-steered molecular dynamics (GMD) (Wells et al., 2007) and collective variables (colvars) (Fiorin et al., 2013) and their respective optimal force constants to mimic the corkscrew motion of the RNA unwinding. To speed up these empirical studies, instead of the all-atom representation, the protein was represented by a molecular dynamics flexible fitting mdf grid map exerting a repulsive gridforce (Trabuco et al., 2009). A variety of colvars were employed to drive the extension of the RNA structure, including distanceZ to pull on the strands to the nsp14-10 active site and the exit point on the nsp13 helicase, as proposed in previous studies (Yan et al., 2021b). We applied a rotational force by using the orientationAngle colvar, while also pulling on the center of mass using distance colvar. The final conformation used for the simulations are a hybrid of the initial double-helical structure of the RNA strands, with the extended single strands which have reached their target points within the complex (Fig. 3B). This structure was obtained by building an attractive 3 dimensional electron density map manually constructed from the backbone of the initial RNA position with the final extended strands obtained from the non-equilibrium simulations.

The all-atom simulations to mimic this process were started from the equilibrated structure described in section 5.1.3. We performed an initial system minimization for 2000 steps and then simulated the system for 0.05 ns to allow the RNA to settle into the grid. After this initial equilibration, the process of unwinding was simulated by pulling on the strands through this grid, which acts as a tunnel, using the distance colvars, with a harmonic force. A soft restraining force was applied on the backbone of the nucleic acids. To avoid distortion of the double-helical shape of the RNA during pulling, the hydrogen bonds formed by the bases were enforced by virtual springs through the NAMD extraBonds feature, with a force of 20 kcal/mol/Å².

5.2 FFEA simulations

5.2.1 Meshing and construction of the RTC dimer for FFEA. Two representative meshes were constructed which correspond to the experimental data available; the cryo-EM density map from EMD-31138, a dimer form of the biologically active RTC, and the atomistic structure 7EGQ.pdb, which was used in the above NAMD atomistic molecular dynamics simulations. In the latter, the monomer RTC complex was formed from two separate meshes, one from the cryo-EM density map EMD-23008, corresponding to 7KRO.pdb, and the other from 7EGQ.pdb for the nsp14-10 complex atomistic information. These structures were used to generate 4 four interacting meshes; two in each monomer (Fig. 2). Meshes were constructed by first removing the dust and smoothing the mask using ChimeraX (Pettersen et al., 2021). The non-manifold elements within the surface were corrected with "MeshLab" (Cignoni et al., 2008), and coarse-grained with "MeshMixer" (Sommer et al., 2017). The 3D meshes were then generated by "Gimsh" (Geuzaine and Remacle, 2009) and made compatible for FFEA simulations with NetGen (Schöberl, 1997).

Following dust-removal and smoothing in ChimeraX, density for the single-stranded RNA was clearly visible in EMD-23008. However, this extremely thin structure is likely to cause instabilities in FFEA due to the small volumetric elements required to represent it. Therefore, the surface elements corresponding to RNA density were deleted, and the resulting surface repaired by filling with additional triangular elements in MeshMixer. The physical presence of the single-stranded RNA was then represented in the FFEA simulations using a harmonic spring.

5.2.2 Visualizing FFEA trajectories using VMD. In order to create informative visualizations that contain multiple simulation modalities VMD was extended with new plugins to enable reading key FFEA simulation data. The new plugins enable VMD to read FFEA simulation and modeling data including tetrahedral mesh nodes, edges, faces, and topology information, as well as information about surface and interior faces and their properties, springs, and time-varying mesh node coordinates. The plugin extensions permit VMD to create visualizations showing the all-atom and FFEA simulations aligned and superimposed, with physical properties associated with the atoms, and FFEA nodes and faces available for use in assigning surface colors for graphical representations of the simulations.

5.3 AI-enabled multi-resolution simulations

We developed a hierarchy of AI methods that would facilitate embedding the high-dimensional (FFEA and AAMD) simulations into a low-dimensional manifold capturing both local (e.g., protein chain/subunit level) and global (e.g., mRTC monomer or dimer level) fluctuations. This required us to carefully consider the balance of simulation workloads and AI tools. We discuss the AI methods first, and then address the workload balance aspects in Sec. 5.4.

5.3.1 Capturing global conformational transitions with anharmonic conformational analysis enabled autoencoders (ANCA-AE). We have previously developed a hybrid machine learning approach that combines linear dimensionality reduction methods (Ramanathan et al., 2011) with a non-linear autoencoder. By minimizing the higher order (linear) correlations in the atomistic fluctuations and then
examining non-linear correlations, we can obtain a succinct description of how global conformational changes are embodied in a simulation. This method, titled ANCA-AE (Clyde et al., 2021), can efficiently handle large dimensions (e.g., in the case of the mRTC, the 6,650 Cα-atoms lead to approximately 20,000 dimensions) and can efficiently run on CPUs. We have shown that ANCA-AE can identify conformational states that share structural and energetic similarities, while characterizing transitional points in the high dimensional landscape. As shown in Fig. 5, we observed (for both FFEA and AAMD, although we only show results from FFEA simulations) that ANCA-AE extracts biophysically meaningful latent coordinates. The ANCA-AE method reveals conformational transitions from the FFEA trajectories, capturing the fluctuations in the interfaces between the nsp10-nsp14 complex and the rest of the RTC subcomponents.

5.3.2 Describing localized conformational transitions with convolutional variational autoencoders (CVAEs). We also leveraged a deep learning algorithm, namely CVAE (Bhowmik et al., 2018) to describe localized conformational changes. The CVAE provides a compact representation of the high dimensional conformational landscape and can also capture intermediate/metastable states from long timescale MD simulations (Bhowmik et al., 2018). It can also be combined with outlier detection algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al., 1996) and the Local outlier factor (LOF) (Breunig et al., 2000) method to identify time points corresponding to rare transition events in the simulation trajectories. We note that both ANCA-AE and CVAE can utilize outlier detection algorithms for this purpose. Since we have previously discussed the CVAE model, we only present scaling results for this model.

However, CVAE is quadratic in time and space complexity (Casalino et al., 2021) and can be prohibitive to train for the entire mRTC. This makes the CVAE ideal to examine protein sub-unit fluctuations, which can be significantly smaller and more tractable for training. To optimize its performance, we implemented a distributed data parallel version of CVAE on the Summit supercomputer as well as a version within a single Cerebras CS-2 deep learning accelerator. By keeping all compute and memory resources on-silicon, the CS-2 provides, in a single system, orders of magnitude more compute power, on-chip memory, memory bandwidth, and communication bandwidth than traditional small-scale processors. This makes it particularly well-suited for accelerating ML models like the CVAE.

5.3.3 Graph neural operator (GNO) networks for characterizing time-dependent conformational changes from AAMD simulations. Additionally, we also implemented the graph neural operator (GNO) as a fast surrogate model to simulate and predict the molecule movements from the AAMD simulation. The applications of deep learning for MD trajectory analysis, prediction and generation have recently exploded (Hoseini et al., 2021). The GNO model (Li et al., 2020) can solve partial differential equations (PDEs) by generalizing finite-dimensional neural networks to the infinite-dimensional operator-learning setting. In this work, we apply the GNO for the first time to capture the time dependent conformational changes governed by the system of equations in a traditional molecular dynamics simulation. As a first step, we show that we can reasonably predict protein backbone conformations up to 5 ps (250 MD simulation timesteps of 2 fs) into the future based on a 50 ps trajectory. This allows us to extract the time-dependent conformational changes of the molecule and analyze the connection between AAMD and FFEA simulation. This approach is capable of achieving an even longer prediction horizon of 500 ps on smaller test systems, and will be described in a separate manuscript. As shown in Fig. 5B, the ability to predict the time-dependent conformational changes in the entire RTC at short time-scales provides a means to accelerate sampling. The structure of the latent space following the RMSD transitions, provides insights into the large conformational changes captured in 5C.
Together, the uniform Balsam API and portable Balsam Sites form a service that manages job distribution. Clients of this service use an API (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5.4 Balsam workflow

Balsam (Salim et al., 2021) (Figure 6) provides a centralized service allowing users to register execution sites from any laptop, cluster, or supercomputer on which they wish to invoke computation. The sites run user agents that orchestrate workflows locally and periodically synchronize state with the central service. Users trigger remote computations by submitting jobs to the REST API specifying the Balsam execution site and location(s) of data to be ingested.

The ability to develop a unified workflow infrastructure that can span multiple computing sites by coupling traditional HPC simulation tasks with AI tasks is a key contribution of this paper. Further, the ability to steer a set of computational simulations based on AI and automatically switch simulation resolution (between all-atom and continuum) is a further jump in capabilities.

A key development is the introduction of a central Balsam service that manages job distribution. Clients of this service use an API to inject and monitor jobs from anywhere. A second key development is the Balsam site, which consumes jobs from the Balsam service and interacts directly with the scheduling infrastructure of a cluster. By relying on a well-defined set of platform interfaces that adapt to machine-specific schedulers, node configurations, and pilot job launch paradigms, the Sites are easily portable across new and heterogeneous systems with minimal implementation effort. Together, the uniform Balsam API and portable Balsam Sites form a distributed, multi-resource execution landscape. The power of this approach was demonstrated using the Balsam API to inject simulation and AI jobs to the Balsam service, targeting simultaneous execution on multiple supercomputers. By strategically combining data transfer bundling and controlled job execution, Balsam achieves higher throughput than direct-scheduled local jobs and effectively balances job distribution across sites to simultaneously utilize multiple remote supercomputers.

The workflow driver agent comprised a lightweight Python process that used the Balsam SDK to track and dispatch NAMD 3 simulations on ThetaGPU and Perlmutter. A Python wrapper of a NAMD 3 replica’s execution enables the agent to seamlessly configure simulation parameters, dispatch runs, monitor data transfers, and send shutdown signals to simulations through a Python API. As simulation conformations arrive at the AI Site, the agent triggers model training and inference runs, whereby the outlier conformers are determined and used to initialize new simulations. In this scheme, a single Python agent exercises complete control over the geographically distributed simulation campaign. The workflow is robust to individual task- and batch job-level faults: a user monitoring the workflow with Balsam can reset failed jobs and provision additional resources by triggering new pilot jobs in the middle of a single agent’s running experiment. Experiments are launched with a simple YAML configuration format that flexibly supports arbitrarily many “simulation pools” on remote HPC systems: these pools comprise a particular simulation ensemble with data transfer settings (e.g. Globus transfer destination) which are passed through the Balsam Site’s Transfer module.

6 HOW PERFORMANCE WAS MEASURED

6.1 NAMD all-atom MD simulations

Production NAMD simulation performance is reported at runtime, directly yielding an achieved ns/day simulation rate. In order to report performance in terms of underlying machine FLOP rates, the marginal FLOPs per step are computed using hardware performance counters. Per-step FLOP performance metrics for NAMD 2.14 were collected on TACC Frontera, using the Intel msr-tools utilities, and the “TACC stats” system. FLOP counts were measured for each NAMD simulation with runs of two different step counts. The results of the two different simulation lengths were subtracted to eliminate counting of startup-associated operations, to provide an accurate estimate of the FLOPs per step for a continuing simulation (Phillips et al., 2002).

6.2 AI methods

As outlined in Sec. 5.3, the ANCA-AE method is quite efficient and for the purposes of this paper, we did not need to optimize its performance. Further, we also note that the hybrid nature of combining linear- and non-linear dimensionality reduction opens up new avenues to assess scaling requirements from large datasets (which is beyond the scope of this paper). Similarly, the newly developed model, GNO was also not optimized for performance. We compared the performance of the GNO model with other baseline

---

Figure 6: Cross-site Balsam Workflow architecture. Users execute scientific workloads by first creating Job resources at a specific execution site through the Balsam REST API. Pilot jobs at Balsam sites on the ALCF-ThetaGPU, and NERSC-Perlmutter supercomputers fetch the appropriate Jobs for execution. The MD Simulation and AI applications can span one or more sites based on the resource needs.

---

https://www.alcf.anl.gov/support-center/theta/theta-thetagpu-overview
https://www.nersc.gov/systems/perlmutter/
https://www.tacc.utexas.edu/systems/frontera
https://github.com/intel/msr-tools
https://github.com/TACC/tacc_stats
neural networks on a 28-node graph with 10,000 time frames. We report the total L2 error on a single GPU trained with 10 epochs.

We measure the throughput performance (samples/s) of the distributed training of CVAE model on the Summit supercomputer. The code is implemented in TensorFlow and parallelized with Horovod library in a data-parallel scheme. We then report the aggregated training throughput from 1 node up to 256 nodes (1536 GPUs). The scaling efficiency is determined by the ratio of ideal throughput (linear scaling) over the measured throughput. For results on the CS-2, we measured the throughput performance (samples/s) of training the CVAE model on a single Cerebras CS-2 system.

### 6.3 Balsam workflow

Balsam records each job state transition (e.g. RUNNING → RUN_DONE) with timestamps and metadata that can be queried to produce workflow-level reports of task throughput and utilization (nodes occupied by running Balsam tasks, as a fraction of the currently provisioned resources). We then quantify the scientific throughput by measuring the cumulative MD simulation time, aggregated over all ensemble pools within an experiment). The Globus data transfer time is captured and recorded as part of the telemetry.

### 7 PERFORMANCE RESULTS

#### 7.1 NAMD all-atom MD simulations

The performance of the NAMD production runs for the “RNA post-unwinding” are shown in Table 2, reporting CPU-based Frontera running NAMD 2 and two different DGX A100 320GB systems running GPU-resident NAMD 3. The first reported result was run on ThetaGPU, configured with one replica per GPU. The second reported result was run on NVIDIA Base Command,10 configured with one replica per node to simulate longer timescales. The Frontera and ThetaGPU measurements are averages. The median was used for NVIDIA Base Command jobs due to outliers; 3 of the 15 nodes achieved around 64ns/day, likely due to system I/O traffic.

NAMD 3 was benchmarked on the “RNA post-unwinding” system from Table 1 to characterize multi-GPU scaling performance. The equilibrated structure was run on a DGX A100 640GB node for 80000 timesteps using 8 CPU cores per GPU, and these simulations included the same parameters (e.g., 8Å cutoff) and file I/O as performed by the production runs, with trajectory output every 5000 steps. The results with estimated TFLOPS performance is reported in Table 3.

The short-range non-bonded interaction calculations, which dominate the computational work per step, have long been optimized in NAMD by interpolating from a lookup table, effectively trading L1 cache reads for FLOPs. Doing this avoids expensive calculation of the error function erf() within the inner loop, as required by the particle-mesh Ewald (PME) method for calculating long-range electrostatics (Essmann et al., 1995). Use of a lookup table together with the CPU-based count of FLOPs per step leads to a very conservative estimate of TFLOPS for GPU-resident NAMD. GPU-based kernels actually produce a greater number of FLOPs per step than CPU-based kernels due to the need to accommodate a wider vector width. In NAMD 3, the non-bonded kernel computes interactions between tiles of 32 atoms, which allows for a high level of data reuse from on-chip memory while avoiding collisions and atomic operations. However, within a tile–tile calculation, a higher number of atom pairs will be outside of the cutoff radius, producing more GPU FLOPs than predicted by the CPU-based FLOP count.

As mentioned earlier, GPU scaling is primarily limited by the PME calculation. PME summation in reciprocal space requires computing 3D FFTs which cannot be decomposed efficiently across multiple GPUs, thus, the calculation is performed on a single GPU. Profiling the 8-GPU case shows that the GPU doing PME is overloaded at 80% utilization, while the rest of the GPUs are at 50% utilization. Strategies to mitigate this load imbalance include reducing the workload of the current PME implementation by distributing those parts of PME that are, in fact, scalable (in particular, the charge-spreading and force-interpolation procedures) and then employing task-based parallelism to reduce the amount of additional work assigned to the GPU doing the truly non-scalable PME work.

#### 7.2 AI methods

As shown in Fig. 7, we achieved about 87% scaling efficiency in the distributed training of CVAE model with input size (512, 512) and (1024, 1024) on the Summit supercomputer. Equivalently, for an input size of 512 × 512, the CS-2 delivers out-of-the-box performance of 24,000 samples/s, or about the equivalent of 110-120 GPUs. For a larger input size of 1024 × 1024, CS-2 delivers 4,700 samples/s, also the equivalent of over 100 GPUs.

For these experiments, we measured performance as-is: without taking time to optimize the model or input functions for the Cerebras architecture. With additional time, we believe the performance from CS-2 would be much higher. Unlike with a typical distributed GPU setup, where hyperparameters need tuning as one scales from a single device to a large cluster of GPUs, the CS-2 is a single, powerful node. This means that we were able to run the CVAE model with the same experimental configuration as for a single GPU, without needing any hyperparameter changes.

Because a single CS-2 here delivers the performance of over 100 GPUs, it is a practical alternative for organizations interested in this workflow who do not have extremely large GPU clusters. By providing focused AI acceleration for the neural network portion of

---

**Table 2: NAMD "RNA post-unwinding" performance results.**

<table>
<thead>
<tr>
<th>System</th>
<th># of nodes per replica</th>
<th>ns/day</th>
</tr>
</thead>
<tbody>
<tr>
<td>TACC Frontera</td>
<td>128</td>
<td>46.37</td>
</tr>
<tr>
<td>NVIDIA DGX A100 320GB</td>
<td>1/8</td>
<td>20.33</td>
</tr>
<tr>
<td>NVIDIA DGX A100 320GB</td>
<td>1</td>
<td>75.37</td>
</tr>
</tbody>
</table>

**Table 3: NAMD 3 GPU-resident scaling on DGX A100 640GB.**

<table>
<thead>
<tr>
<th># of GPUs</th>
<th>ns/day</th>
<th>Scaling efficiency</th>
<th>Estimated TFLOPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.7767</td>
<td>100%</td>
<td>1.3274</td>
</tr>
<tr>
<td>2</td>
<td>36.1360</td>
<td>87%</td>
<td>2.3087</td>
</tr>
<tr>
<td>4</td>
<td>56.3996</td>
<td>68%</td>
<td>3.6033</td>
</tr>
<tr>
<td>8</td>
<td>75.2359</td>
<td>45%</td>
<td>4.8067</td>
</tr>
</tbody>
</table>

We evaluate Balsam’s efficacy utilizing multiple HPC supercomputers run on both the Perlmutter and ThetaGPU systems. The AI training and inference models, in this case the ANCA-AE, were run on a dedicated ThetaGPU node. The trajectory information from the simulations is transferred via Globus from Perlmutter to ThetaGPU. This data together with the data from the local simulations at ThetaGPU are used to both re-train as well as to run inference. Based on the inference, promising simulation configurations are queued up with the Balsam service for launch and shutdown. The agent signals simulations to shutdown and restart from outlier conformations (that keep the protein-protein interfaces intact within the RTC simulations) reveal the dynamic nature of how such concerted motions are mediated across the entire RTC via protein-protein interfaces.

The implicit coupling between AAMD and FFEA simulations for integrative biology will be invaluable for making use of large experimental data that are difficult to transfer.

We demonstrated the work flow across heterogeneous computational architectures, diverse HPC sites – two supercomputers, namely ThetaGPU (ALCF) and Perlmutter (NERSC). With the balkanization of heterogeneous architecture and software platforms, the workflow represents a necessary capability that can be leveraged across domains where heterogeneous tasks and hardware platforms are the norm (for e.g., climate modeling, materials simulation, etc.). Further, our resource utilization is near 100% and optimally distributed for AI vs. HPC (simulation) jobs.

The capability to perform multi-site data analysis and simulations for integrative biology will be invaluable for making use of large experimental data that are difficult to transfer.

We have shown that balsam is able to sustain close to 100% utilization on Perlmutter. Quick training and inference jobs, shown as stars on the timeline, are launched asynchronously on ThetaGPU upon receiving preprocessed training data from Perlmutter. The agent signals simulations to shutdown and restart from outlier conformations which leads to brief periods of inactivity on Perlmutter.

The ability to perform multi-site data analysis and simulations for integrative biology will be invaluable for making use of large experimental data that are difficult to transfer.

We demonstrated the workflow across heterogeneous computational architectures, diverse HPC sites – two supercomputers, namely ThetaGPU (ALCF) and Perlmutter (NERSC). With the balkanization of heterogeneous architecture and software platforms, the workflow represents a necessary capability that can be leveraged across domains where heterogeneous tasks and hardware platforms are the norm (for e.g., climate modeling, materials simulation, etc.). Further, our resource utilization is near 100% and optimally distributed for AI vs. HPC (simulation) jobs.

The capability to perform multi-site data analysis and simulations for integrative biology will be invaluable for making use of large experimental data that are difficult to transfer.
ACKNOWLEDGMENTS

We thank the Argonne Leadership Computing Facility supported by the DOE under DE-AC02-06CH11357, the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory supported by the DOE under Contract DE-AC05-00OR22725, and the National Energy Research Scientific Computing Center at Lawrence Berkeley National Laboratory supported by the DOE under Contract No. DE-AC02-05CH11231. 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 MCB-20024). NAMD and VMD are funded by NIH P41-GM104661. 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-SC0019232). 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 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 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.

REFERENCES


