AI driven integrative biology for accelerating therapeutic discovery against SARS-CoV-2

**A. Scientific/Technical Goals**

Our proposed work seeks to address both the fundamental biological mechanisms of the virus and the disease, while simultaneously targeting the entire viral proteome to identify potential therapeutics. Our proposed work will develop machine learning (ML), deep learning (DL) and artificial intelligence (AI) techniques to:

(1) Identify, and build accurate three-dimensional structural models of the SARS-CoV-2 proteome by integrating experimental structural and systems biology datasets **[Task 1]**,
(2) Accelerate adaptive conformational sampling of the viral proteins to potentially identify novel binding sites/ pockets that can be targeted by small molecules **[Task 3]**,
(3) Rapidly filter, rank, and search for small molecules across widely available chemical libraries and to integrate virtual screening (computational drug discovery techniques) techniques with experimental high throughput screening **[Task 2]**,
(4) Enable multi-scale, multi-resolution simulations of the SARS-CoV-2 viral envelope, and specific proteins **[Task 4]**, and
(5) Characterize the evolutionary 'traits' of the virus including identification of epitopes and the viral genome that can be targeted for vaccine design **[Task 1]**.

The immediate impact of our current research is to build an ecosystem of open source AI/ML tools and conventional physics based simulations that can accelerate timely response for treating such pandemics. We have made significant progress across the aforementioned goals, including the development of AI/ML tools for rapidly filtering chemical space to identify small molecules that can bind to various viral protein targets, adaptive conformational sampling using molecular dynamics (MD) simulations, and building all-atom models for the entire viral envelope. Further, our approaches leverage open source software and tools that have already been tested on leading high performance computing facilities across the country. In addition, all of the data, computational tools and software will be released into the public domain to enable easy access, sharing and dissemination for further research.

Fig. 1: AI-driven multi-scale modeling of SARS-CoV-2 proteome for small molecule discovery.

As illustrated in Fig. 1, the central aspect of our effort is focused on AI-driven techniques that enable rapid filtering and ranking across compound libraries (task 2), with the potential to fast-track over billion compounds, which uses an active learning strategy to 'seed' good binders to the active sites of various protein targets (identified in task 1) followed by successive 'fast' refinement using AI-driven adaptive sampling strategies (task 3). Simultaneously, we propose to utilize existing structural data to build all-atom (AA), and coarse-grained (CG) models for various viral targets and the entire viral envelope (task 4).

**B. Estimate of Compute, Storage and Other Resources**

The project is organized around five primary scientific tasks, summarized as workflows. Each workflow makes use of distinct computational resources and can be composed in parallel, leading to effective usage of these resources (across distributed facilities). A brief description of each workflow is provided below, including a summary of the open source tools/ frameworks it uses. Computational resources requested are provided in Table 1.

**[Task 1] Workflow 3: SARS-CoV-2 target identification** The primary goal of this task is to obtain genomic, structural, and systems biology data to obtain full three-dimensional (3D) structural models of proteins from both the viral proteome and host proteins that interact with them. It has been quite challenging to obtain information regarding the host proteins (that the viral proteins interact with) and characterize missing residues/loops within these 3D structures.

**[Task 2] Workflows 0 and 1: Rapidly filter, rank, and search for small molecules across widely available chemical libraries for target proteins**Accessing large-scale libraries could potentially give us interesting chemical space. However, current computational tools are limited in how they can scale to use these large-scale libraries. We will filter candidates from the collection using RL-Dock. We have developed RL-Dock, an agent-based docking protocol that uses advancements in reinforcement learning (RL) to drive a ligand onto a protein in a computational environment where the protein may be moving. Our method can interface with other AI tools as well as standard physics-based models (e.g., DOCK) and can effectively evaluate small-molecules that may target specific proteins within the COVID-19 proteome.

**Workflow 1a**: We will evaluate our predictions using emerging uncertainty scoring methods, which we developed while predicting cancer cell line response to drug treatment. Details can be found in [Xia 2018]. **Workflow 1b:** Results from the candidate filtering and uncertainty scoring provide not only a prioritized list compounds for experimental screening, but as well input to active learning methods that aim to understand what area of the chemical space we need to increase our sampling of to improve the performance of our AI models. To ensure we are using highly tuned AI techniques, we will use hyperparameter (HP) tuning approaches as well as network search strategies that have been developed as part of our DOE exascale computing project - CANDLE.

**[Task 3]**: **Workflow 2a: Accelerate adaptive conformational sampling of target proteins**

MD simulations can be easily trapped within local minima and therefore can become less productive in sampling new molecular states. Various adaptive MD simulations have been proposed. Common to each approach is the start of an ensemble of MD runs, which is then managed iteratively based on some statistical criteria. We posited that a variational autoencoder could be used to interface with an ensemble of MD runs and make decisions to continue or terminate a simulation based on whether the run sampled different conformer states, or spawn new simulations from a less sampled configuration with available

hardware (GPUs). Our implementation, called DeepDriveMD enables us to sample the conformational space spanned by the protein more comprehensively than other techniques.

**[Task 4] Workflow 2b: Multi-scale, multi-resolution simulations of the SARS-CoV-2 viral**
A further goal of the proposal is to extend our DeepDriveMD to enable multi-scale modeling of the entire viral envelope. Dr. Amaro's group has built all-atom models of the Spike protein, and have initiated full system-scale runs of the viral envelope with the primary motivation of finding potentially novel binding sites on the virus that can be targeted. We are planning to utilize these two approaches to provide a better understanding how the virus attacks the host system, and elucidate novel binding modes.

### C. Support Needs

This is a rapidly evolving project and the exact distribution of compute resources across the different workflows will be determined by intermittent science results, as well as the effectiveness of the methods. Further, some workflows (e.g., coarse graining) have not been benchmarked yet, but drawing upon extensive experience we provide the following best effort estimates:

| | Target Platform<br>F = Frontera - Longhorn<br>S = Summit; L = Lassen | Configuration<br>(node-count, duration, number of runs) | Estimated Total node-hours |
|---|---|---|---|
| Workflow 0 | Frontera | 128 CPUs, 12, 50 | 75K node hours |
| Workflow 1 | Frontera-Longhorn | 64 CPUs,12,100 + 64 GPUs, 12, 100 | 75K node hours + 75K GPU hours |
| Workflow 2 | F, S, L | 144GPUs,12hrs,100 | 200K GPU hours |
| Workflow 2-a | F, S, L | 256 GPUs,24hrs,10 | 75K GPU hours |
| Workflow 2-b | F, S | 256 GPUs, 48 hrs, 5<br>1024 GPUs, 12 hrs, 2 | 300K GPU hours |
| Workflow 3 | SDSC - Comet | 64 GPU, 6hrs, 50 | 20K GPU hours |

### D. Team and Team Preparedness

The team is comprised of practising computational scientists and computer scientist who work on a diverse set of high-performance computing platforms, including but not limited to the DOE and NSF leadership machines as well as a wide range of supercomputers (e.g., XSEDE, NERSC, SuperMUC etc.) Each Investigator on their own is an awardee of a substantial (peer reviewed) computational allocation. In this project we are converging to solve an end-to-end problem that standalone, no one could solve. The team has collective experience on a range of supercomputers, including many of those being requested (e.g., Frontera, Summit, Comet). We have been recent recipients of multiple DOE INCITE, ALCC, NSF PRAC and XRAC awards. As can be seen from Table of resource requests, our workloads are modest size (typically 128 nodes), but require a significant number O(100) of such workloads. Our experience is managing the concurrent execution of a large number of workloads, as well as multi-stage workflows. Our project is "shovel-ready" and will start production runs on Day 1 of the project. In addition to listed Investigators, the primary team members include: (1) Dr. Matteo Turilli (Research Professor); (2) Dr. Heng Ma (Post-doc); (3) Andre Merzky (Staff); (4) Anda Trifan (DOE Graduate Research Fellow); (5) Dr. Hyungro Lee (Research Scientist). Each of these are active users of and developers of software for

supercomputers across the DOE and NSF complex. We note that we will request Lassen (LLNL) for team members with proper security or nationality credentials.