Beyond the Scales: A physics-informed machine learning approach for more efficient modeling of SARS-CoV-2 spike glycoprotein

David Liang (dliang7234@gmail.com)
Ward Melville High School https://orcid.org/0000-0003-3895-9228
Ziji Zhang
Stony Brook University
Miriam Rafailovich
Stony Brook University
Marcia Simon
Stony Brook University
Yuefan Deng
Stony Brook University
Peng Zhang
Stony Brook University

Research Article

Keywords: coarse-graining, machine learning, multiscale modeling, SARS-CoV-2, S-protein, molecular dynamics

Posted Date: October 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1011812/v1

Beyond the Scales: A physics-informed machine learning approach for more efficient modeling of SARS-CoV-2 spike glycoprotein

David Liang[1*], Ziji Zhang[2], Miriam Rafailovich[3], Marcia Simon[4], Yuefan Deng[2], and Peng Zhang[2]

Ward Melville High School[1], Departments of Applied Mathematics and Statistics[2], Materials Science and Chemical Engineering[3], and Oral Biology and Pathology[4], Stony Brook University, NY 11790, United States

ABSTRACT

This paper presents a physics-informed machine learning approach to the derivation of a bottom-up coarse-grained model of the SARS-CoV-2 spike glycoprotein from all-atomic molecular dynamics simulations. The machine learning procedure employs a force-matching scheme in the optimization of interaction parameters, where the force-matching scheme is combined in methodology with the initialization of the interaction parameters by the traditional iterative Boltzmann inversion method. The force-matched machine learning procedure is constructed based on two physics-informed layers: one is the Harmonic layer consisting of bond, angle, and dihedral terms as bonded potentials; the other is the Lennard-Jones layer consisting of the non-bonded Lennard-Jones potential. Coarse-grained validation simulations are performed with the learned parameters to test the derived bottom-up coarse-grained model. The simulations are able to reach the microsecond time scale with stability. The physics-informed learning approach yields simulation speeds nearly 40,000 times faster than conventional all-atomic simulations while maintaining comparable simulation accuracy. Additionally, through examination of the non-bonded Lennard-Jones parameters and the radial distribution function analysis, the learning approach matches pairwise distances of the ground-truth data with greater accuracy than the conventional iterative approach method.

Keywords—coarse-graining, machine learning, multiscale modeling, SARS-CoV-2, S-protein, molecular dynamics.

1. INTRODUCTION

The outbreak of SARS-CoV-2 in 2019 and its continued persistence have led to millions of deaths globally [1], initiating great investigation efforts on its molecular structure and mechanisms of infection. There exist numerous unique spike glycoproteins that cover the outer surface of the virion. These proteins function as glycosylated trimers largely responsible for the binding of the virus to the host cell receptor angiotensin-converting enzyme 2 (ACE2), mediating cell entry [2]. The S-protein of the virus is comprised of two domains: the first one is the S1 domain which is on the top of the S-protein and contains the receptor-binding domain (RBD) responsible for binding to ACE2 as an initial step toward cell entry; the second one is the S2 domain that forms the stalk of the S-protein, which mediates cell-fusion and integration of the virus into host cells [3-6].

Efforts toward analyzing conformational states, binding, and functions of the virus have been critical to uncovering developments with regards to therapeutics and vaccines [7]. Currently, prominent techniques of electron microscopy and X-ray crystallography have been proven useful in unveiling protein structural models at high resolutions [8, 9]. All-atomic molecular dynamics (AAMD) simulations have provided a vital insight into the biological properties of systems down to the atomic- and femtosecond-level resolution, deepening understanding into the biological functions, interactions with other molecules, and conformational states [10, 11]. Studies are currently underway in uncovering specific mechanisms of action of the SARS-CoV-2 virion or possible therapeutics [12-14]. However, many such practical and large-scale applications of AAMD simulations are challenged by the computational expense when dealing with these larger proteins or structures. The S-protein, containing over twenty thousand atoms, is no exception to this limitation [15].

Because of the complexity of the S-protein, multiscale modeling (MSM) [16-18], which has been shown advantageously on modeling large molecules over greater timescales, is adapted to this investigation. In the MSM framework, coarse-graining (CG) has been well established within literature for simulating complex, high-definition systems using simplified, lower-resolution representations, thus simulating with greater computational efficiency [19-22]. There exists a broad spectrum of coarse-graining resolutions, resulting in various definitions of models of interactions. Among these definitions are "physics-based" approaches that utilize all-atom models of interactions in defining united-atom potentials [23, 24] and "knowledge-based" approaches that derive interaction models based on known protein structures [25-27].

Our motivation is to derive "physics-informed" coarse-graining models to enable the stable simulation of the SARS-CoV-2 S-protein across more relevant spatial and temporal scales that are inaccessible to conventional all-atomic simulations. Our physics-informed CG models are categorized as the aforementioned "physics-based" approach.

In this work, we construct a "bottom-up" CG model of the SARS-CoV-2 S-protein from classical all-atomic models of interactions. While "physics-based" transferable CG models like MARTINI [28] have shown feasibility in a variety of applications, we use more aggressive coarse-grain modeling which involves molecular species/groupings specific to the mapping of the S-protein structure, i.e. a bottom-up CG model. We derive and optimize the interaction potentials of the CG model using a combination of iterative and physics-informed machine learning (ML) approaches. The AAMD simulations are conducted on powerful supercomputers to help obtain large amounts of data to derive the associated CG potentials. The experimental outcomes show computing speed nearly 40,000 times faster than that of the AAMD simulations while retaining comparable structural accuracy. The contributions of this work can be summarized as follows.

- We present an innovative application of supervised ML in the derivation of a CG model;
- We combine ML with molecular dynamics towards greater efficiency, achieving a speed of coarse-grained simulations of 40,000 times faster than that of conventional all-atom models while maintaining comparable structural accuracy;
- The greater efficiency presents the timeliness of the research in producing long-term simulations of SARS-CoV-2 that can help illuminate protein mechanisms and impacts of environmental changes.

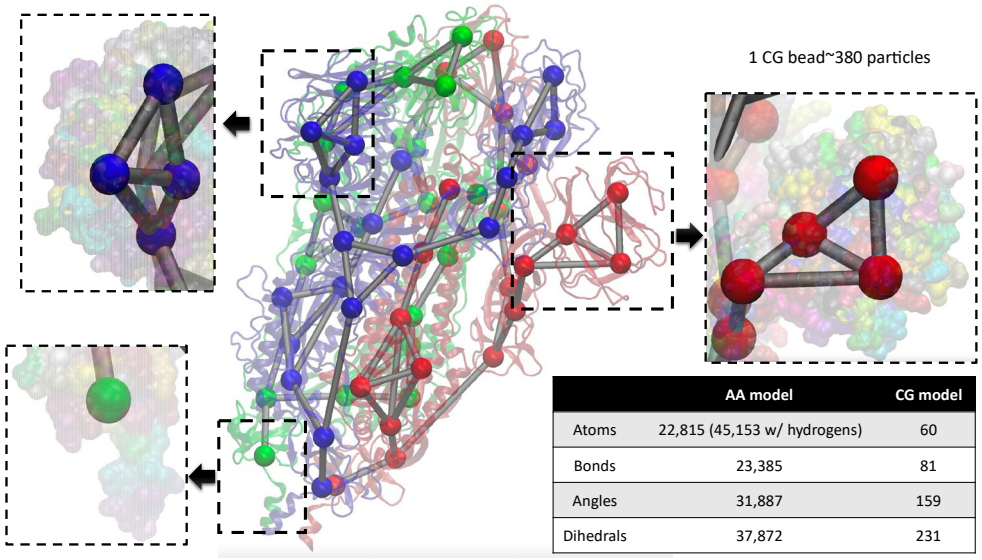
The rest of this paper is organized as follows. Section 2 details the proposed physics-informed bottom-up CG model based on the ideas introduced above and the implementation of the model to conduct the simulations. Section 3 reports the experimental outcomes of our CG model with comparison to the AAMD simulations. Section 4 draws conclusions from the reported results, followed by discussions of our future research topics along the direction of multiscale modeling of biomolecular systems.

2. METHODOLOGY

In the sections of (2.1) and (2.2) below, some background information is given to help the presentation of our physics-informed bottom-up CG model in section (2.3).

2.1. Coarse-Grained Structure Development

The full SARS-CoV-2 S-protein model was obtained from the protein data bank 6VXX and was run through Nanoscale Molecular Dynamics (NAMD) software [5, 29]. The all-atom system consists of 22,815 atoms in the 6VXX structure, with a total of 45,153 atoms including the hydrogens during the simulation. We utilized the Shape-Based Coarse Graining approach [30] to aggressively reduce the model to 60 atoms, maintaining the homotrimeric structure with 20 atoms per chain, see Fig. 1. The coarse-grained beads were distributed using a self-organizing neural network with a stochastic learning algorithm. The learning parameters are as follows: we used 12000 learning steps with an initial epsilon value of 0.3 and an initial lambda value of 12.0; we used a final epsilon value of 0.05 and a final lambda value of 0.01; the bonds were determined from the all-atom structure.



	AA model	CG model
Atoms	22,815 (45,153 w/ hydrogens)	60
Bonds	23,385	81
Angles	31,887	159
Dihedrals	37,872	231

Fig. 1: Illustration of structural visualization and statistics for all-atom vs. CG models.

2.2. All-Atomic Simulations

To obtain the ground-truth data for our learning procedure, we first conducted all-atomic simulations. Our simulations were conducted on the AiMOS supercomputer, a heterogeneous system architecture that includes IBM POWER9 CPUs connected to NVIDIA GPUs, and the Seawulf computing cluster at Stony Brook University. We utilized the CHARMM-36 force field [31] in describing the S-protein system in a vacuum canonical ensemble at 310K. The simulation began with 10 ps of energy minimization and was run for 400 ps to reach a stable state with a 1 femtosecond timestep. Instead of running on a single long-term simulation, we ran two hundred short simulations in parallel, with starting states randomly sampled and transformed from an equilibrated trajectory. Random frame sampling within the stable state in addition to random translational and rotational transformations was used to generate initial states for two hundred separate subsequent simulations. From these simulations, frames containing coordinate and force data were collected every one femtosecond. We accumulated a total of 9.7905 nanoseconds of the ground-truth all-atom data, which upon mapping yielded 97,905 data frames of coordinates and forces, representing our ground-truth data.

2.3. Coarse-Grained Modeling

In addition to the developed coarse-grained structure and constructed ground-truth data, our proposed physics-informed CG model follows a serial multiscale approach, as shown by Fig. 2. In the first box of "Data Collection", spatial and temporal mapping schemes are employed to map the AAMD simulations to the reduced-resolution coarse-grained structure. In the second box of "Parameter Optimization", the resulting ground-truth data is used to generate the initial parameterization of the CG model. The model parameters are then further refined by the physics-informed supervised ML approach. The initialization strategy employs the iterative Boltzmann inversion (IBI) approach on the bonded parameters [32, 33] and the Visual Molecular Dynamics (VMD) software for the non-bonded terms [34]. The ML approach employs a force-matching scheme using the ground-truth position and force data. In the last box of "Validation Analysis", the learned parameters are implemented in a coarse-grain molecular dynamics (CGMD) validation simulation for comparison with a separate continuous long-term AAMD vacuum simulation. The simulations are evaluated and compared in terms of simulation accuracy and computation speed. More details are provided in the following sections.

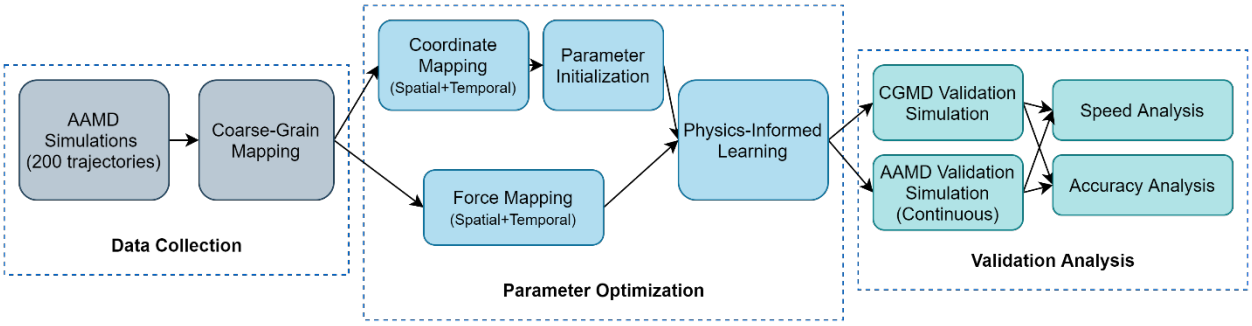


Fig. 2: Illustration of our coarse-grain modeling pipeline including data collection, parameter optimization, and validation analysis.

2.3.1. Coarse-grained mapping

The raw trajectory obtained in sections (2.1) and (2.2) was processed by mapping the extracted atomic coordinate and force data to the coarse-grained scale. Spatial mapping was conducted by computing the center of mass and the sum of forces for each atom group, constituting a bead, according to the following equations:

$$X_{I,CG} = \frac{\sum_i w_i x_{i,AA}}{\sum_i w_i} \quad (1)$$

$$F_{I,CG} = \sum_i f_{i,AA} \quad (2)$$

where $X_{I,CG}$ and $F_{I,CG}$ represent the calculated position and force, respectively, of bead I, $x_{i,AA}$ and $f_{i,AA}$ represent the position and force, respectively, of atom i within the atom group constituting bead I, and w_i represents the mass of atom i as a weighting factor.

In addition to the spatial mapping, temporal averaging was performed by the averaging of both coordinates and forces across the temporal dimension every 100 frames. Using this mapping scheme, we processed the ground-truth data for the initialization of our model parameters.

2.3.2. Parameter initialization

We utilized the traditional IBI method in generating initial bonded values or weights for our proposed ML network to train the data. Diverging from the conventional implementation, we incorporated the refinement of dihedral parameters in addition to the bonds and angles. From the ground truth simulations, we were able to extract distribution functions $P(x)$ of variable x representing the bond lengths, bond angles, or torsion angles. Utilizing the Boltzmann relation:

$$U(x) = -k_B T \ln P(x) \quad (3)$$

we obtained *U(x)* representing the potential function, where k_B is a parameter and *T* represents the temperature. Furthermore, the bonded parameters can be modeled as harmonics:

$$U(x) = \frac{1}{2}k(x - x_0)^2 \quad (4)$$

where x_0 represents the respective equilibrium measurement and k represents the harmonic constant. Thus, we can illustrate the relation between distribution functions and harmonic constants as follows:

$$\langle x^2 \rangle - \langle x \rangle^2 = \frac{k_B T}{2k} \quad (5)$$

where the equilibrium measurement x_0 is equal to the average position $\langle x \rangle$. A coarse-grain simulation was conducted as a trial run using these initial guesses, where the Boltzmann inversion relation was once again applied to extract the potentials. To update the potentials to match the reference or ground-truth distribution, constants were scaled to match the trial distribution with the reference. The process is an iterative refinement until trial distribution matches the reference within some tolerance.

Additionally, we utilized the VMD software in deriving the initialization of the Lennard-Jones (LJ) interaction potential. In this procedure, each bead i was assigned an LJ strength ϵ_i based on

$$\epsilon_i = \epsilon_{max} \left(\frac{SASA_i^{hphob}}{SASA_i^{tot}} \right)^2 \quad (6)$$

where $SASA_i^{hphob}$ and $SASA_i^{tot}$ represent the hydrophobic and total solvent-accessible surface areas of domain i, respectively, and ϵ_{max} is the user-controlled maximum energy for LJ potential well-depth, which was selected to be 20 kcal/mol. The LJ potential radius r_i (with the minimum of R_{min_i}) is given by the radius of gyration of the group of atoms constituting bead i, which is increased by a user-defined addition, e.g. an increment of 1 Å was selected in this work. The LJ energy constant ϵ_{ij} and the equilibrium distance $R_{min_{ij}}$ for a pairwise interaction are defined in Eq. (7) and Eq. (8), respectively. These parameters thus constitute the LJ potential, U_{LJ}, between pairs of beads as shown in Eq. (9).

$$\epsilon_{ij} = \sqrt{\epsilon_i * \epsilon_j} \quad (7)$$

$$R_{min_{ij}} = \frac{R_{min_i}}{2} + \frac{R_{min_j}}{2} \quad (8)$$

$$U_{LJ} = \epsilon_{ij}\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - 2\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right] \quad (9).$$

2.3.3. Physics-informed model

We constructed a force-matching model as a means of preserving thermodynamic consistency in minimizing the error between the instantaneous ground-truth forces and predicted forces [35-38]. The error is defined as:

$$Loss = \langle (F_{CG} + \nabla U_{CG})^2 \rangle \quad (10)$$

where F_{CG} represents the instantaneous force and U_{CG} represents the coarse grain potential and ∇ is the gradient operator.

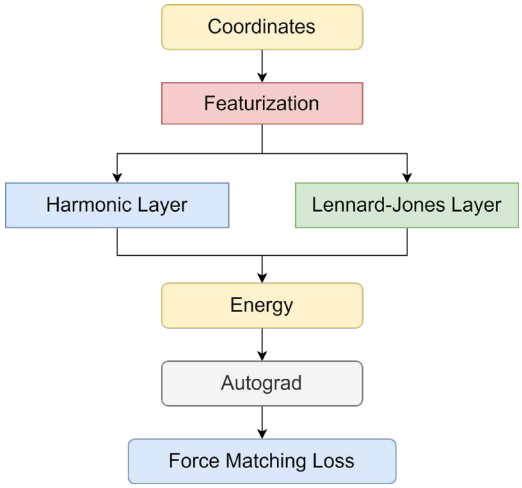


Fig. 3: Physics-informed model architecture.

Firstly, the model contains an initial featurization layer that provides the physics-informed layers with the pairwise distances, bond lengths, bond angles, and torsion angles extracted from the input coordinates, as displayed in Fig. 3. The model was constructed based on two physics-informed layers: one is the Harmonic layer comprised of bond, angle, and dihedral terms as bonded potentials; and the other is the Lennard-Jones layer consisted of the LJ potential accounting for the weak dipole attraction between distant atoms and the hard-core repulsion between close atoms. Within the Harmonic layer, the trainable parameters include the harmonic constants for each of the aforementioned bonded potentials, whereas, in the Lennard-Jones layer, the trainable parameters are the bead strength ϵ_i and the minimum radius, R_{min_i}, for each unique bead i.

There exist 471 and 40 trainable parameters that comprise the bonded and non-bonded interactions, respectively, in the physics-informed model. The parameters are associated with the Lennard-Jones layer or LJ potential of Eqs. (7)-(9) and the Harmonic layer with the dihedral potentials of Eqs. (11) or (12), both of which will be presented in the following paragraphs.

For the dihedral potentials, we represented them in two ways: harmonic representation of Eq. (11) and periodic representation of Eq. (12). The harmonic form represents the dihedral potential in the same manner as bonded

potentials, where the trainable constants are analogous to spring constants. The periodic representation accounts for the periodicity of dihedrals, where the phase shift angle, ϕ, was adjusted to fit the equilibrium value as the potential minima.

We visualize the torsion angle distribution to depict the unimodality in Fig. 4 and thus confirm our choice of $n = 1$ as the multiplicity for the periodic representations, as well as $n = 0$ for the harmonic representation. The distributions in Fig. 4 present more common conformations with the yellow color, where the means are the respective equilibrium states.

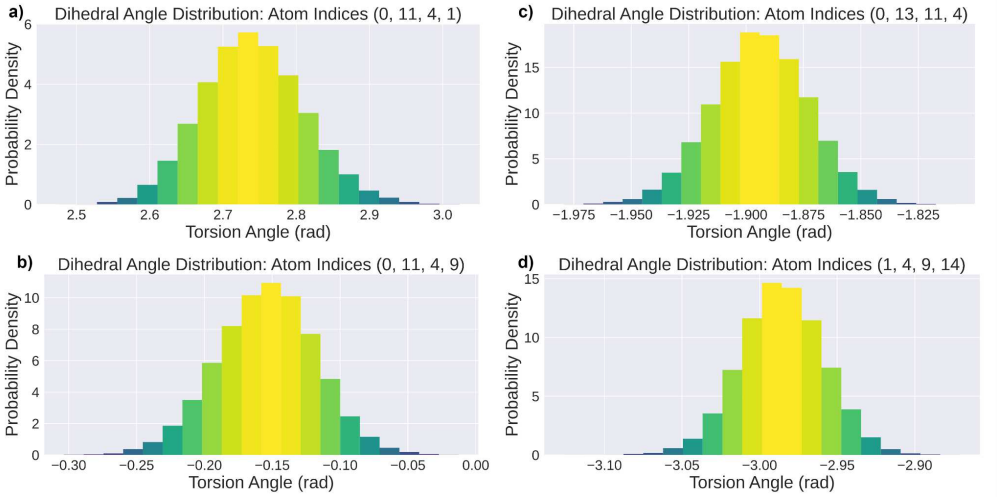


Fig. 4: Examples of torsion angle distributions for four dihedrals of atom indices of (a): (0, 11, 4, 1); (b): (0, 11, 4, 9); (c): (0, 13, 11, 14); and (d): (1, 4, 9, 14).

Based on the above analysis, the resulting energy in Eq. (10) can be calculated according to:

$$U_{CG} = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_a(\theta - \theta_0)^2 + \sum_{dihedrals} k_d(\psi - \phi)^2 + \sum_{i<j}^{atoms} \epsilon_{ij} \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] \quad (11)$$

$$U_{CG} = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_a(\theta - \theta_0)^2 + \sum_{dihedrals} k_d(1 + cos(n\psi - \phi)) + \sum_{i<j}^{atoms} \epsilon_{ij} \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] \quad (12)$$

where k_a, k_b and k_d are constant parameters, r is bond distance, $\theta$is bond angle, ψ is torsion angle, and ϕ was defined above as the torsion phase shift angle, which acts as an equilibrium angle in the harmonic representation, see Eq. (11).

The calculated energy was passed through a gradient layer to compute the derivatives with respect to the input coordinates, thus the force predictions $\hat{f}$ can be calculated by Eq. (13):

$$\hat{f} = -\nabla_x U_{CG} \quad (13).$$

From these predictions, we can determine the loss as a mean-squared error function between the calculated and the mapped ground-truth forces, as given in Eq. (10). Validation of our above presented physics-informed bottom-up CG approach is described in the following section. The validation analysis is shown as the third box of the pipeline of Fig. 2.

2.4. CGMD Validation

To measure the performance of our approach, we evaluated a validation CGMD simulation across the metrics of accuracy and speed. Using the optimized parameters, we ran the CGMD simulations of one microsecond for both the harmonic and periodic dihedral parameter sets respectively.

With regards to accuracy analysis, torsional analysis was applied in the form of free energy surface plots and the free energy was plotted along two dihedral quadruples, where the free energy profiles provide insight into the conformational states. From the plots, we compared our validation simulations with the ground-truth training data using the dihedral pairs belonging to S-protein RBD and S2 domain. Radial distribution functions (RDF) were applied as well in providing insight into the distribution of particles around certain particles. Thus, we can utilize these plots as a mode of comparison between the ground-truth training AAMD simulations (i.e. the reference baseline for comparison purposes) and the CGMD validation simulations (i.e. our presented CGMD model) for evaluating structural accuracy.

Additionally, we incorporated the root-mean-square-deviation (RMSD) analysis to monitor the structural stability of the compared models throughout their respective trajectories. In this analysis, we ran a separate AAMD vacuum simulation (as the baseline for the purpose of comparison) in order to obtain a continuous trajectory of data.

In addition to the accuracy analysis, we further performed speed analysis as a measure of the overall model's performance and efficiency. Using the Seawulf cluster, we ran the CGMD validation simulation for 5 microseconds and compared its simulation speed with the continuous 0.1 nanosecond AAMD validation simulation (i.e. the one used in our RMSD analysis as the reference baseline).

The validation experimental outcomes are reported in the section below.

3. RESULTS

In section (3.1) below, we will focus on the performance of our machine learning for those parameters associated with our CG model. The learned parameters will be compared to the reference distributions derived from the ground-truth data. With the learned model parameters, we will report the accuracy of our CGMD model simulations vs. the baseline AAMD model simulations in section (3.2) and the simulation speed in section (3.3).

3.1. Parameter Learning

Using the obtained 97,905 coordinates and force frames, the parameter initialization for bonds, angles, and dihedrals, respectively, proceeded with 3 iterations. For each iteration, the trial simulations were conducted with 10 femtosecond timesteps, minimized for 500 picoseconds, and simulated for 4 ns. Figure 5 illustrates the IBI refinement of the parameters to match the reference distributions to reasonable accuracy after three iterations.

Fig. 5: Illustration of IBI method's initialization of parameters (harmonic dihedral representation) for 3 iterations. Bonds (a). Angles (b). Dihedrals (c).

There exist 511 total learnable parameters, including the LJ potential parameters, which are learned with the physics-informed model configured with Adam optimizer with a learning rate of 0.001, a batch size of 256 for 10 epochs. Figure 6 displays the loss plot over the training process. Both training and validation losses approached convergence after 4 epochs. The optimization of each individual bonded and non-bonded parameter over the 10 epochs is visualized in Fig. 7. The results of Fig. 7 agree very well with that of Fig. 6.

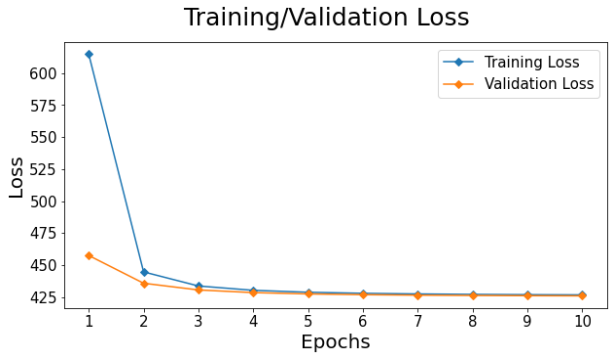


Fig. 6: Training and validation loss vs. epochs.

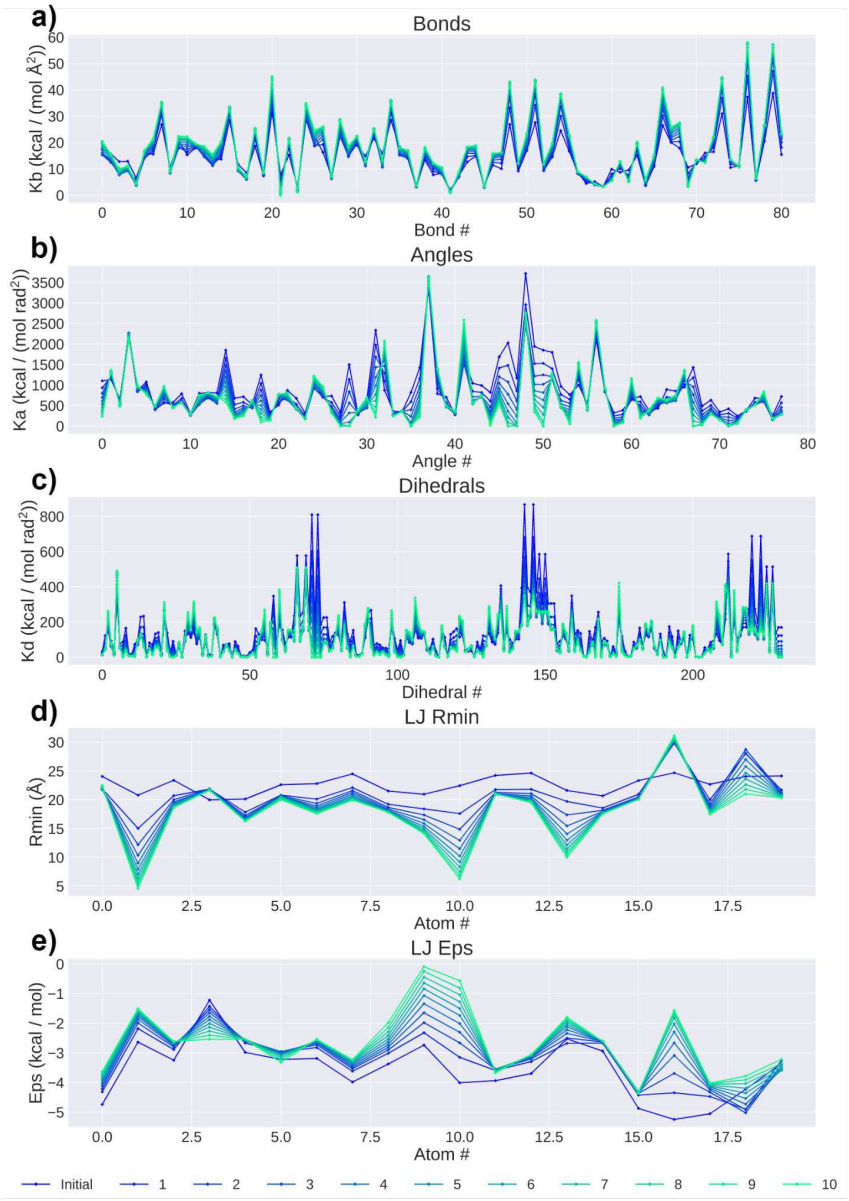


Fig. 7: Physics-informed parameter learning (harmonic dihedral representation) for over 10 epochs. Bonds (a). Angles (b). Dihedrals (c). LJ Rmin (d). LJ Epsilon (e).

3.2. CGMD Validation: Accuracy

To test the efficacy of our CG model, we conducted validation simulations using the learned parameters of section (3.1). The simulations were configured with a timestep size of 10 femtoseconds, as opposed to the 1 femtosecond time-step used in our ground-truth all-atom simulations. The resulting simulations, for both the harmonic and periodic dihedral representations, reached stability for one microsecond. We used this data to conduct our accuracy analysis (as well as speed analysis to be presented in section (3.3) below) in comparison to the ground-truth data. We evaluated the validation accuracy using the torsion analysis, the RDF plots, and the RMSD analyses.

3.2.1. Torsion analysis

During our accuracy analysis, we evaluated the free energy profiles as a function of dihedral angles. The plots were used to analyze and compare the torsion angles as a representation of the protein conformational states. We

display two separate pairs of torsional angles for such analysis: one is located on the receptor-binding domain, and the other is located in the S2 subunit, see Fig. 8.

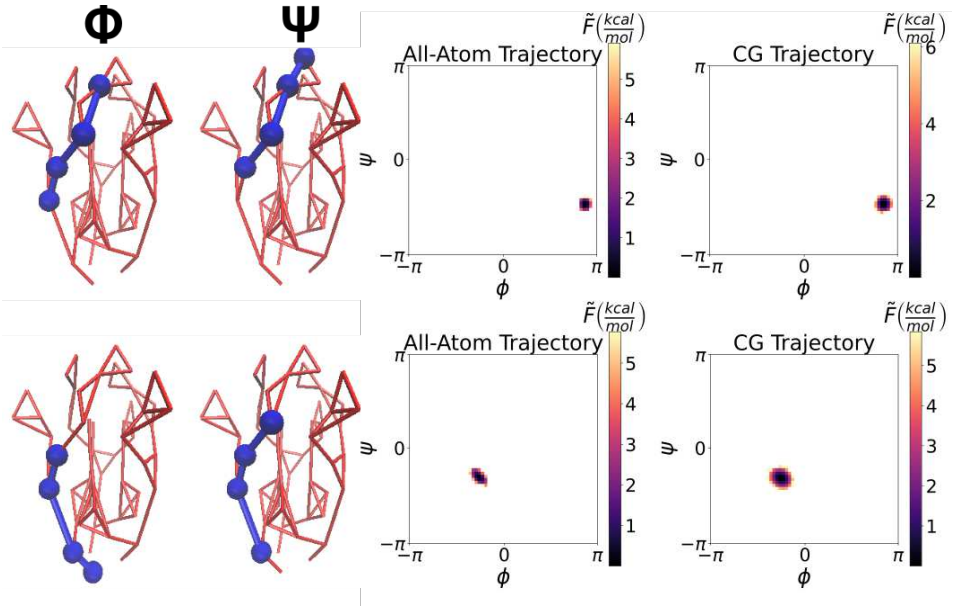


Fig. 8: Free energy profiles of RBD pair (top) and S2 subunit pair (bottom). (Harmonic dihedral representation).

Figure 8 demonstrated that the resulting torsion analysis on our CGMD simulations matches the ground-truth training data (or the AAMD simulations) with high accuracy.

3.2.2. Radial-distribution function measure

We implemented the RDF measure to compare the mapped ground-truth statistics (or the AAMD simulations) with the performance of our CGMD simulations, see Fig. 9 and Fig. 10.

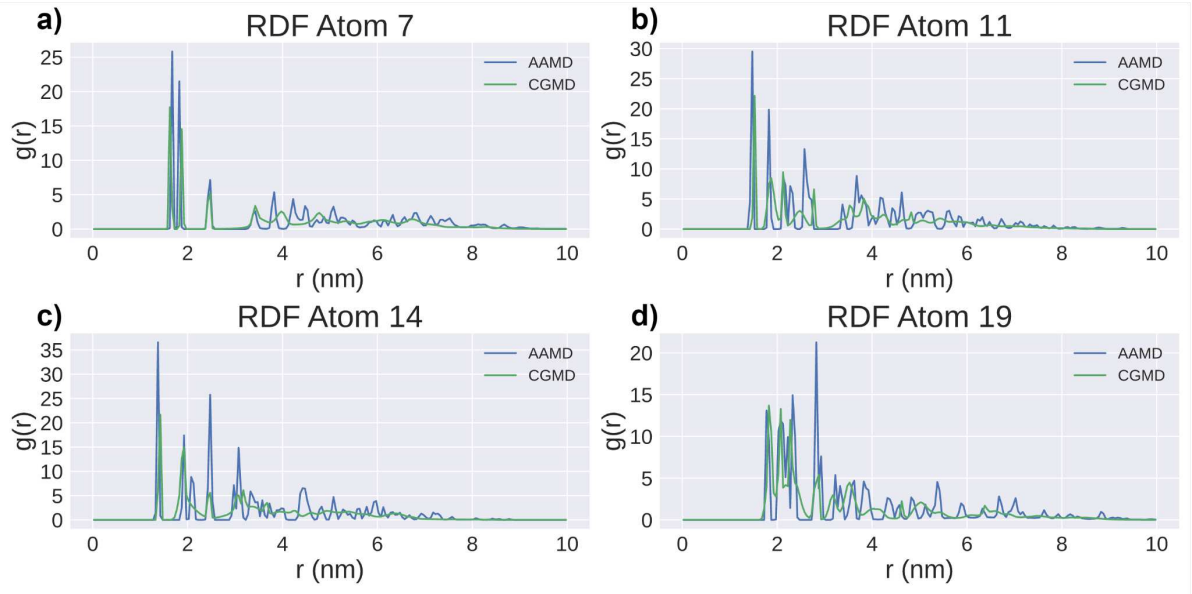


Fig. 9: RDF plot from single reference atom comparison of our CGMD simulations vs. the baseline continuous AAMD simulations. (Harmonic dihedral representation). Atom 7 vs all (a). Atom 11 vs all (b). Atom 14 vs all (c). Atom 19 vs all (d).

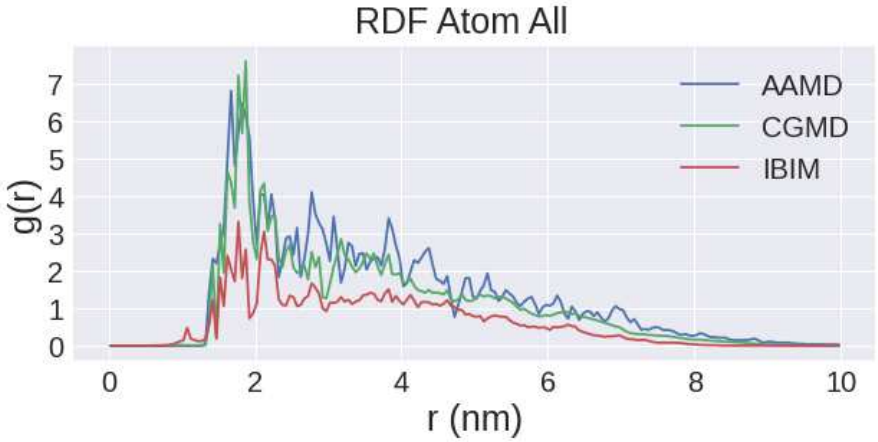


Fig. 10: RDF plot of all atoms for comparison of CGMD vs AAMD. (Harmonic dihedral representation).

As illustrated in Fig. 9 and Fig. 10, our physics-informed machine learning approach is able to resemble the reference plots with reasonable accuracy, as it can capture the more significant peaks within the RDF. However, there does indicate some loss of resolution and accuracy in the smaller peaks at larger distances. Investigation on this observation of some loss in some small peaks will be one of our future research topics.

In addition to the use of RDF measures on the statistics above, we further expanded the RDF measures on the performance of our ML procedure for model parameterization. As seen in section (3.1) above, the ML procedure indicates very noticeable refinements on the model parameterization, particularly on the non-bonded LJ potential terms. Within this refinement is the very noticeable decrease in both the epsilon and the associated well-depth terms. Upon further investigation, it is shown that the model's calculated energies begin with positive repulsion, gradually becoming negative by the end of the training, demonstrating the proper optimization to match the distances of the ground-truth data. In comparison with the IBI trial results above in section (3.1), specifically on the atom pair between atom numbers 18 and 47, the learned distances are more consistent with the ground-truth result, as shown in Fig. 11.

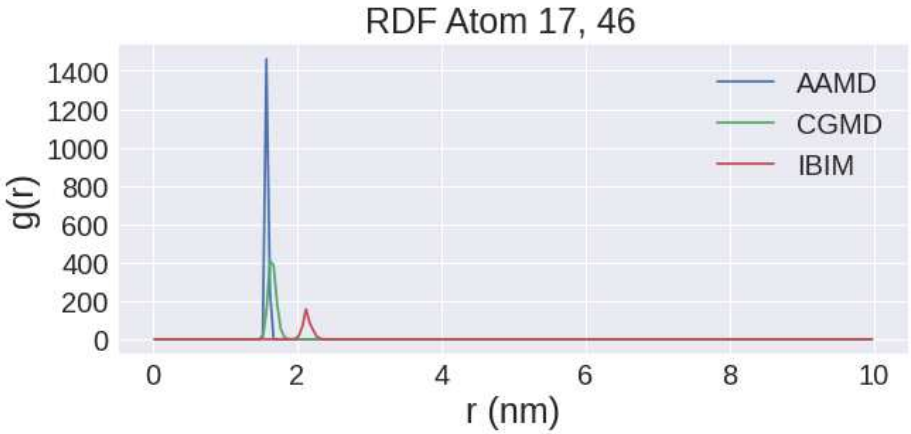


Fig. 11: RDF plot between the atoms 17 and 46.

Within the RDF plots, our CG model is able to capture the positions and peaks in the respective pairs with comparable accuracy to the ground-truth model. Further analysis indicates stability for the entirety of the microsecond, indicating the feasibility of our approach for long-term modeling of the SARS-CoV-2 S-protein.

3.2.3. RMSD

We analyzed the evolution of our presented CG model trajectory by calculating the RMSD values using the starting structure as a reference frame, see Fig. 12. The RMSD reveals the overall stability and conformational change of the whole protein. Protein coordinates were recorded every 10 picoseconds and the RMSD was calculated on the aligned trajectory. Figure 12 presents the RMSD of our CGMD simulations alongside the baseline continuous AAMD simulations.

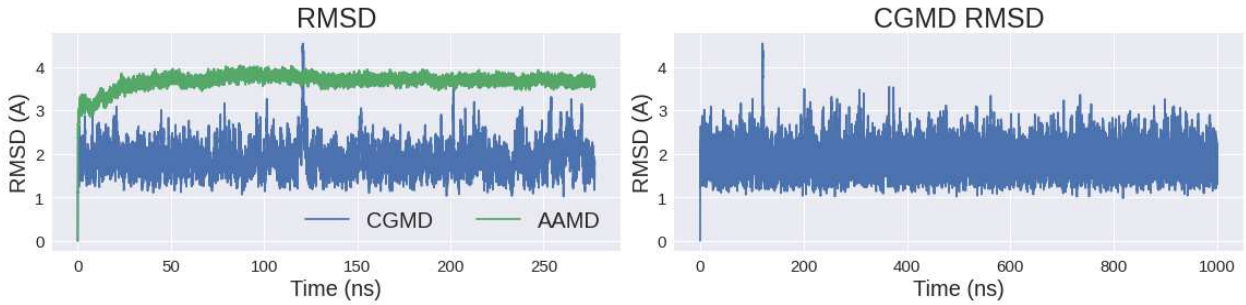


Fig. 12: RMSD comparison between our CGMD simulation and the reference AAMD simulation (right). RMSD of one microsecond CGMD simulation (left). (Harmonic dihedral representation)

Our presented CGMD simulations appear to have greater fluctuations in comparison to the reference AAMD simulations but remain consistent throughout the full microsecond of simulation, indicating long-term structural stability. Investigation on this observation of fluctuations will be another topic of our future research interests. It is worth noting that convergence to a higher value is a known drawback of the RMSD metric and, therefore, does not indicate inaccuracy within our model.

In addition to the RMSD analysis above, we also visualized the protein structural conformations of the starting and ending states of the AAMD and CGMD simulations as shown in Fig. 13. The protein maintains the overall structure, however, the slight deviations in the AAMD and CGMD structures will be investigated in a future study.

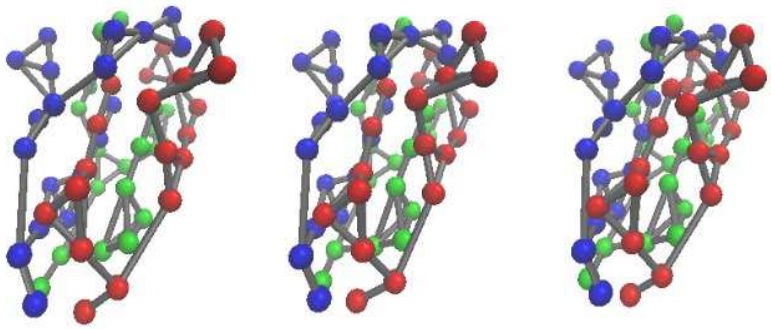


Fig. 13: CGMD structure visualization. Starting state (left). AAMD trajectory final frame state (middle). CGMD trajectory final frame state (right).

The following section will focus on the comparison studies in terms of simulation efficiency.

3.3. CGMD Validation: Speed

Both the reference AAMD ground truth simulations and our CGMD validation simulations were conducted on the Seawulf cluster, where each computing node consists of two Intel Xeon E5-2690v3 CPUs. By using the parallel NAMD package on 1 node with 24 CPU cores, the AAMD simulations with 1 femtosecond as time step size produced 0.243 nanoseconds/day while the CGMD simulations with 10 femtoseconds as time step size produced 9532.6 nanoseconds/day. The experimental outcomes indicate that our CGMD validation simulations have a speed nearly 40,000 times faster than that of the reference AAMD simulations. Detailed measurements are presented in Table 1.

Table 1: Validation simulation comparisons using 24 CPU cores.

Simulations	Time step size	Total steps	Simulated time	Simulating time	Simulation speed
AAMD	1 fs	100,000	0.1 ns	35,557 s	0.243 ns/day
CGMD	10 fs	500,000,000	5 μs	45,318 s	9,532.6 ns/day

4. CONCLUSION AND DISCUSSION

This work represents an implementation of artificial intelligence-enabled modeling towards more efficient multiscale CGMD validation simulations. The presented physics-informed ML approach to the model parameterization includes two phases: firstly, using all-atom simulations to generate the ground-truth data for parameter learning; secondly, using the learned parameters to run long-term coarse-graining simulations. Our proposed physics-informed bottom-up CGMD model simulations were compared to the ground-truth AAMD model simulations.

The comparison between our coarse-graining modeled validation simulations and the mapped all-atom modeled simulations yields high accuracy in agreement within the free energy profiles, indicating a resemblance of the conformation.

Aside from the accuracy, our simulation model is significantly faster than the all-atom simulation model. With the aggressive coarse-graining approach, our model is able to achieve the simulation speed nearly 40,000 times faster than that of the all-atom simulations. This significant speedup lends itself to more aggressive coarse-graining of the protein structure, or more accurately reproducing the structural properties of the S-protein compared to other CG models. Our work presented here underscores the following contributions toward more efficient multiscale modeling:

- Our approach demonstrates feasibility and advantages with the application of supervised ML in the derivation of a coarse-graining model;
- In combining ML with molecular dynamics, we immensely accelerate simulation speed compared to conventional all-atom models while maintaining stability and structural accuracy;
- The gained efficiency can elucidate protein mechanisms and render a great impact on future simulational studies of environmental changes by relieving the ongoing concerns about timeliness.

During the validation simulations, we observed some deviations from our expectation and investigations on these deviations will be our future research tasks:

- The observation of some loss in small peaks in the RDF analysis;
- The observation of the fluctuations in the RMSD analysis;
- The observation of the slight deviations in the AAMD and CGMD structures.

In addition, the application of this approach in simulating the S-protein under various environmental conditions, including solvation, will be investigated as another future research topic.

ACKNOWLEDGEMENT

The project is supported by the SUNY-IBM Consortium Award, IPDyna: Intelligent Platelet Dynamics, FP00004096 (PI: Y. Deng, Co-I: P. Zhang). All simulations were conducted on the AiMOS supercomputer at Rensselaer Polytechnic Institute through an IBM Faculty Award FP0002468 (PI: Y. Deng), and the Seawulf cluster at Stony Brook University.

The project is sponsored by Stony Brook University's OVPR & IEDM COVID-19 Seed Grant, PIs: P. Zhang, Y. Deng, M. Rafailovich, and M. Simon.

DATA AVAILABILITY

All original data are available upon request.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by David Liang, Ziji Zhang and Peng Zhang. Funding support and research guidance were given by Peng Zhang, Yuefan Deng, Miriam Rafailovich and Marcia Simon. The first draft of the manuscript was written by David Liang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

DECLARATIONS

Ethics approval: The manuscript is prepared in compliance with the Ethics in Publishing Policy as described in the Guide for Authors.
Consent to participate: The manuscript is approved by all authors for publication.
Consent for publication: The consent for publication was obtained from all participants.
Conflict of interest: The authors declare that they have no conflict of interest.

REFERENCES

1. Johns Hopkins Coronavirus Resource Center (n.d.). COVID-19 Map - Johns Hopkins Coronavirus Resource Center. Johns Hopkins Coronavirus Resource Center.
2. Letko, M., Marzi, A., & Munster, V. (2020). Functional assessment of cell entry and receptor usage for SARS-CoV-2 and other lineage B betacoronaviruses. *Nature microbiology*, 5(4), 562–569. https://doi.org/10.1038/s41564-020-0688-y
3. Huang, Y., Yang, C., Xu, X. F., Xu, W., & Liu, S. W. (2020). Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. *Acta pharmacologica Sinica*, *41*(9), 1141–1149. https://doi.org/10.1038/s41401-020-0485-4
4. Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., & Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, *581*(7807), 215–220. https://doi.org/10.1038/s41586-020-2180-5
5. Walls, A. C., Park, Y. J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell*, 181(2), 281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058
6. Liu, S., Xiao, G., Chen, Y., He, Y., Niu, J., Escalante, C. R., Xiong, H., Farmar, J., Debnath, A. K., Tien, P., & Jiang, S. (2004). Interaction between heptad repeat 1 and 2 regions in spike protein of SARS-associated coronavirus: implications for virus fusogenic mechanism and identification of fusion inhibitors. *Lancet (London, England)*, *363*(9413), 938–947. https://doi.org/10.1016/S0140-6736(04)15788-7
7. K. Chan, S. Sridhar, R. Zhang, H. Chu, A.Y. Fung, G. Chan, J. Chan, K. To, I. Hung, V.C. Cheng, K. Yuen, Factors affecting stability and infectivity of SARS-CoV-2. *J. Hosp. Infect.* 106, 226–231 (2020)
8. Nogales E. (2016). The development of cryo-EM into a mainstream structural biology technique. *Nature methods*, *13*(1), 24–27. https://doi.org/10.1038/nmeth.3694
9. Sztain, T., Amaro, R., & McCammon, J. A. (2020). Elucidation of cryptic and allosteric pockets within the SARS-CoV-2 protease. *bioRxiv : the preprint server for biology*, 2020.07.23.218784. https://doi.org/10.1101/2020.07.23.218784
10. M. Song, P. Zhang, C. Han, Z. Zhang, Y. Deng, Long-time simulation of temperature-varying conformations of COVID-19 spike glycoprotein on IBM supercomputers, *supercomputing conference 2020 (SC20)*, Research Posters Track (2020).
11. Rath, S. L., & Kumar, K. (2020). Investigation of the Effect of Temperature on the Structure of SARS-CoV-2 Spike Protein by Molecular Dynamics Simulations. *Frontiers in molecular biosciences*, *7*, 583523. https://doi.org/10.3389/fmolb.2020.583523
12. J. He, H. Tao, Y. Yan, S. Y. Huang, and Y. Xiao, "Molecular Mechanism of Evolution and Human Infection with SARS-CoV-2," (in eng), *Viruses,* vol. 12, no. 4, Apr 10 2020, doi: 10.3390/v12040428.
13. Babuji, Y., Blaiszik, B., Brettin, T., Chard, K., Chard, R., Clyde, A., ... & Wagner, R. (2020). Targeting SARS-CoV-2 with AI-and HPC-enabled lead generation: a first data release. *arXiv preprint arXiv:2006.02431*.
14. Woo, H., Park, S. J., Choi, Y. K., Park, T., Tanveer, M., Cao, Y., Kern, N. R., Lee, J., Yeom, M. S., Croll, T. I., Seok, C., & Im, W. (2020). Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *The journal of physical chemistry. B*, *124*(33), 7128–7137. https://doi.org/10.1021/acs.jpcb.0c04553
15. Liang, D., Song, M., Niu, Z., Zhang, P., Rafailovich, M., & Deng, Y. (2021). Supervised machine learning approach to molecular dynamics forecast of SARS-CoV-2 spike glycoproteins at varying temperatures. *MRS advances*, 1–6. Advance online publication. https://doi.org/10.1557/s43580-021-00021-4
16. Sheriff, J., Wang, P., Zhang, P., Zhang, Z., Deng, Y., Bluestein, D., "In Vitro Measurements of Shear-Mediated Platelet Adhesion Kinematics as Analyzed through Machine Learning", *Annals of Biomedical Engineering* 2021. DOI: 10.1007/s10439-021-02790-3
17. Zhang, Z., Zhang, P., Han, C., Cong, G., Yang, C-C., Deng, Y., "AI Meets HPC: Learning the Cell Motion in Biofluids", *Supercomputing Conference 2020 (SC20)*, Research Posters Track, November 16–19, 2020. DOI: 10.13140/RG.2.2.18340.40321
18. Zhang, Z., Zhang, P., Wang, P., Sheriff, J., Bluestein, D., Deng, Y., "Rapid Analysis of Streaming Platelet Images by Semi-unsupervised Learning", *Computerized Medical Imaging and Graphics*, 2021. DOI: 10.1016/j.compmedimag.2021.101895

19. Moore, T. C., Iacovella, C. R., & McCabe, C. (2014). Derivation of coarse-grained potentials via multistate iterative Boltzmann inversion. *The Journal of chemical physics*, 140(22), 224104. https://doi.org/10.1063/1.4880555
20. Leong, T., Voleti, C., & Peng, Z. (2021). Coarse-Grained Modeling of Coronavirus Spike Proteins and ACE2 Receptors. *Frontiers in Physics*, NA-NA.
21. Yu, A., Pak, A. J., He, P., Monje-Galvan, V., Casalino, L., Gaieb, Z., Dommer, A. C., Amaro, R. E., & Voth, G. A. (2021). A multiscale coarse-grained model of the SARS-CoV-2 virion. *Biophysical journal*, 120(6), 1097–1104. https://doi.org/10.1016/j.bpj.2020.10.048
22. Ramabadran, A., Narayanan, A., Zhang, D., Zhang, Z., Simon, M., Rafailovich, M., Deng, Y., & Zhang, P. (2021). Coarse-grained modeling for efficient simulation of SARS-CoV-2 spike glycoprotein. *American Chemical Society (ACS)*. https://doi.org/10.1021/scimeetings.1c00157
23. Scheraga, H. A., Khalili, M., & Liwo, A. (2007). Protein-folding dynamics: overview of molecular simulation techniques. *Annual review of physical chemistry*, 58, 57–83. https://doi.org/10.1146/annurev.physchem.58.032806.104614
24. Kar, P., & Feig, M. (2014). Recent advances in transferable coarse-grained modeling of proteins. *Advances in protein chemistry and structural biology*, 96, 143–180. https://doi.org/10.1016/bs.apcsb.2014.06.005
25. Sippl M. J. (1995). Knowledge-based potentials for proteins. *Current opinion in structural biology*, 5(2), 229–235. https://doi.org/10.1016/0959-440x(95)80081-6
26. Carlsen, M., Koehl, P., & Røgen, P. (2014). On the importance of the distance measures used to train and test knowledge-based potentials for proteins. *PloS one*, 9(11), e109335. https://doi.org/10.1371/journal.pone.0109335
27. Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-Grained Protein Models and Their Applications. *Chemical reviews*, *116*(14), 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163
28. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P., & de Vries, A. H. (2007). The MARTINI force field: coarse grained model for biomolecular simulations. *The journal of physical chemistry. B*, *111*(27), 7812–7824. https://doi.org/10.1021/jp071097f
29. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kalé, L., & Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, *26*(16), 1781–1802. https://doi.org/10.1002/jcc.20289
30. Freddolino, P. L., Shih, A. Y., Arkhipov, A., Ying, Y., Chen, Z., & Schulten, K. (2009). Application of residue-based and shape-based coarse-graining to biomolecular simulations. *Coarse-graining of condensed phase and biomolecular systems*, 299-315.
31. Huang, J., & MacKerell, A. D., Jr (2013). CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *Journal of computational chemistry*, 34(25), 2135–2145. https://doi.org/10.1002/jcc.23354
32. Reith, D., Pütz, M., & Müller-Plathe, F. (2003). Deriving effective mesoscale potentials from atomistic simulations. *Journal of computational chemistry*, 24(13), 1624–1636. https://doi.org/10.1002/jcc.10307
33. Agrawal, V., Arya, G., & Oswald, J. (2014). Simultaneous iterative boltzmann inversion for coarse-graining of polyurea. *Macromolecules*, *47*(10), 3378-3389.
34. Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1), 33–28. https://doi.org/10.1016/0263-7855(96)00018-5
35. Izvekov, S., & Voth, G. A. (2005). A multiscale coarse-graining method for biomolecular systems. *The journal of physical chemistry. B*, 109(7), 2469–2473. https://doi.org/10.1021/jp044629q
36. Noid, W. G., Chu, J. W., Ayton, G. S., Krishna, V., Izvekov, S., Voth, G. A., Das, A., & Andersen, H. C. (2008). The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of chemical physics*, 128(24), 244114. https://doi.org/10.1063/1.2938860
37. Wang, J., Olsson, S., Wehmeyer, C., Pérez, A., Charron, N. E., de Fabritiis, G., Noé, F., & Clementi, C. (2019). Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS central science*, 5(5), 755–767. https://doi.org/10.1021/acscentsci.8b00913
38. Husic, B. E., Charron, N. E., Lemm, D., Wang, J., Pérez, A., Majewski, M., Krämer, A., Chen, Y., Olsson, S., de Fabritiis, G., Noé, F., & Clementi, C. (2020). Coarse graining molecular dynamics with graph neural networks. *The Journal of chemical physics*, 153(19), 194101. https://doi.org/10.1063/5.0026133