MRS Advances (2021) 6:362–367
https://doi.org/10.1557/s43580-021-00021-4

ORIGINAL PAPER

Supervised machine learning approach to molecular dynamics forecast of SARS-CoV-2 spike glycoproteins at varying temperatures

David Liang[1] · Meichen Song[2] · Ziyuan Niu[2] · Peng Zhang[2] · Miriam Rafailovich[2] · Yuefan Deng[2]

Received: 17 December 2020 / Accepted: 31 January 2021 / Published online: 17 February 2021

Abstract

Molecular dynamics (MD) simulations are a widely used technique in modeling complex nanoscale interactions of atoms and molecules. These simulations can provide detailed insight into how molecules behave under certain environmental conditions. This work explores a machine learning (ML) solution to predicting long-term properties of SARS-CoV-2 spike glycoproteins (S-protein) through the analysis of its nanosecond backbone RMSD (root-mean-square deviation) MD simulation data at varying temperatures. The simulation data were denoised with fast Fourier transforms. The performance of the models was measured by evaluating their mean squared error (MSE) accuracy scores in recurrent forecasts for long-term predictions. The models evaluated include k-nearest neighbors (kNN) regression models, as well as GRU (gated recurrent unit) neural networks and LSTM (long short-term memory) autoencoder models. Results demonstrated that the kNN model achieved the greatest accuracy in forecasts with MSE scores over around 0.01 nm less than those of the GRU model and the LSTM autoencoder. Furthermore, it demonstrated that the kNN model accuracy increases with data size but can still forecast relatively well when trained on small amounts of data, having achieved MSE scores of around 0.02 nm when trained on 10,000 ns of simulation data. This study provides valuable information on the feasibility of accelerating the MD simulation process through training and predicting supervised ML models, which is particularly applicable in time-sensitive studies.

Keywords Machine learning · Molecular · Simulation · Computation

Introduction

The outbreak of SARS-CoV-2 in 2019, having persisted for over a year, has led to over 70 million recorded cases and over a million deaths globally [1]. As the virus continues to spread, it is vital to better understand and explore the effects of environmental factors. Much of the work toward understanding the effects of environmental factors focuses on the infectivity and stability of the virus under varying conditions [2, 3]. In terms of the impact of temperature, the SARS-CoV-2 virus is demonstrated to be more thermally stable compared to previous SARS and MERS coronavirus outbreaks [4]. Furthermore, it is now evident that this virus has the ability to survive and spread in both the cold and warm temperatures of winter and summer. The spike glycoprotein (S-protein), a prominent structural protein found on the surface of coronaviruses, is largely responsible for virus entry into cells. Thus, the infectivity and stability of the virus can be analyzed through the modeling of the changes in the spike glycoprotein, for instance, at temperatures at which it may denature [5].

By using all-atom molecular dynamics (MD) simulations, we can model complex nanoscale interactions of the virus and explore the effects of temperature on the S-protein. However, due to the size of the protein and the sheer number of residues, the high computational requirements of all-atom MD simulations limit capabilities in long-term modeling. This work explores a machine learning (ML) solution to accelerating the MD simulation data gathering process. In this work, we use different supervised ML methods to generate predictions of long-term properties of SARS-CoV-2 S-protein through the analysis of its μs-scale backbone RMSD (root-mean-square deviation) MD simulation data at varying temperatures. These RMSD data provide us with information on the distances between the particles of the protein and can demonstrate the protein's stability over time.

✉ David Liang
dliang7234@gmail.com

[1] Ward Melville High School, East Setauket, NY 11733, USA

[2] Stony Brook University, Stony Brook, NY 11790, USA

We analyze and evaluate the performance of k-nearest neighbors (kNN) and neural network models in recurrent forecasting capabilities, highlighting their feasibility in accelerating the MD simulation data gathering process.

Experiment design

In this study, we ran full-atom MD simulations on the SARS-CoV-2 spike glycoprotein model, which was obtained from the protein data bank PDB 6VXX, through GROMACS. The structure of the protein was immersed in a water box with dimensions of $21 \times 21 \times 21$ nm^3. The total number of atoms was 805,218, of which 45,156 (5.6%) were of the protein and 760,047 (94.4%) were water molecules. The simulation was run with the CHARMM27 force field for describing the system of the S-protein and the water molecules. Simulations were conducted at varying temperatures, namely 3 °C, 20 °C, 37 °C, 60 °C, 80 °C, and 95 °C, with the isothermal-isobaric (NPT) ensemble with a time step size of 2.5 fs [6].

The RMSD data of atomic positions of the S-protein was extracted from the simulation trajectories of each temperature value. In this study, we used simulation data from 750 ns to 2500 ns, yielding 17,500 values for each temperature setting. The data processing was characterized primarily by the denoising of the data with fast Fourier transforms (FFT) through filtering the Fourier terms or frequencies from the power spectrum. As shown in Fig. 1, we denoised the data by retaining 50 frequencies with the greatest power values from the power spectrum. In addition to the data denoising, the time-series data were normalized between 0 and 1 for the training procedure. The training dataset was created by first organizing the data into input windows and output forecasts. For instance, we organized the data to have a lookback window of 300 units and a multi-step output of 50 units for the GRU (Gated Recurrent Unit) network. The first 15,000 sets of data for each temperature were compiled into a training dataset, on which the supervised ML models were trained. The supervised models consisted of a kNN model, a GRU network, and an LSTM (long short-term memory) autoencoder. The kNN model was trained on extracted features of 10 interval means and a standard deviation value from 200 step input windows, learning to predict the next time step through feature similarity. The neural network models, as shown in Fig. 2, incorporated LSTM and GRU, which aimed to capture order dependence within time-steps as a means of making accurate forecasts. The LSTM autoencoder aims at implementing the unsupervised aspect of an autoencoder to extract useful information from the input sequences. Because these models were trained on the RMSD values of multiple temperature simulations, k-means clustering was used to cluster each input window by two features of mean and standard deviation into categorical classes. A k-nearest neighbors model would then analyze feature similarity to predict the classes of input sequences in forecasting. These categorical classes were inputted and concatenated into the models in Fig. 2. These two models also incorporated multi-step outputs in their forecasts [7]. All of the supervised models were evaluated on their accuracy in recurrent forecasts in that predictions would be used as input sequences for the next prediction step. The metric used in measuring the accuracy was (MSE) mean squared error (MSE).

Furthermore, this study explored the effect of training sizes on the forecasts of high-performing models as a means of determining approximately how much simulation data are necessary to provide accurate and stable predictions. We experimented with reducing training data sizes to 10,000 units and 7500 units and evaluated the resultant model's performance.

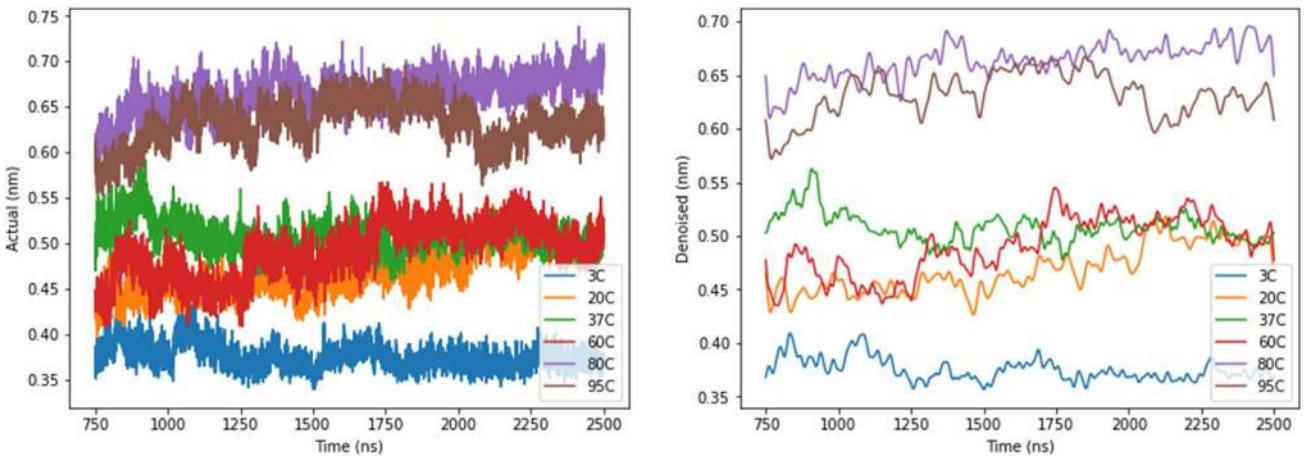

Fig. 1 Actual and denoised RMSD (nm) vs time (ns) for various temperatures

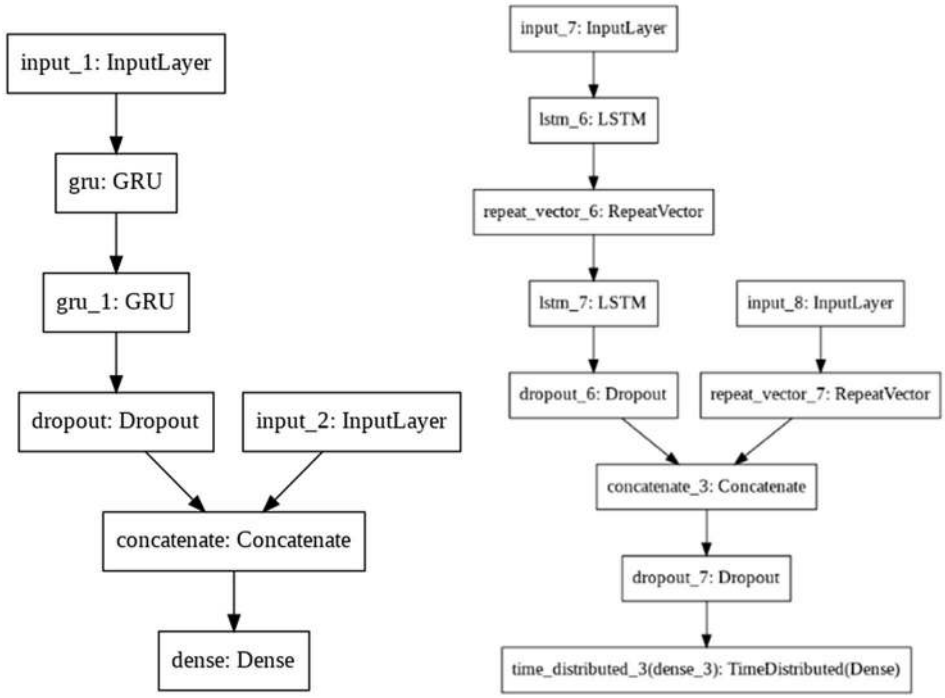

Fig. 2 Neural network model architectures. **a** Left: GRU Network. **b** Right: LSTM Autoencoder

Results

The training data consisted of 15,000 simulation data values from 750 to 2250 ns, while the forecast was evaluated with simulation data from 2250 to 2500 ns. For the GRU neural network, we used 2 stacked GRU layers, with 30 and 20 GRU units, respectively. The first GRU layer also had the return sequence parameter. A dropout layer of 0.3 was used to reduce overfitting. Window sequences of 300 values each were clustered into 3 different groups with k-means clustering using mean and standard deviation features. A k-nearest neighbors model was trained to classify the input sequences, where the class predictions were concatenated with the GRU layer outputs. The GRU model had an output size of 50 units and was trained for 6 epochs with a batch size of 1. For the LSTM Autoencoder model, the same k-nearest neighbors class predictions were used as the input data and concatenated with the LSTM layer outputs. The autoencoder has a lookback of 120 values and an output size of 30. Both LSTM layers used in the autoencoder were composed of 50 units and the output layer took the form of a time-distributed dense layer. Dropout layers of 0.3 and 0.2 are also used to reduce overfitting. This LSTM autoencoder model was trained for 8 epochs with a batch size of 1. Adam optimizers and MSE loss functions were used in both neural network models. The k-nearest neighbors model was trained on features of 10 interval means and a standard deviation value for windows of size 200. The single-step predictions each specified the change or difference in the next value and were computed with a neighbor parameter of 4 and a distance weighted function.

Experiment results suggest the kNN regression model outperformed both the GRU neural network and the LSTM autoencoder models in forecasting at temperatures of 3 °C, 20 °C, 37 °C, 60 °C, and 95 °C in terms of MSE metrics, as shown in Table 1. The kNN model

Table 1 Statistical model train (750–2250 ns) and forecast (2250–2500 ns) MSE (nm) at varying temperatures

Temperature (Celsius)	kNN		GRU		LSTM Autoencoder	
	Train	Forecast	Train	Forecast	Train	Forecast
3	2.468e−08	9.830e−03	8.482e−03	4.442e−03	3.909e−03	4.627e−02
20	1.663e−08	8.172e−03	5.220e−03	2.907e−02	2.765e−03	4.837e−02
37	4.916e−08	6.944e−03	5.835e−03	2.721e−02	1.900e−03	4.919e−02
60	1.626e−08	8.974e−03	4.845e−03	2.813e−02	2.419e−03	5.000e−02
80	1.252e−08	2.882e−02	3.801e−03	2.406e−02	1.716e−03	1.560e−01
95	2.105e−09	2.121e−02	6.321e−03	3.078e−02	1.756e−03	1.390e−01

MSE performance was around 0.02 nm less than that of the GRU model and over 0.04 nm less than that of the LSTM autoencoder. In Fig. 3, the kNN regression models were demonstrated to have captured and maintained much of the patterns in the training set, while the neural network models appear to have converged toward a single value.

Discussion

This work essentially demonstrates the feasibility of using ML to predict long-term properties of SARS-CoV-2 S-proteins from MD simulation data. Our findings show that the kNN model's performance is associated with its sensitivity in capturing patterns and structures

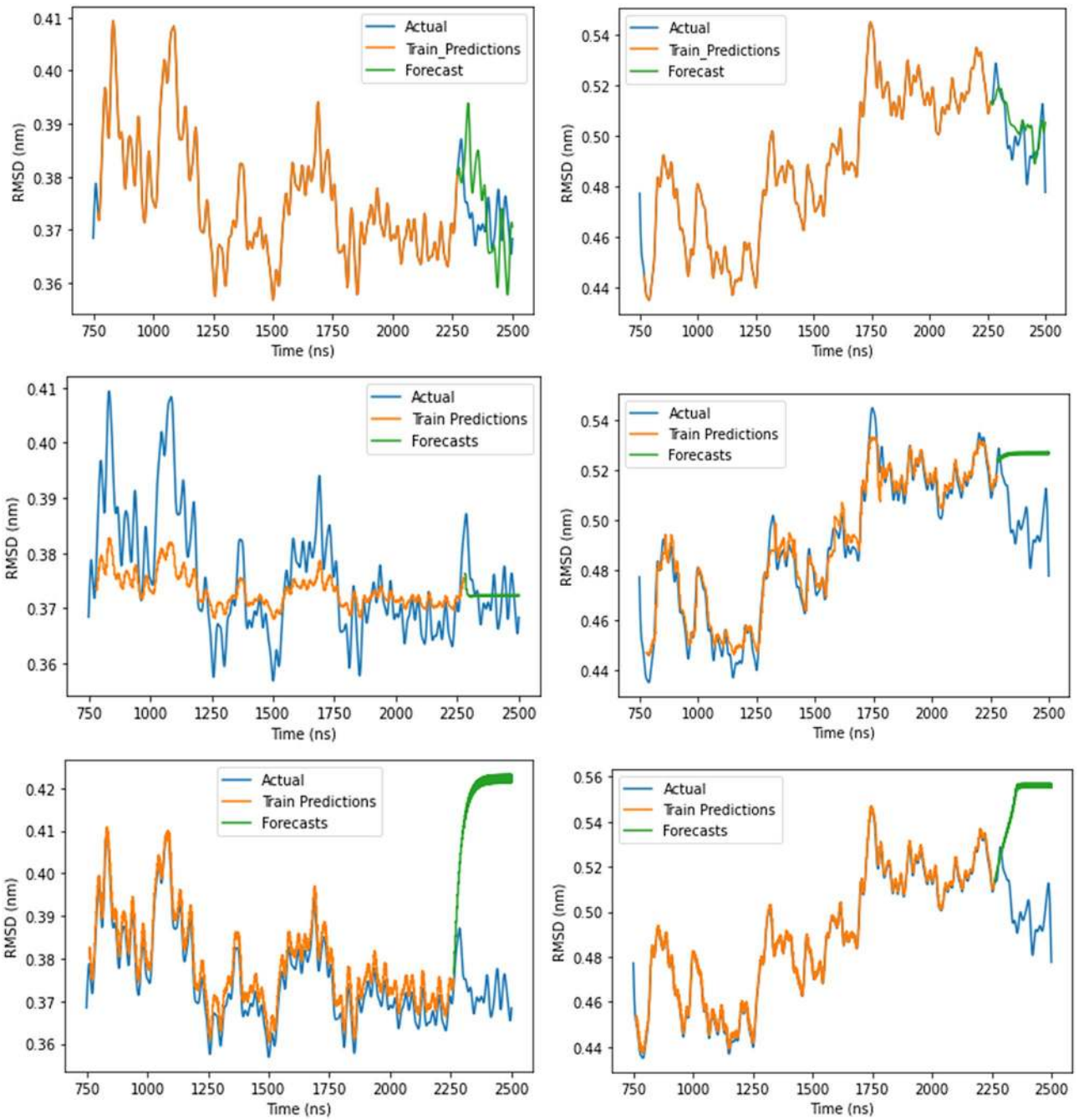

Fig. 3 Supervised model predictions for 3 °C and 60 °C. Left: 3 °C. Right: 60 °C. **a** Top: kNN. **b** Middle: GRU Network. **c** Bottom: LSTM Autoencoder

Table 2 kNN train and forecast MSE (nm) at varying training data sizes

Temperature (Celsius)	Train (15,000) 750.0–2250.0 ns	Forecast 2250.0–2500.0 ns	Train (10,000) 750.0–1750.0 ns	Forecast 1750.0–2500.0 ns	Train (7500) 750.0–1500.0 ns	Forecast 1500.0–2500.0 ns
3	2.468e−08	9.830e−03	2.568e−08	9.905e−03	4.559e−09	1.030e−02
20	1.663e−08	8.172e−03	1.281e−08	2.501e−02	2.287e−07	1.844e−02
37	4.916e−08	6.944e−03	8.280e−08	1.956e−02	8.132e−09	1.170e−02
60	1.626e−08	8.974e−03	5.154e−08	2.527e−02	1.240e−08	3.345e−02
80	1.252e−08	2.882e−02	8.904e−10	1.853e−02	5.182e−08	2.311e−02
95	2.105e−09	2.121e−02	1.551e−08	3.478e−02	1.063e−09	2.807e−02

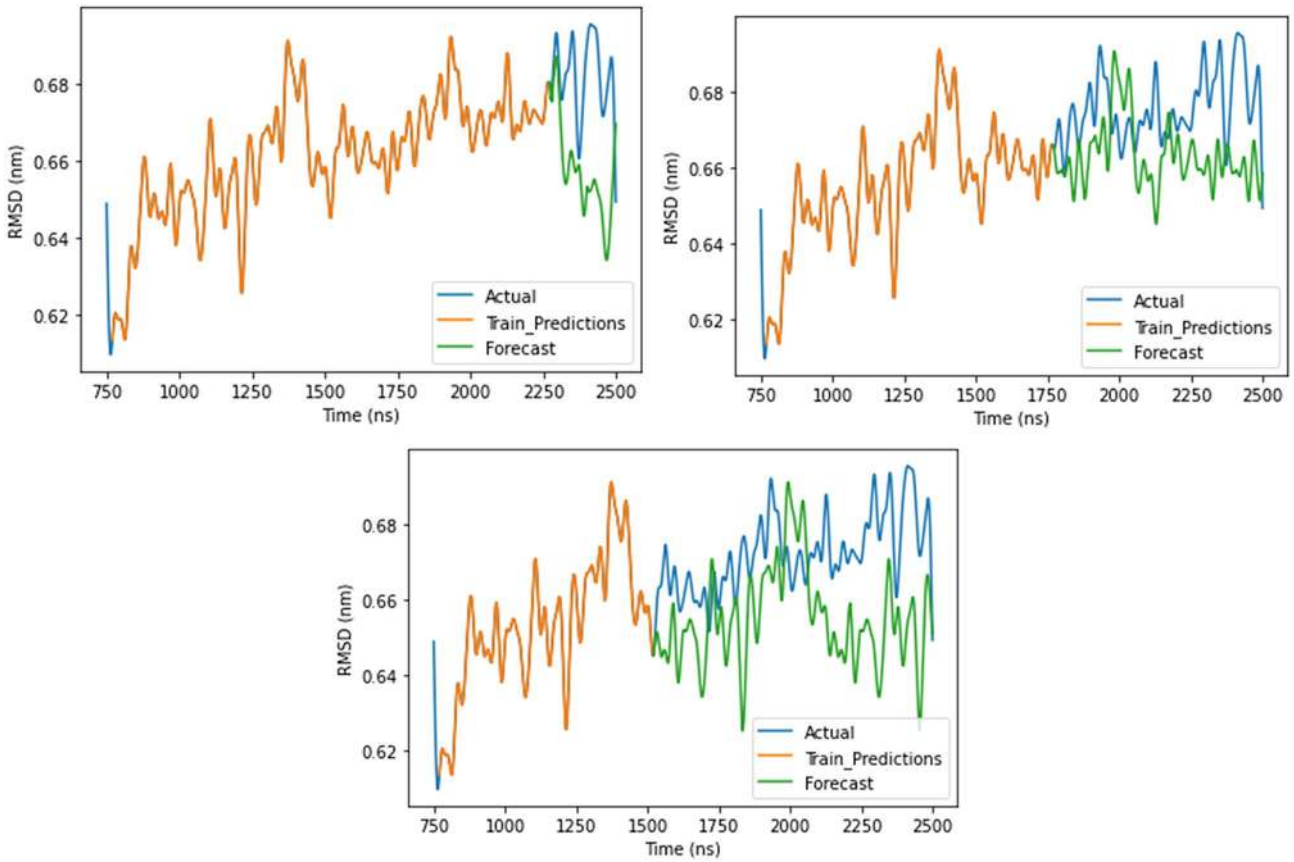

Fig. 4 kNN forecasts for 80 °C. **a** Top Left: Trained on 750.0–2250.0 ns. **b** Top right: trained on 750.0–1750.0 ns data. **c** Bottom: 750.0–1500.0 ns data

within the data. This is due to the kNN's use of instance-based learning, characterized primarily by the lack of a training period. Instead of creating abstraction from the data and learning as done with the neural networks, kNN models merely store the data and form predictions based on similarity with the new instance, or nearest neighbors [8]. The advantage of this sensitivity in this application is demonstrated with its ability to capture more patterns, rather than converging to single values as displayed with the neural network models. Furthermore, the warm-up stage (first 750 ns) of the simulation was not considered so long-term memory was not necessary in our model, as the RMSD was stabled after the warm-up. While the GRU and LSTM models perform well in prediction with long-term memory, they appear to be too complicated with their parameters for our model. The kNN model, being the best performing model, provides a fast approach toward accelerating the gathering of simulation data given relatively small amounts of training data. Experimental results in Table 2 demonstrate

that the kNN model exhibits better forecast performance when given larger data sizes. When given 15,000 training values, the model was able to achieve MSE scores over 0.01 nm less than those when trained on 10,000 values and over 0.02 less than those when trained on 7500 values. Nevertheless, the model exhibits relatively good accuracy even when given small data sizes, as shown in Fig. 4, demonstrating the effectiveness in using small training sizes to accelerate the process of gathering simulation data.

Conclusion

In this work, we investigate the performance of different supervised ML approaches toward accelerating simulation data gathering, specifically GRU neural networks, LSTM autoencoders, and kNN models. The kNN model demonstrated the greatest success and effectiveness toward forecasting simulation data, as it is shown to have had the greatest accuracy, while capturing the most patterns and trends within the data. The kNN model is demonstrated to be sensitive to the data size, with greater performance when given more data, yet even when given small amounts of training data, it still performs relatively well with MSE scores around 0.02 nm when given 7500 training values. This study provided valuable information on how to accelerate the MD simulation process through training and predicting supervised ML models. These findings are applicable to various other MD simulations, extending beyond the SARS-CoV-2 simulation conducted in this study and are shown to be useful in areas where time is of the essence. We are aware of the existence of other mutations and variants, including the Value of Concern (VOC) 202012/01 and the 501Y.V2 strain and we believe that this study is also applicable to those different strains.

Acknowledgments This work is supported by the Garcia Summer Research Program at Stony Brook University. The work is also sponsored by the OVPR&IEDM COVID-19 Seed Grant, PIs: P. Zhang, Y. Deng, M. Rafailovich, M. Simon. The simulations were conducted on the AiMOS supercomputer through an IBM Faculty Award FP0002468 (PI: Y. Deng).

References

1. Johns Hopkins Coronavirus Resource Center (n.d.). COVID-19 Map - Johns Hopkins Coronavirus Resource Center. Johns Hopkins Coronavirus Resource Center.
2. J. Tan, K.H. Verschueren, K. Anand, J. Shen, M. Yang, Y. Xu, Z. Rao, J. Bigalke, B. Heisen, J.R. Mesters, K. Chen, X. Shen, H. Jiang, R. Hilgenfeld, pH-dependent conformational flexibility of the SARS-CoV main proteinase (M(pro)) dimer: molecular dynamics simulations and multiple X-ray structure analyses. J. Mol. Biol. **354**(1), 25–40 (2005). https://doi.org/10.1016/j.jmb.2005.09.012
3. 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)
4. H.A. Aboubakr, T.A. Sharafeldin, S.M. Goyal, Stability of SARS-CoV-2 and other coronaviruses in the environment and on common touch surfaces and the influence of climatic conditions: a review. Transbound. Emerg. Dis. (2020). https://doi.org/10.1111/tbed.13707
5. S.L. Rath, K. Kumar, Investigation of the effect of temperature on the structure of SARS-CoV-2 spike protein by molecular dynamics simulations. Front. Mol. Biosci. **7**, 583523 (2020). https://doi.org/10.3389/fmolb.2020.583523
6. 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).
7. Y. Wang, S. Zhu, C. Li, C. Research on multistep time series prediction based on LSTM. IN: 2019 3rd International Conference on Electronic Information Technology and Computer Engineering (EITCE) (pp. 1155–1159). IEEE (2019).
8. E. Keogh, Instance-based learning, in *Encyclopedia of Machine Learning*. ed. by C. Sammut, G.I. Webb (Springer, Boston, 2011). https://doi.org/10.1007/978-0-387-30164-8_409