Communications in Physics, Vol. 34, No. 2 (2024), pp. 99-114
DOI: https://doi.org/10.15625/0868-3166/19607

Comparing receptor binding properties of SARS-CoV-2 and of SARS-CoV virus by using unsupervised machine learning models

Cao Phuong Cong[1,2], **Hien T. T. Lai**[1], **Ly H. Nguyen**[1,3], **Anh D. Phan**[7], **Agata Kranjc**[4,5,6], **Tien-Cuong Nguyen**[2], **Duc Nguyen-Manh**[8] **and Toan T. Nguyen**[1,2†]

[1]*Key Laboratory for Multiscale Simulation of Complex Systems, University of Science, Vietnam National University, 334 Nguyen Trai Street, Thanh Xuan, Hanoi 11400, Vietnam*

[2]*Faculty of Physics, University of Science, Vietnam National University, 334 Nguyen Trai Street, Thanh Xuan, Hanoi 11400, Vietnam*

[3]*Center for Environmental Intelligence and College of Engineering & Computer Science, VinUniversity, Gia Lam, Hanoi 12400, Vietnam*

[4]*Institute for Neuroscience and Medicine (INM-9), Forschungszentrum Jülich, Jülich, 52425, Germany*

[5] *Laboratoire de Biochimie Théorique, UPR 9080 CNRS, Université de Paris, 13 rue Pierre et Marie Curie, F-75005, Paris, France*

[6] *Institut de Biologie Physico-Chimique-Fondation Edmond de Rotschild, SL Research University, 75005, Paris, France*

[7]*Faculty of Materials Science and Engineering, Phenikaa Institute for Advanced Study, Phenikaa University, Hanoi, 12116, Vietnam*

[8]*CCFE, United Kingdom Atomic Energy Authority, OX14 3DB, Abingdon, UK*

E-mail: †toannt@hus.edu.vn

Received 12 December 2023
Accepted for publication 13 May 2024
Published 6 June 2024

Abstract. *This work continues our recent molecular dynamics investigation of the three systems of the human ACE2 receptor interacting with the viral RBDs of SARS-CoV virus and two variants of SARS-CoV-2 viruses. The simulations are extended and analyzed using unsupervised machine learning models to give complementary descriptions of hidden features of the viral binding mechanism. Specifically, the principle component analysis (PCA) and the variational autoencoder (VAE) models are employed, both are classified as dimensionality reduction approaches with different focuses. The results support the molecular dynamics results that the two variants of SARS-CoV-2 bind stronger and more stable to the human ACE2 receptor than SARS-CoV virus does. Moreover, stronger bindings affect the structure of the human receptor, making it fluctuate more, a sensitive feature which is hard to detect using standard analyses. Unexpectedly, it is found that the VAE model can learn and arrange randomly shuffled protein structures obtained from molecular dynamics in time order in the latent space representation. This result potentially has promising application in computational biomolecules. One could use this VAE model to jump forward in time during a molecular dynamics simulation, and to enhance the sampling of protein configuration space.*

Keywords: Coronaviruses, human ACE2, unsupervised machine learnings, enhanced sampling, molecular dynamics, variable autoencoder.

Classification numbers: 87.10.Tf; 87.15.ap.

1. Introduction

By the end of 2019, the Severe acute respiratory syndrome coronavirus 2 (also known as SARS-CoV-2, or 2019-nCoV) was detected in Wuhan city, China, and spread rapidly to all over many countries and territories, forcing The World Health Organization to declare a public health emergency only three months later [1]. Because of the extremely fast spread rate, fast mutation rate and the toxicity of the SARS-CoV-2, scientists are rushing to find a cure for severe acute respiratory syndrome caused by the virus. Although the pandemic has subdued significantly lately, research into this viral disease remains active as ever to thoroughly understand the viral mechanism, and be able to prevent or to prepare for future similar viral pandemic due to coronaviruses.

The structure of coronavirus can be divided into two parts, namely core and shell. The viral genome is contained in the core, while the viral shell is a combination of fat lipids, envelope proteins, and spike proteins, in which spike proteins play an important role in the entry of the RNA viral genome into the host cell. The receptor-binding domain (RBD) is a subunit of the spike glycoprotein (also known as protein S) attached to the viral outer shell [2, 3]. RBD recognizes and binds to human cells through a receptor called Angiotensin Converting Enzyme 2 (ACE2) (Figure 1) [4]. After that, the coronavirus is incorporated into the host cell to release the viral RNA into the cytoplasm.

According to several studies [5–10], the RBD of SARS-CoV (that caused the SARS epidemic in June 2003) and SARS-CoV-2 have significant similarities in genome sequence and also use the same cellular entry receptor, namely ACE2. Therefore, there is a strong need [7, 11] to properly understand the binding mechanism between the ACE2 receptor with both of these viruses in the coronavirus family to find important similarities as well as differences. Not only does this knowledge assist strongly in developing antibodies or antiviral drugs based on the binding features

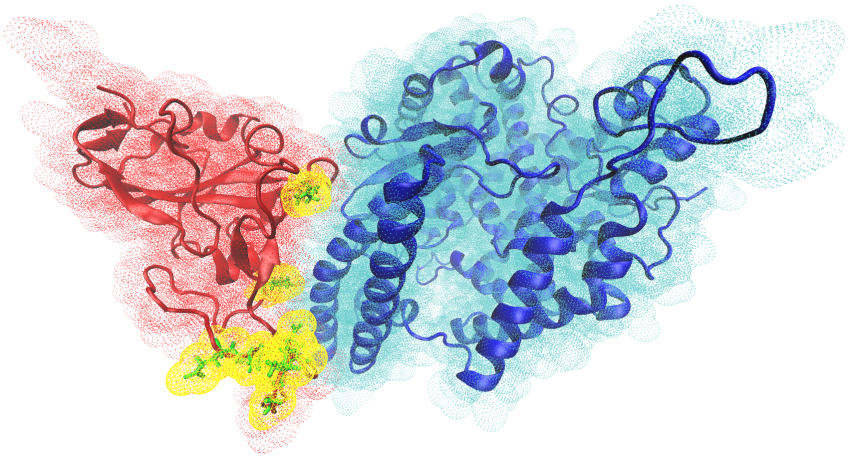

Fig. 1. The binding of coronavirus spike protein (red color) to human ACE2 receptor (blue color). The location of four significant mutations of the viral RBD is shown in yellow (see the Discussion section for more information).

of the RBD of the SARS-CoV-2 spike protein, it also provides a solid foundation for future tracking of diseases caused by coronaviruses. Noticeably, not all therapeutic antibodies or antivirals work well with different viruses of the same strain. This is because of the difference in structure caused by mutations between virus variants [7, 11]. Therefore, to evaluate the reliability of the physical picture, we study the interaction mechanism of the SARS-CoV-2 coronavirus with ACE2 in comparison with the interaction mechanism of other coronaviruses, specifically the original SARS-CoV.

In our previous study [10], a molecular dynamics simulation of the complex of the viral RBD bound to the human ACE2 receptor has been carried out, which stressed several aspects of the physical mechanism of stronger binding of the new coronavirus. In this work, we extend these simulations further, and use various machine learning approaches to study characteristics of the binding complex, in a complementary approach to conventional molecular dynamics analysis. The trajectories obtained from the molecular dynamics simulation are used as input for the principal component analysis (PCA) and for a variational autoencoder model, both are dimensionality reduction unsupervised machine learning methods, to extract hidden features of the binding dynamics. Unlike deep autoencoder [12] where each input sample is mapped into a single point, the VAE's latent space is expected to have similar features of phase space of thermodynamics and statistical physics, where the volume of a region of phase space is proportional to the time spent by the system in that region with the same energy according to the ergodic hypothesis. In the term of latent space, the area of latent space is expected to be proportional to the time spent by our system in that area, assuming that our simulated system is in equilibrium. Therefore, the VAE is expected to have more physical insights than the standard deep autoencoder, hence it is used as our choice of this class of machine learning model.

The latent space of VAE can be seen as a representation of the input data, where each dimension captures some aspects of the data variation. In some sense, it is similar to the "collective" principal coordinates of PCA. However, it is not always easy to interpret physical meanings like

the case of PCA, because the dimensions may not correspond to intuitive or meaningful concepts. The latent space may have different properties depending on the training data, and the optimization process. The reduced dimensionality of the latent space therefore represents hidden features of the system investigated. Only when combined with other physics analyses, one may be able to interpret the data more clearly. As an unexpected result, it is found that the VAE model can learn and arrange randomly shuffled protein structures obtained from molecular dynamics in time order in the latent space representation. This result potentially has promising application in computational studies of biomolecules. One could use this VAE model to jump forward in time during a molecular dynamics simulation, and to enhance the sampling of protein configuration space.

2. Materials and methods

2.1. Sequence alignment and molecular dynamics simulation

Sequence alignment is a method of arranging two or more genome sequences in order to achieve maximum similarity. It is often used to study the evolution of sequences from a common ancestor, especially biological sequences such as protein sequences or DNA, RNA sequences. Incorrect matches in the sequence correspond to mutations and gaps correspond to additions or deletions. In this work, the sequence alignment is used for the viral RBDs of both SARS-CoV virus and SARS-CoV-2 viruses to elucidate the common features as well as the viral mutations during the time of more than a decade. From the point of view of a biophysicist, understanding the physics of amino acids at mutations can provide important clues to the binding mechanism [10]. In this work, the primary sequences of the proteins were aligned by means of ClustalW webserver [13] using BLOSUM matrix [14] and then visually analyzed in order to find mutations between the SARS-CoV and SARS-CoV-2.

For atomistic simulation, the starting structure is obtained from experimental structures with PDB ID 2AJF (for SARS-CoV RBD), 6M0J and 6VW1 (for SARS-CoV-2 RBD). From these complexes, chain A (the human ACE2) and chain E (viral RBD) were prepared and corrected manually to properly describe some particular structural elements. The experimental structures of ACE2 and RBD have some missing residues in their central parts (D615 for ACE2 and A522 for RBD). These residues were added using homology modeling methods [15–17]. The molecular dynamics (MD) simulations were performed by GROMACS/2018.6 software package [18]. Proteins and ions were described by CHARMM-36 force-field [19] and glycans by GLYCAM06 force-field [20]. The TIP3P [21] model was used for water molecules. The physiological electrolyte concentration of the solution simulated is 150 mM NaCl. The simulation box size was chosen so that the proteins in neighboring periodic boxes are at least 3 nm apart from each other. Since the electrostatic screening length at 150 mM NaCl concentration is about 7 Å, this 3 nm distance is more than enough to eliminate the finite size effect due to the long range electrostatic interactions among proteins in neighboring simulation boxes, and small enough to keep the size of the system manageable with our current computational resources.

The temperature of 310 K and the pressure of 1 atm were maintained by the Nose-Hoover thermostat [22, 23] and the Parrinello-Rahman barostat [24]. The Particle Mesh Ewald (PME) method [25] is used to treat the long-range electrostatic interaction with a real space cutoff of 1.2 nm. The van der Waals interactions were also cut off at 1.2 nm, with the appropriate cut-off corrections added to pressure and energy. To speed up computations, all hydrogen bonds were

constrained by the LINCS method [26]. The systems were equilibrated in NPT ensemble for 1 ns, then simulated for 2 μs of production run.

2.2. Principal Component Analysis

Principal component analysis (PCA, also called covariance analysis) is a very common and powerful tool not only in machine learning but also in general data analysis. PCA is an unsupervised learning technique for pre−processing and reducing the dimensionality of high-dimensional datasets while maintaining the original structure and connections. In our case of systems of RBD−ACE2 complex, PCA is a powerful tool for analyzing protein dynamics because of the big data of a large number of atoms of proteins over a long time of simulation.

During the simulation, we consider a subgroup of N atoms. Denote q_1, ... , q_{3N} as the coordinates of the atoms, the covariance matrix of the $\sigma_{3N\times 3N}$ of $3N$ atoms has elements

$$\sigma_{ij} = \langle (q_i - \langle q_i \rangle)(q_j - \langle q_j \rangle) \rangle, \tag{1}$$

where i and j are the indexes of coordinates, $\langle \cdot \rangle$ denotes the time-average operator.

We obtain $3N$ eigenvectors $\mathbf{v}^{(k)}$ and $3N$ eigenvalues λ_k by diagonalizing σ with

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{3N}. \tag{2}$$

The modes of collective motion and their amplitudes are specified by the eigenvectors and eigenvalues of σ. The larger the value of λ_k is, the more significant that mode of motion contributes to the overall motion of the system. The principal components k^{th} has the form

$$V_k = \mathbf{v}^{(k)} \cdot \mathbf{q} = v_1^{(k)} q_1 + v_2^{(k)} q_2 + \cdots + v_{3N}^{(k)} q_{3N}. \tag{3}$$

Over a long time of simulation, not every fluctuation and deviation of atoms in the protein are equally important. The dynamics of the protein is dominated by only a few component motions, or a few $\mathbf{v}^{(k)}$ with largest eigenvalues.

In our system, the number of atoms of the RBD-ACE2 complexes is rather large (around 12500 atoms). If all atoms are used for PCA, the covariance matrix will have a size of about 37500×37500, making the calculation not only computationally expensive but also somewhat redundant. We select only two groups of atoms from the RBD−ACE2 complexes, namely the backbone of the viral RBD protein and the backbone of the ACE2 receptor for analyses.

2.3. Variational autoencoders

Variational autoencoders (VAE) is an advanced technique of machine learning in general and deep learning in particular. Just like PCA, VAE is also an unsupervised learning technique for dimensionality reduction of high-dimensional datasets. It belongs to the family of autoencoder methods. VAE is the combination of deep autoencoder (DAE) and variational Bayesian methods [27]. It essentially consists of two main parts: encoder and decoder (Fig. 2). The encoder is the first half of the VAE neural network. The encoder aims to condense the input information of protein structure by passing it through a funnel-like fully connected neural network. The latent space generated from the encoder is just the representation of the condensed input information. The decoder is the last half of the VAE neural network. In contrast to the encoder, the decoder aims to use the encoder output and reconstruct the input data. In the form of the loss function, the reconstructed data will then backpropagate from the VAE's neural network. The key difference of VAE from other variants of autoencoder is that it maps each input sample into an area with

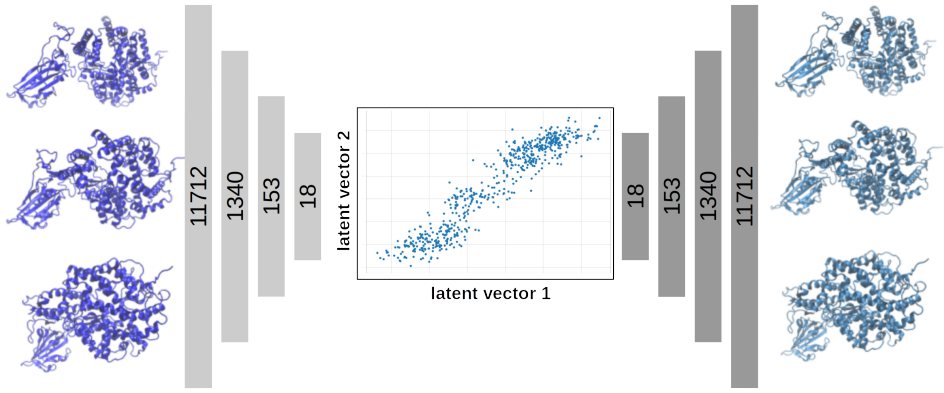

Fig. 2. (Color online) Illustration of VAE structure used in this work for systems. The left outer most layer are the input layer with size equal to the degrees of freedom of the system. The right outer most layer is the output layer with the same size as the input layer. In between, there are four hidden layers for the encoder (left) and four for the decoder (right). The model is trained to minimize the difference in the output (generated) versus the input configuration.

a Gaussian distribution in the latent space, instead of a single point. VAE provides a statistical way to describe the dataset's samples in latent space. This key difference of VAE is also the reason why it is chosen to investigate our systems instead of other variants of autoencoder such as the deep autoencoder [12]. The VAE's latent space is expected to have similar features of phase space of thermodynamics and statistical physics, where the volume of a region of phase space is proportional to the time spent by the system in that region with the same energy according to the ergodic hypothesis. In the term of latent space, the area of latent space is expected to be proportional to the time spent by our system in that area, assuming that our simulated system is in equilibrium. In the case of DAE, each input sample is mapped into a single point in the latent space. Therefore, theoretically, there are no constraints for two relative mapped points of two extremely different protein structures. Hence, there may be less physical meaning in the case of using DAE.

Just like PCA, only the backbone atoms of the RBD-ACE2 complexes are chosen for the input data for feeding to the VAE model. Besides, positions of C_β are selected additionally for supplying the information of residue's directions for VAE. The total number of atoms is 3906. Conventionally, the raveled distance matrix of atom coordinate is used as the input data. However, in this case, the raveled distance matrix is an array with a length of more than 15×10^6 (3906×3906), a huge number for building and optimizing VAE model. Therefore, in this case, the distance matrix is improved to keep the distances of atoms having less than three neighbor atoms in between. Accordingly, the number of distances drops to 11712, which is reasonable for our machine learning purpose. The system structures are extracted from the 2 μs trajectory every 1 ns.

The number of the input layer is equal to the number of distances between selected atoms, that is 11712. The numbers of nodes of each layer in the VAE's encoder are chosen to decrease

gradually, i.e. 1340, 153, 18, and 2. Similarly, layers of the VAE's decoder have 2, 18, 153, 1340, and 11712 nodes respectively.

In this work, the code of VAE is built and developed in Python 3.8. The data preprocessing procedure is carried out using the MDAnalysis library [28] for easy reading, writing, and analyzing trajectories from MD simulations in GROMACS formats. For machine learning, the Keras package [29] with Tensorflow library [30] is used. The total trainable parameters for our model is about 32 million parameters. The source code of VAE model can be provided to interested readers upon a reasonable request.

3. Results and discussions

3.1. Sequence alignment and molecular dynamics simulation

The viral genome sequence is not as stable and conservative as the human ACE2 receptor. The viral genome sequences change over time because of a number of mutating events. The viral receptor-binding domain sequence alignments are described in Fig. 3. The sequence identities between 2AJF with 6VW1 and 6M0J are 83.3% and 71.1% respectively. The sequence identity between 2 variants 6VW1 and 6M0J is 85.7%. These identity scores imply that the RBD of all investigated viruses has similar structures and sequences. This result can be explained by the fact that the investigated viruses all belong to the coronavirus family. Also from the sequence alignments, it is observable that there are many mutations and they are evenly distributed throughout the viral sequences. We choose to focus on mutations between SARS-CoV and SARS-CoV-2 that are not-conservative and appear in both SARS-CoV-2 variants. There are four significant mutation

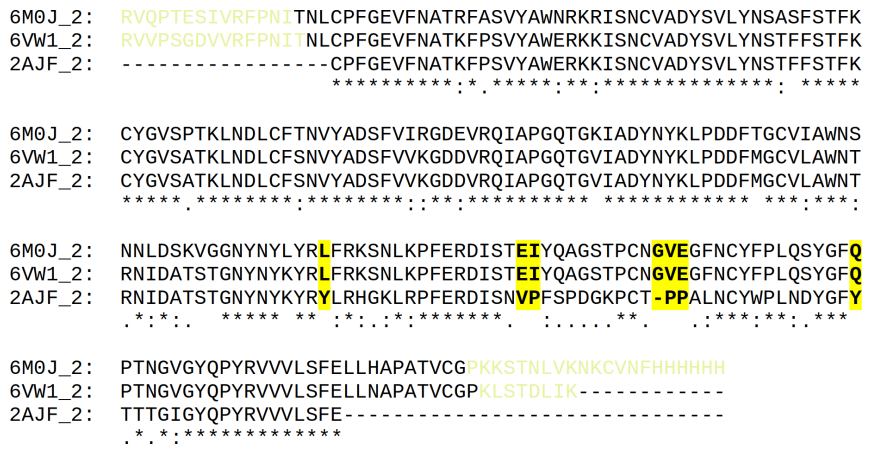

Fig. 3. (Color online) The sequence alignments of the viral RBD of 6VW1 and 6M0J for two variants of SARS-CoV-2 virus, and of 2AJF for SARS-CoV virus. Below the viral RBD protein sequences is a key denoting conserved sequence (*), conservative mutations (:), semi−conservative mutations (.), and non−conservative mutations ( ). Some missing residues (in the crystal structures) at the terminals of sequences are faded out.

positions that are close to the binding interface with the receptor ACE2. These four mutation positions are colored yellow as already shown in Fig. 1. They may hold an important explanation for the stable and strong binding of the SARS-CoV-2 viruses, as our previous work already suggested

by using molecular dynamics analyses [10]. Here we only demonstrate a few properties needed for discussion of our results with unsupervised machine learning models, namely the standard deviation and fluctuations of the backbone atoms. The results are shown in Figs. 4 and 5.

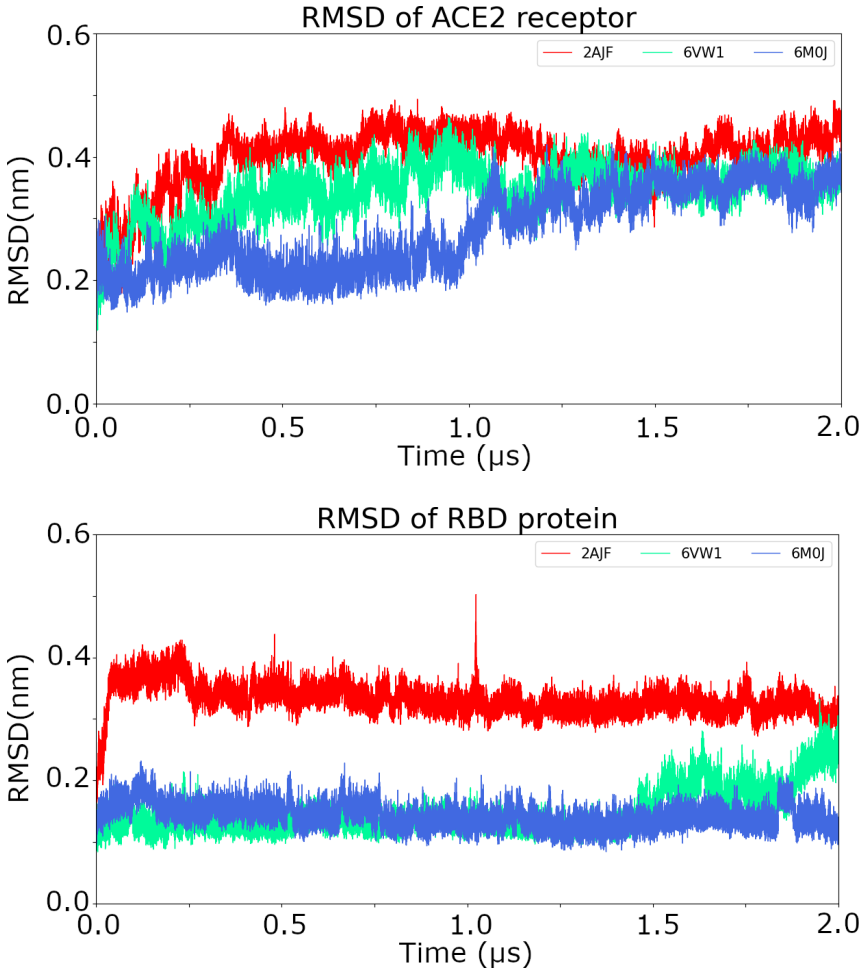

Fig. 4. (Color online) The root mean square deviations of the backbone of the human ACE2 receptor and of the viral RBD protein.

In general, Fig. 4 revealed that the RMSDs of the viral RBD is less variable than the RMSDs of the human ACE2 receptor over 2 μs of simulation. The saturation time of the viral RBD is about 50 ns in comparison with the much higher saturation time of the ACE2 receptor which is nearly 400 ns. Thus, the first 500 ns of simulation was dropped for the subsequent equilibrium analyses. Not only the saturation time but also the magnitudes of the deviation of the receptors tend to slightly increase and be higher than that of the viral binding domains. These observations can be explained by the fact that the ACE2 receptor has almost 600 residues whereas the viral RBD is smaller and has only about 200 residues. Apparently, the size of the receptor makes it less variable

than the RBD even though the RMSD is almost always less than 4 Å and all systems are supposed to be stable.

Among the viruses, the RMSDs of the RBD of the two new viruses are around 1.5 Å, dramatically lower than that of SARS-CoV virus which is approximately 3 Å. In the same trend as RBD, the ACE2 in the new virus systems show lower RMSD values, especially in the 6M0J system, than the ACE2 in the old virus system. Therefore, the 6M0J and 6VW1 systems are evidently more stable than the 2AJF system.

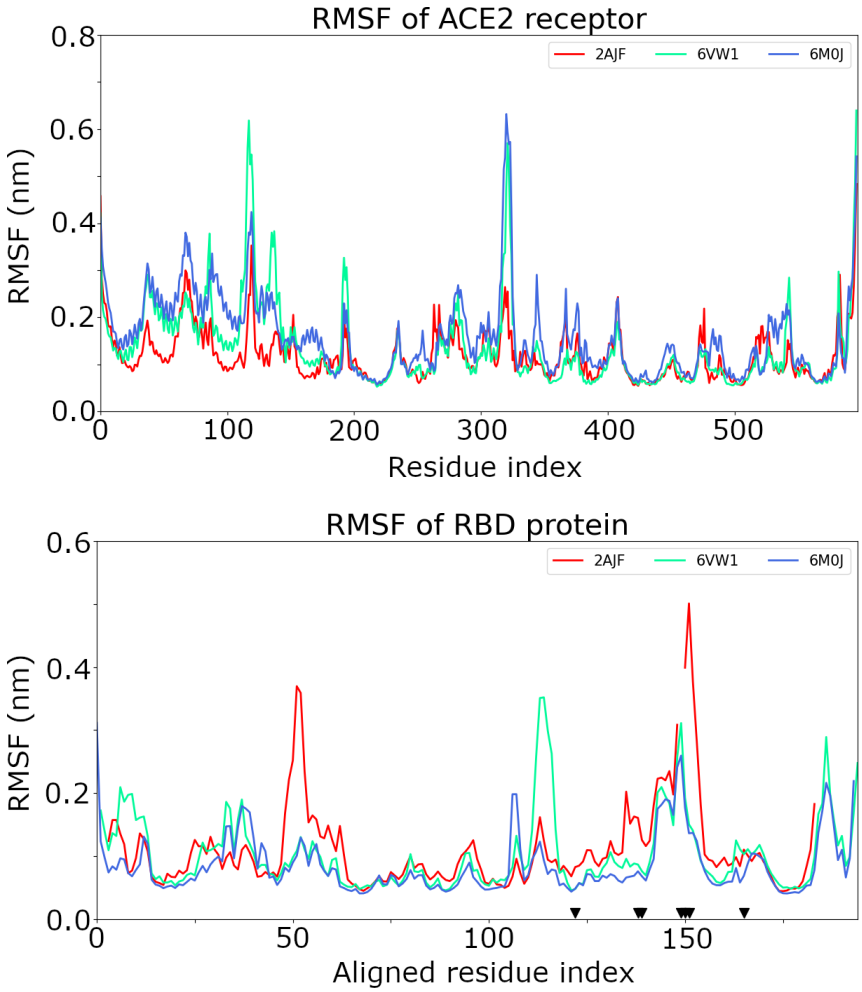

Fig. 5. (Color online) The root mean square fluctuations of the backbone of the human ACE2 receptor and of the viral RBD protein.

RMSF is often used to measure the stability of the protein backbone. For the human ACE2 receptor, the overall RMSF trends of all three systems are similar with each other (Fig. 5). This is reasonable because all three ACE2 structures from simulated systems have identical sequences and structures. Some small changes can be explained by the two following reasons. The first

change is that residues at the N- and C-termini have few constraints causing a higher magnitude of fluctuation. The second change, more importantly, is that the structures of proteins used for the simulations are from the experiments where the proteins are frozen and crystallized. Therefore, when simulated at 310 K, the protein structures slightly change and have thermal fluctuations. These fluctuation differences are insignificant because most residues have RMSF differences less than 1 Å.

For the viral RBD protein, the residue indexes are aligned according to the sequence alignments shown in Fig. 3. The RMSF of the viral RBD of the two SARS-CoV-2 variants are very close to each other implying that the two viruses behave very similarly especially from residues larger than 120. The binding interface can be determined from residue 120 to residue 170. Therefore, there is almost no difference in the behavior of the two SARS-CoV-2 viruses.

The second observation is that all four significant mutations of SARS-CoV-2 viruses make the viruses more stable with fewer thermal fluctuations at their positions. The four mutations occur at residues 123, 139-140, 150-152, and 166 (denoted as small black downward triangles in Fig. 5)). The RMSF of RBD of SARS-CoV-2 viruses at these residues is smaller than that of SARS-CoV virus. Especially at residue 150-152, the RMSF of RBD of SARS-CoV virus is almost twice as much as that of the new viruses. This significant change is supposedly caused by the insertion of Glycine amino acid combined with the substitution of two Proline amino acids in the SARS-CoV sequence making the backbone more flexible to move closer to the ACE2 receptor and to attach to it more tightly.

The third observation is that there are still some abnormal peaks of RMSF such as at around residue 113 of RBD of 6VW1 system or at around residue 50 of RBD of 2AJF system. However, these residues are relatively far from the binding interface region, thus having no remarkable effect on those four significant mutations. Discounting these, the interface mutations comprise most significant differences between the RMSFs of the viral proteins. The most distinguished mutation is the substitution of -PP for GVE. Because Glycine (G) is a small, light, and soft residue, the insertion and substitution make the sequence softer (mass), longer (geometry), and easier to reach the receptor. This mutation clearly causes the most significant difference in Fig. 5 of viral protein around residue index 150, with both of the SARS-CoV-2 viral proteins being much more stable.

3.2. Principal Component Analysis

To analyze the dynamics of proteins throughout a long simulation, PCA is performed on all three systems. The first 500 ns of the simulations is dropped to ensure that the system is well equilibrated. The trace of the covariance matrix is an effective measurement of the overall backbone flexibility of the human ACE2 receptor and the viral RBD protein. The trace of the covariance matrix of the protein backbones of our three systems, namely 6VW1, 6M0J and 2AJF, are computed and displayed in Table 1. From this table, the traces of ACE2 receptors increase steadily from the 2AJF system to the 6M0J system. This result is reasonable because it is in line with our RMSD and RMSF calculations shown above. Fig. 4 showed that the ACE2 displacement difference between the first and the last system configuration of the 6M0J system is the largest and that of the 2AJF system is the smallest. This infers that the ACE2 receptor is more flexible in the 6M0J system than in the 2AJF system. Opposite to the trace of the ACE2 receptor, the traces of the viral RBD decrease from the 2AJF system to the 6M0J system. By a similar argument, this

result is expected. This result also showed that the viral RBD of the two new viruses is more stable than the SARS-CoV virus RBD.

Table 1. The trace of the co-variance matrix of the projections of the protein backbones on the two largest principal components.

		2AJF	6VW1	6M0J
Trace	ACE2 receptor	10.118	15.500	19.197
(nm^2)	viral RBD	3.012	2.729	1.820

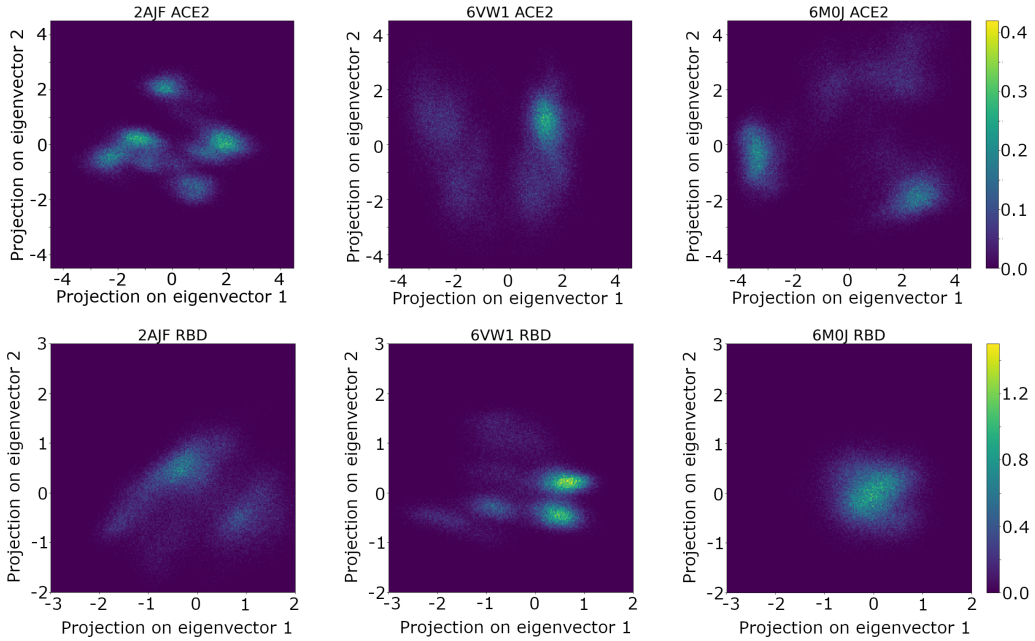

Fig. 6. (Color online) The probability density in the plane of the two largest principal components from the PCA of the backbones structure of proteins. Note that the colorbar scale (right most figures) is the same for all RBD or ACE2, but they are different between RBD and ACE2. Similarly, for the range of axes, the same range is used for all RBD or ACE2, but they are different between RBD and ACE2.

Figure 6 shows the probability density of the two largest principal components from the PCA of the backbones structure of proteins. The projections on the third-highest principal show some simple normal distribution, thus are ignored and not shown here. From Fig. 6, the brightness of an area indicates the likelihood that the system configuration localizes in this area. The sharper and brighter the region is, the more preferable the system stays at that region, and vice versa.

The first set of three subfigures (the first row) of Fig. 6 displays the sharpest and brightest area of the ACE2 receptor in the 2AJF system. Despite a large number of spots, the ACE2 receptor is comparatively stable. The two remaining subfigures of 6VW1 and 6M0J systems show the dim and blurred spots implying that even considering the two most outstanding motion modes, the

systems still tend to vibrate freely. The second set of three subfigures (the second row) of Fig. 6 tells a contrary story. The probability density of the RBD protein backbone of the 2AJF system is very dim and unclear. On the other hand, the probability density shown in the two remaining subfigures of 6VW1 and 6M0J is brilliant and very sharp. It means that the RBD protein in 6VW1 and 6M0J systems is very localized and stable throughout the simulation. These results are in good agreement with the observations that the RBD of SARS-CoV-2 viruses are more stable than that of SAR-CoV virus from previous works [10,31–33]. Our PCA analyses indicate an additional effect, that their ACE2 receptor vibrates harder. A reasonable explanation for this is that the RBD protein binds stronger and more stable to the ACE2 in 6M0J and 6VW1 systems making them vibrate together.

Overall, our PCA unsupervised learning analyses have enhanced molecular dynamics analyses. Not only it confirms the stronger and more stable bindings of viral RBD to human ACE2 in the case of SARS-CoV-2 viruses, but also demonstrates the slightly destabilizing effect that stronger binding has on the human ACE2 for which our previous molecular dynamics could not sensitively detect.

3.3. Variable autoencoder model

Let us now move to VAE analyses, another unsupervised learning model in the same approach of dimensionality reduction. Fig. 7 shows the latent space projection of variational autoencoder trained on the distance matrix of the RBD-ACE2 complex of the 6M0J system. In our first thought, we expect the latent space to represent the classified system configurations in clusters with reference to the potential energy. However, the results shown in Fig. 7 tell a totally different story. From running VAE, one sees that the latent space of VAE is able to represent protein structures linearly according to their simulation time instead of their potential energy (Figures 7b and 7d). In MD simulation of proteins of biophysical systems, the simulation time 1 ns is not too big but still long enough for proteins to perform some significant changes in their structure. During the VAE training, the input protein structures are 1 ns-simulation-apart from each other and shuffled. However, VAE somehow can learn and organize the data representation in the latent space linearly in time instead of energy (Fig. 7b). This result is unexpected, and can have very promising applications in computational biomedicine, such as predicting the next system configurations based on some simulated configurations, speeding up simulation time, or potentially enhancing sampling of the configuration space. Nevertheless, this result is not completed yet and still needs further investigations and extension to other systems.

The second observation is that, there are two clusters of the representative data in the latent space as the histogram of latent space projections clearly shows. Fig. 7c shows two distinct regions of the representative data (colored deep dark blue). On the other hand, these two regions exactly match two different simulation stages according to Fig. 7b. Combining the results, we can draw a conclusion that during the simulation, the system is moving from a local minimum state to another local minimum state. This moving is extremely difficult to observe in previous MD analyses. Moreover, it is native that there are still thermal fluctuations in a local minimum causing the different values of potential energy of the system (shown in Fig. 7d). The upper region is bigger and covers about two-third color spectrum or two-third of simulation time equally. Therefore, this local minimum state movement happens recently and there also needs further simulations and investigations.

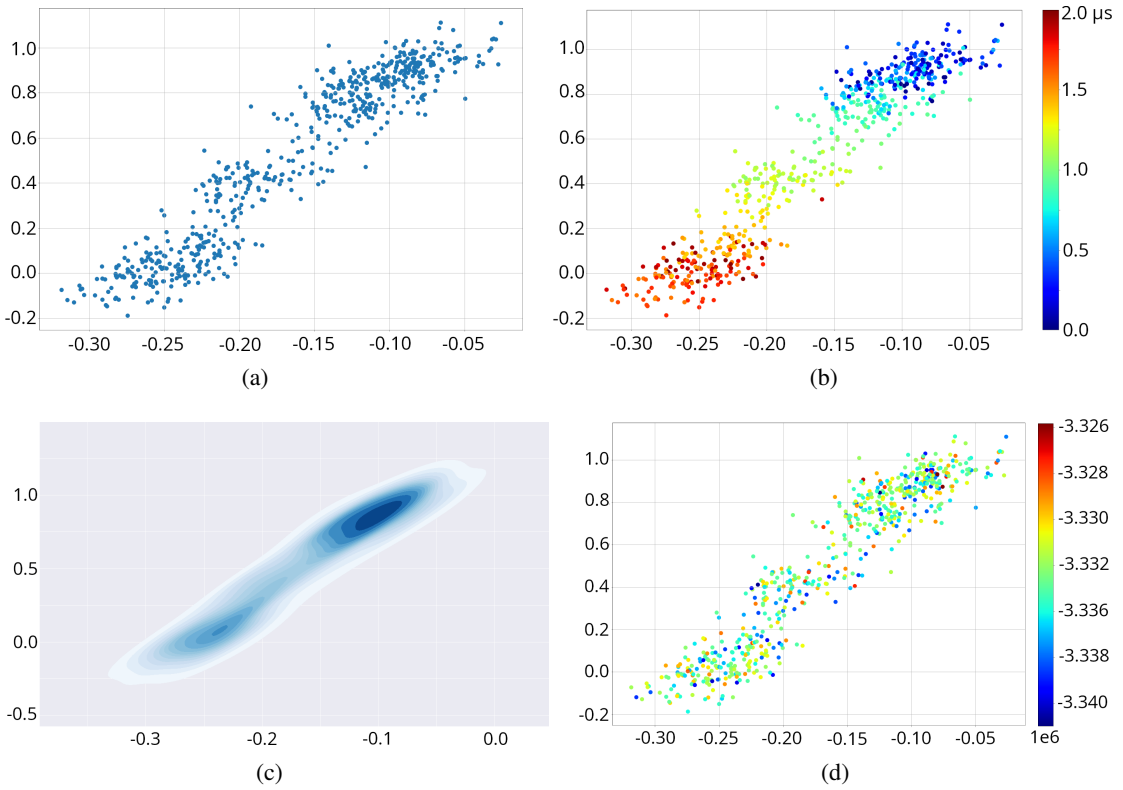

Fig. 7. (Color online) Latent space projection of variational autoencoder trained on the distance matrix of RBD-ACE2 complex of 6M0J. The training procedure was done after 60 epochs. Fig. (a) shows the projection without labeling data. Fig. (b) and Fig. (d) show the projection with data labeled to time frame and potential energy respectively. Fig. (c) shows the histogram of the projection points from Fig. (a).

Finally, Fig. 8 shows interesting dependence of the VAE results when it is trained using different numbers of epochs. As the number of epochs increases, the distribution of representative points in the latent space tends to form a line. In other words, the first latent vector seems to linearly depend on the second latent vector at epoch 100. Therefore, the representation in the latent space would become less informative and less useful. This is a well-known knowledge in training machine learning models: the model becomes overfitted at a large number of epochs because the model starts memorizing the data instead of learning it. We plan for more investigations in the future to improve our model, optimizing its parameters and hyperparameters and extend it to a broader spectrum of systems.

4. Conclusions

In this work, three systems of the human ACE2 receptor interacting with the viral RBDs of SARS-CoV virus and two variants of SARS-CoV-2 viruses are modeled and simulated and analysed using unsupervised machine learning models. Both standard and machine learning analyses

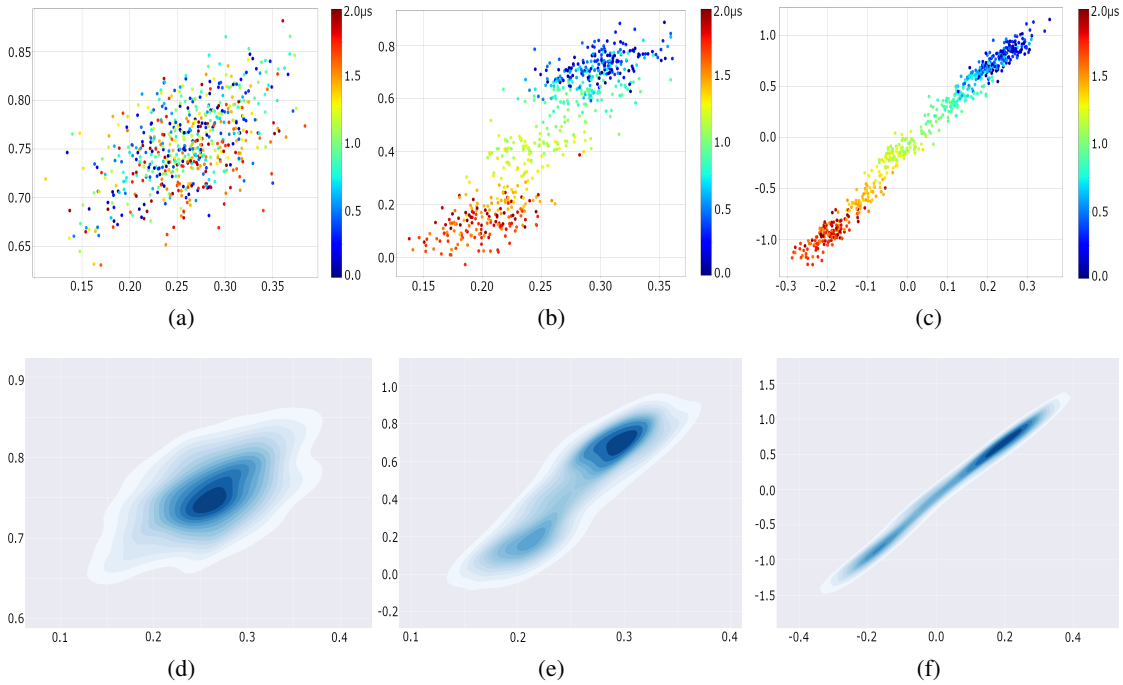

Fig. 8. (Color online) Latent space projection of variational autoencoder trained on the distance matrix of RBD-ACE2 complex of 6M0J. The training procedure was done after 10, 50 and 100 epochs. In the first row, all the projections are shown with data labeled to time frame. In the second row, the histrogame of distribution in the latent space are shown.

agree with each other and support the picture that the two variants of SARS-CoV-2 show stronger binding and form more stable complex with the human ACE2 receptor than SARS-CoV virus does. Moreover, the stronger bindings can affect the structure of the human receptor, making it fluctuate more, hence slightly becoming less stable. This is a sensitive feature which is hard to detect using standard analyses.

Even though, protein structures obtained from molecular dynamics simulation are randomly shuffled before feeding to the VAE model, the VAE interestingly can learn and arrange them in time order in representation in the latent space. This result is unexpected, but the application of this result is very promising. One could use VAE for jumping forward in time during a MD simulation, as well as for enhanced sampling of configuration space. Nevertheless, our results are reportedly preliminary, more rigorous investigation to optimize parameters and hyperparameters of the model are needed in the future.

Acknowledgements

The authors acknowledge the financial support of the Vietnam National University - Hanoi grant number QG.20.82. We thank Dr. Paolo Carloni for many critical comments on the manuscript, and Dr. Pham Tien Lam for valuable assistance with the Variational Autoencoder model.

References

[1] C. Wang, P. Horby, F. Hayden and G. Gao, *A novel coronavirus outbreak of global health concern*, The Lancet **395** (2020) 470.

[2] S. Belouzard, J. Millet, B. Licitra and G. Whittaker, *Mechanisms of coronavirus cell entry mediated by the viral spike protein*, Viruses **4** (2012) 1011.

[3] S. Sieczkarski and G. Whittaker, *Dissecting virus entry via endocytosis*, The Journal of general virology **83** (2002) 1535.

[4] W. Li, M. Moore, N. Vasilieva, J. Sui, S. Wong, M. Berne *et al.*, *Angiotensin-converting enzyme 2 is a functional receptor for the sars coronavirus*, Nature **426** (2003) 450.

[5] P. Zhou, X. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang *et al.*, *A pneumonia outbreak associated with a new coronavirus of probable bat origin*, Nature **579** (2020) 270.

[6] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan *et al.*, *Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor*, Nature **581** (2020) 1.

[7] Y. Wan, J. Shang, R. Graham, R. Baric and F. Li, *Receptor recognition by novel coronavirus from wuhan: An analysis based on decade-long structural studies of sars*, Journal of Virology **94** (2020) e00127.

[8] W. Tai, L. He, X. Zhang, J. Pu, D. Voronin, S. Jiang *et al.*, *Characterization of the receptor-binding domain (rbd) of 2019 novel coronavirus: implication for development of rbd protein as a viral attachment inhibitor and vaccine*, Cellular & Molecular Immunology **17** (2020) 1.

[9] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen *et al.*, *Sars-cov-2 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor*, Cell **181** (2020) 271.

[10] H. T. Lai, L. H. Nguyen, A. D. Phan, A. Kranjc, T. T. Nguyen and D. Nguyen-Manh, *A comparative study of receptor interactions between sars-cov and sars-cov-2 from molecular modeling*, Journal of molecular modeling **28** (2022) 305.

[11] K. Andersen, A. Rambaut, W. Lipkin, E. Holmes and R. Garry, *The proximal origin of sars-cov-2*, Nature Medicine **26** (2020) 1.

[12] M. T. Degiacomi, *Coupling molecular dynamics and deep learning to mine protein conformational space*, Structure **27** (2019) 1034.

[13] J. D. Thompson, T. J. Gibson and D. G. Higgins, *Multiple sequence alignment using clustalw and clustalx*, Current protocols in bioinformatics (2003) 2.

[14] D. W. Mount, *Using blosum in sequence alignments*, Cold Spring Harbor Protocols **2008** (2008) pdb.top39.

[15] A. Šali and T. L. Blundell, *Comparative protein modelling by satisfaction of spatial restraints*, Journal of Molecular Biology **234** (1993) 779.

[16] E. Krieger, S. B. Nabuurs and G. Vriend, *Homology modeling*, Methods of biochemical analysis **44** (2003) 509.

[17] H. T. Lai, D. M. Nguyen, T. T. Nguyen *et al.*, *Homology modeling of mouse nlrp3 nacht protein domain and molecular dynamics simulation of its atp binding properties*, International Journal of Modern Physics C (IJMPC) **31** (2020) 1.

[18] H. J. Berendsen, D. van der Spoel and R. van Drunen, *Gromacs: a message-passing parallel molecular dynamics implementation*, Computer physics communications **91** (1995) 43.

[19] J. Huang and A. D. MacKerell Jr, *Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data*, Journal of computational chemistry **34** (2013) 2135.

[20] K. N. Kirschner, A. B. Yongye, S. M. Tschampel, J. González-Outeiriño, C. R. Daniels, B. L. Foley *et al.*, *Glycam06: a generalizable biomolecular force field. carbohydrates*, Journal of computational chemistry **29** (2008) 622.

[21] Y. Sun and P. A. Kollman, *Hydrophobic solvation of methane and nonbond parameters of the tip3p water model*, Journal of computational chemistry **16** (1995) 1164.

[22] W. G. Hoover, *Canonical dynamics: Equilibrium phase-space distributions*, Phys. Rev. A **31** (1985) 1695.

[23] S. Nosé, *A molecular dynamics method for simulations in the canonical ensemble*, Mol. Phys. **52** (1984) 255.

[24] M. Parrinello and A. Rahman, *Polymorphic transitions in single crystals: A new molecular dynamics method*, J. Appl. Phys. **52** (1981) 7182.

[25] T. Darden, D. York and L. Pedersen, *Particle mesh ewald: An n·log (n) method for ewald sums in large systems*, The Journal of chemical physics **98** (1993) 10089.

[26] B. Hess, H. Bekker, H. J. Berendsen and J. G. Fraaije, *Lincs: a linear constraint solver for molecular simulations*, J. Comput. Chem. **18** (1997) 1463.
[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes." 2013.
[28] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf and O. Beckstein, *Mdanalysis: a toolkit for the analysis of molecular dynamics simulations*, Journal of computational chemistry **32** (2011) 2319.
[29] F. Chollet *et al.*, *Keras: Deep learning library for theano and tensorflow*, URL: https://keras. io/k **7** (2015) T1.
[30] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015.
[31] J. Shang, G. Ye, K. Shi, Y. Wan, C. Luo, H. Aihara *et al.*, *Structural basis of receptor recognition by sars-cov-2*, Nature **581** (2020) 221.
[32] A. C. Walls, Y.-J. Park, M. A. Tortorici, A. Wall, A. T. McGuire and D. Veesler, *Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein*, Cell **181** (2020) 281.
[33] D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona *et al.*, *Cryo-em structure of the 2019-ncov spike in the prefusion conformation*, Science **367** (2020) 1260.