Allosteric Regulation of SARS-CoV-2 Spike Protein Revealed by Contrastive Machine Learning

Yong Wei,[1] Amy X. Chen,[2] Yuewei Lin,[3] Tao Wei,[4] Baofu Qiao[5]

[1] Department of Computer Science, High Point University, High Point, NC, 27268
[2] Thomas Jefferson High School for Science and Technology, Alexandria, VA, 22312
[3] Computational Science Initiative, Brookhaven National Laboratory, Upton, NY 11973
[4] Department of Chemical Engineering, Howard University, Washington, DC, 20059
[5] Department of Natural Sciences, Baruch College, City University of New York, New York, NY, 10010

Abstract: Allosteric regulation is common in protein-protein interactions and thus can be used in drug design. Nevertheless, the mechanism of allosteric regulation remains elusive for most proteins, including SARS-CoV-2 spike protein, despite extensive experimental endeavors over the past years. In the present computational study, the route of allosteric regulation of SARS-CoV-2 is examined by all-atom explicit solvent molecular dynamics simulations in conjunction with contrastive machine learning. It was found that peptide binding to the polybasic cleavage sites, especially the one at the first monomer of the trimeric SARS-CoV-2, activates the fluctuation of the spike protein's backbone. This fluctuation eventually propagates to a nitrogen-terminal domain and its neighboring receptor-binding domain, remarkably weakening the latter's binding affinity to the human cell receptor ACE2. Our study justifies the presence of allosteric regulation in SARS-CoV-2, paving the way for the systematic design of allosteric antibody inhibitors.

Introduction:

Allosteric regulation refers to the mechanism that an event (e.g., ligand binding) at one place of a protein leads to an influence on a remote location of the protein, such as the local mobility and interactions with the other molecule.[1, 2, 3, 4, 5, 6] In addition to the design of drugs which directly bind the active sites of proteins, allosteric regulation provides a new route for drug design.[7, 8, 9] Nevertheless, our current understanding of allosteric regulation is remarkably limited and its molecular mechanism remains unrevealed at this moment due to the protein's complicated folded structures. [10] It thus limited the progress of allosteric regulation-based drug design.

Coronavirus disease 2019 (COVID-19), due to infection of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has caused a global pandemic for three years, leading to over 6.6 million deaths and 0.65 billion confirmed cases worldwide according to the report of the World Health Organization (WHO) (https://covid19.who.int/) by the end of December of 2022. SARS-CoV-2 virus attacks human cells via the binding of its spike protein with the angiotensin-converting enzyme 2 (ACE2) receptors, which are highly expressed on the surface of type II cells.[11, 12, 13] The coronavirus spike protein, typically known as the spike-protein, is a trimeric glycoprotein, which plays a key role in binding receptors,[14] appears on the virus surface as outward-facing 23 nm molecular "spikes".[15] A spike protein is composed of three monomers, each composed of around 1270 amino acids.[16] Therefore, each trimeric spike protein has around 3800 amino acids, where there exist a huge amount of protein-protein interactions, standing for one of the most complicated examples in folded proteins.

The allosteric regulation of SARS-CoV-2 spike protein has been previously observed experimentally[17, 18, 19] and in computer simulations [20]. Specifically, Chi, et al., [17] found that antibody 4A8, which was isolated from recovered patients, binds the nitrogen-terminal domains (NTDs) of the spike proteins. These NTDs are around 4-8 nm away from the binding interface between the spike receptor-binding domain (RBD) and human cell receptor ACE2. Another antibody CR3022,[18] also isolated from a recovered patient, was found to target a highly conserved epitope of SARS-CoV-2 (and SARS-CoV), which is distal from the RBD. Similarly, antibody 47D11 [19] binds to non-RBD sites of SARS-CoV-2 (and SARS-

CoV). These experimental observations support the presence of allosteric regulations in the spike protein of SARS-CoV-2 (and likely SARS-CoV). Meanwhile, using all-atom explicit solvent molecular dynamics simulations, Qiao and Olvera de la Cruz [20] examined the polybasic cleavage sites (PCSs, $R_{682}RAR_{685}$) on the spike protein, which are unique to SARS-CoV-2 compared to the other lineage B coronavirus.[21, 22, 23] Even though the PCSs are around 10 nm away from the RBD-ACE2 binding interface, their presence was shown to be able to enhance the binding affinity between the spike RBD and ACE2 and a neutralizing peptide targeting the PCSs also decreased the RBD-ACE2 bind affinity.[20] Therefore, these existing experimental and simulation works have supported the presence of the allosteric effect in the SARS2 spike protein. Nevertheless, the mechanism remains unexplored. The pathway of signal transmission from the peptide-binding site to the remote RBD-ACE2 binding site and the structures in transition are unknown.

Machine learning (ML) has been proven to be a very powerful means in understanding protein structures. Simulation data of protein molecule structure evolution is usually unlabeled. Therefore, these data are interpreted in an unsupervised manner. Generative models such as autoencoders have been used to obtain features from protein contact maps, aiming to provide data to clustering [24, 25] to find stages of structure evolution trajectory. Self-supervised learning is another category of ML models that can produce good feature representations from unlabeled data for downstream tasks [26]. Backbone models that generate the feature representations of data are trained by solving "pretext" tasks, such as predicting rotations [27], relative patch locations [28], learning inpainting [29], solving jigsaw puzzles [30], and image coloring [31]. However, these hand-crafted pretext tasks depend empirically on ad-hoc heuristics, which limits the generality of the representation of data. In this work, contrastive learning [31] is adopted to extract feature representations of contact maps of the SARS-CoV-2 spike protein obtained by the all-atom explicit solvent MD simulations. The extracted contact map features are then grouped by the kmeans clustering algorithm to reveal the stages of spike protein structure evolution when it binds with human ACE2 receptor and further reveals the mechanism of allosteric regulations.

In the present work, we examined the pathway for allosteric regulation in the SARS-CoV-2 spike protein. All-atom explicit solvent molecular dynamics (MD) simulations were carried out along with the contrastive machine learning (ML) approach. Both methods have proven successful in examining allosteric effects in protein-protein interactions.[10, 32] The large structure of the SARS-CoV-2 provides a perfect example for ML calculations. The hybrid ML-atomistic simulation approach reveals the route in the allosteric regulation in the SARS-CoV-2 spike protein, which will be greatly beneficial for our understanding of the mechanism of allosteric regulations.

Results:

Stages of SARS-CoV-2 spike protein structure evolution trajectory obtained via contrastive learning and clustering

Fig. 1 shows the results of machine learning analysis of the entire trajectory of SARS-CoV-2 spike protein molecular structure evolution in the process of protein-ACE2 binding. Feature vectors of spike protein contact maps are extracted by the backbone feature extractor, which is a deep resnet50 model in this work. The contact map feature vectors are grouped using the kmeans algorithm. To find the optimal number of clusters *k*, cluster numbers ranging from 1 to 25 were tried, the elbow method and the average silhouette scores were utilized (**Fig. 1a and 1b)**. It was found that the optimal number of clusters $k = 6$, indicated by both the elbow method and the silhouette scores except $k = 2\ or\ 3$, which cannot present protein structural evolution with sufficient details and were not adopted. As shown in **Fig. 1c**, transitions between folding stages occurred at 9.2 ns, 12.0 ns, 28.3 ns, 50.1 ns, 65.4 ns and 92.2 ns, respectively, suggesting significant changes in protein structure.

Route of allosteric regulation in SARS-CoV-2 spike protein

To examine the pathway of the allosteric regulations in the spike protein, we first carried out all-atom explicit solvent molecular dynamics (MD) simulation on the spike-ACE2 complex. In this simulation, each of the three positively charged polybasic cleavage sites was associated with one negatively charged tetrapeptide EELE (Glu-Glu-Leu-Glu), which was found to be able to destabilize RBD-ACE2 binding.[20] We hypothesize that the strong electrostatic attractions between the positively charged PCS domains and the negatively charged EELE tetrapeptides afford a local structural fluctuation that might eventually trigger a global conformational adaptation of the entire spike protein, which is the objective of this work. The initial structure of the spike-ACE2-EELE complex was the same as that in our previous work,[20] whereas different random seeds were employed in conducting the equilibration and productions simulations. The production simulation was carried out for a duration of 100 ns. By saving one structure for every 0.1 ns simulation time, we saved 1000 frames. These frames were then analyzed using the contrastive machine learning method, which identified a total of 6 stages with their center indexes (frame number) as 48, 164, 330, 463, 639, and 832.

Using the 48th frame as the reference structure, we calculated the other five stage-centric frames' root-mean-square deviations (RMSD). RMSD stands for the least-square fit between the target structure and the reference structure (frame 48) after structural alignments. We are interested in the following motifs of PCS-A, PCS-B, PCS-C, NTD-A, NTD-B, NTD-C, and RBD-C, where PCS stands for polybasic cleavage sites, NTD for nitrogen terminal domain, and RBD for receptor-binding domain, and -A, -B, -C are indicating the three monomers of the trimeric spike protein. The calculated RMSDs are presented in **Table 1**, alongside the ones for the entire spike protein. In each column of **Table 1**, we highlighted the notable variances of RMSD in red and blue for the increase and decrease in RMSD upon consecutive inter-stage transitions. For instance, in the column corresponding to the 48→832 transition, the RMSD of PCS-A drops substantially from 1.15 nm to 0.63 nm when compared to its counterpart in the precedent 48→639 transition, making 0.63 nm highlighted in blue. By contrast, the RMSD of NTD-B increases significantly from 1.53 nm to 2.29 nm, which is thus highlighted in red. As shown in **Fig. 2**a, the EELE tetrapeptide fell off the PCS-B motif upon the 48→164 transition, drastically raising the RMSD of PCS-B to 1.37 nm. Similarly, during the successive 48→330 transition, another EELE tetrapeptide dissociated from PCS-C (**Fig. 2b**), leading to a large structural fluctuation of PCS-C as exhibited by an increase in its RMSD from 0.56 nm to 0.91 nm. Nevertheless, the greatest increase in RMSD from 0.48 nm to 1.35 nm was found on PCS-A, which can be ascribed to the structural modulation by the preserved EELE/PCS-A adduct (**Fig. 2b**). Another large amplification of RMSD from 0.25 nm to 1.09 nm was observed on NTD-A. Given the large distance between NTD-A and the activated PCS-A of around 3.3 nm, we rule out the short-range through-space coupling between them. If one considers the fact that both NTD-A and PCS-A motifs belong to the same monomer chain and thus are somehow connected via chemical bonds, the long-range through-bond coupling is more likely the driving force for their pronounced structural correlation. On the other hand, the short separation between NTD-B and NTD-C by 2.1 nm furnishes their through-space coupling, leading to their simultaneously elevated RMSD values from ~0.8 nm to ~1.2 nm. In the 48→463 transition, the only appreciable RMSD change is the one for PCS-C that restores from 0.91 nm to 0.50 nm. This restoration can be explained by the PCS-C's random thermal fluctuations since its originally bound EELE tetrapeptide had drifted further away and showed no sign of reversible binding (**Fig. 2c**). Interestingly, this Brownian motion of PCS-C also significantly promotes its RMSD to 1.33 nm in the 48→639 transition wherein remarkable changes of RMSD were observed on NTD-C and RBD-C, too. Since PCS-C, NTD-C, and RBD-C are far apart by at least 4.0 nm (Fig. 2d), the through-bond coupling is seemingly the predominant cause for their collective fluctuations. In the final 48→832 transition, the entire protein's RMSD reached its plateau value of ~0.8 nm, suggesting the conclusion of its global response to the EELE binding after ~80 ns. As a result, the considerable changes in RMSD on PCS-A and NTD-B correspond to local stochastic oscillations. Surprisingly, the EELE/PCS-A adduct is the only one that survived nearly the whole simulation of 100 ns long, whereas EELE/PCS-B and EELE/PCS-C broke up after ~15 ns and ~30 ns, respectively. More interestingly, among the three PCS motifs, PCS-A is the farthest from RBD-C (**Fig. 1**), which directly

binds to the ACE2 receptor. Therefore, it would be valuable to explore the response of the spike protein when only its PCS-A motif is patched.

Inspired by these observations, we then carried out another all-atom explicit solvent MD simulation, where only one tetrapeptide EELE was binding PCS-A. Equilibration simulations, and a 100 ns production simulation were conducted (see Methods for details). When only one EELE tetrapeptide was bound to the PCS-A motif, substantial structural changes were observed on the spike protein over 1,000 frames extracted from the 100 ns MD trajectory. Using our contrastive machine learning, a total of 10 stages were clustered for the 1,000 frames, featuring stage center indexes (frame numbers) of 26, 91, 222, 351, 467, 567, 644, 729, 860, and 954, respectively. Therefore, the 26th frame was chosen as the reference structure for the other 9 stage-centric frames to perform structural alignments before their RMSDs were evaluated and presented in **Table 2**. In the first transition (*i.e.*, 26→91), PCS-A has the greatest RMSD of 0.75 nm among all seven motifs of our interest. This large structural fluctuation of PCS-A is well expected because it was patched by an EELE tetrapeptide (**Fig. 3a**), which is hypothesized to initiate a global response of the spike protein that eventually weakens its binding to the ACE2 receptor. After ~13 ns, the second transition (*i.e.*, 26→221) took place to activate RBD-C, which exhibited an RMSD of 1.00 nm, whereas PCS-A was deactivated with a substantially reduced RMSD of 0.24 nm. Therefore, long-range signal transduction appears feasible as the PCS-A and RBD-C motifs are over ~10 nm apart from each other (**Fig. 3b**) despite nearly unvaried structures of the other 5 motifs of interest (**Table 3**). In the third transition (*i.e.*, 26→351), NTD-B begins to resonate with its neighboring RBD-C, featuring collective activations as shown in **Fig. 3c**. This resonance can be ascribed to the short-range coupling between NTD-B and RBD-C as their shortest distance is only 2.63 nm. As a result, both of their RMSD values significantly increased by ~0.40 nm. In contrast, this short-range coupling cannot explain the simultaneous RMSD increase of NTD-B and PCS-B in the fourth transition (*i.e.*, 26→467) because these two motifs are well separated by ~5 nm. Although the structural change of NTD-B could be partially modulated by the neighboring RBD-C, the strong correlation between NTD-B and PCS-B is more likely expressed through the through-bond coupling. In the fifth transition (*i.e.*, 26→567), PCS-A is activated again by the EELE tetrapeptide, raising its RMSD value from 0.45 nm to 0.91 nm. In a similar fashion to the 26→91 transition when PCS-A was activated for the first time, the RMSD of NTD-B is reduced drastically from 1.34 nm to 0.73 nm. Nevertheless, when PCS-A is deactivated again in the sixth transition (26→567) by dropping its RMSD to 0.58 nm, the RMSD of NTD-B restores to 1.17 nm probably due to the through-bond coupling from PCS-B, which maintains a high RMSD value of 1.13 nm. Moreover, the RMSD of RBD-C surges to 1.70 nm, presumably because of the through-space coupling from the already activated NTD-B if one considers their shortest distance of 1.64 nm. In the seventh transition (*i.e.*, 26→567), the three key motifs, namely, PCS-A, NTD-B, and RBD-C, all exhibit remarkably increasing RMSDs by at least 30%, affording a rather large RMSD of 0.82 for the entire protein. In particular, the RMSD of RBD-C reaches its plateau value of 2.15 nm, considerably disrupting its binding with the ACE2 receptor, as also evidenced by our binding affinity simulation. Similar to our previous work,[20] the RBD-C/ACE2 binding affinity was characterized using the short-range coulomb and van der Waals interactions between RBD-C and ACE2 in the presence of tetrapeptide EELE. The ACE2/RBD-C binding affinity was found to drop sharply from 740 kJ/mol in the absence of any EELE tetrapeptide to 440 kJ/mol where PCS-A is patched by an EELE tetrapeptide. During the eighth (*i.e.*, 26→860) and ninth (*i.e.*, 26→954) transitions, the only notable change is the RMSD variance of NTD-B owing to random thermal fluctuations that do not seem to be driven by other motifs as their RMSDs are nearly invariant. Therefore, the global response of the spike protein effectively converged upon the seventh transition, which occurred ~70 ns after the EELE tetrapeptide bound to the PCS-A motif.

Discussion:

We demonstrate the route of allosteric effects in the spike protein of SARS-CoV-2 (**Fig. 5**) The EELE tetrapeptides prefer the binding to the polybasic cleavage site on the first monomer of the trimeric spike protein. The fluctuation of the spike protein was activated upon the binding of the EELE tetrapeptide.

This fluctuation propagates from the binding site (PCS-A) to the polybasic basic cleavage site B (PCS-B), as indicated by the green arrow in **Fig. 5**. The fluctuation consequently propagates to the nitrogen-terminal domains NTER-B and NTER-C, as well as RBD on the monomer C of the trimer spike protein (orange arrows in **Fig. 5**). NTER-B and NTER-C are adjacent to RBD-C, which is in direct contact with the human cell receptor ACE2.

Also demonstrated in **Fig. 5, NTDs (at least NTD-B and NTD-C)** correlate with RBD-C. In other words, NTDs and RBD-C share the same route of allosteric regulation for the SARS-CoV-2 spike protein. This is believed to explain the efficacy of antibody 4A8.[17] Antibody 4A8, which was discovered from recovered patients, binds NTDs of SARS-CoV-2 spike protein, the mechanism of which has never been examined before. The current simulation work and the experimental evidence of the efficacy of antibody 4A8 suggest the feasibility of the design of allosteric neutralizing antibodies targeting NTD.

In summary, by coupling contrastive machine learning and all-atom explicit solvent MD simulations, we have revealed the route of allosteric regulation in the spike protein of SARS-CoV-2. Impressively, the nitrogen-terminal domains are found to share the same route of allosteric regulations as the receptor-binding domain that in direct contact with the human cell receptor ACE2. It thus also supports the feasibility in designing allosteric drugs targeting the nitrogen-terminal domains of SARS-CoV-2 spike protein. This work thus sheds insights into the fundamental understanding of allosteric regulations in protein-protein interactions as well as into the rational design of allosteric drugs.

Methods:

All-atom explicit solvent MD simulations. We carried out two all-atom simulations on SARS-CoV-2 spike-ACE2 complex, the first simulation with three tetrapeptides EELE, each binding to one of the three PCSs,[20] and the second with only one tetrapeptide EELE binding to the first PCS (PCS-A). These simulations were performed using the package GROMACS (version 2019.6) [33] at the Texas Advanced Computing Center. Like our previous work,[20] the CHARMM 36m potential [34] was used, along with the recommended CHARMM TIP3P water model [35] with the water structures constrained using the SETTLE algorithm. [36]

The SARS-CoV-2 spike protein-ACE2 binding structure was downloaded from the Zhang-Server.[14] The spike protein-ACE2 complex was reconstructed using the C-I-TASSER model [37] based on the protein identification number QHD43416 [38] for the spike protein. The SARS-CoV-2/ACE2 complex with 68,608 atoms was then solvated in a water box with a size of 16 nm×18 nm×24 nm. A salt concentration of 0.15 M was applied. In the first simulation, three tetrapeptides EELE (Glu-Glu-Leu-Glu) were placed near the three PCS motifs of spike protein in hopes that a structural change of the protein could be activated by the electrostatic binding between the negatively charged tetrapeptides and the positively charged PCS motifs. This simulation discovered a stable binding between the first PCS and the adjacent tetrapeptide. Consequently, in the second simulation, only one tetrapeptide EELE was placed next to the PCS-A motif.

The energy minimization of the whole system was first conducted using the steepest descent algorithm to remove possible close contact between different molecules. Subsequent equilibrations were conducted for one simulation of 1 ps using the canonical ensemble (constant number of particles, volume, and temperature) and another simulation of 1 ps using the isothermal-isobaric ensemble (constant number of particles, pressure, and temperature, NPT). The velocity-rescale temperature coupling and the Berendsen pressure coupling were applied.

Afterward, the solvated system was equilibrated for another 10 ns under the NPT ensemble with the Nosé-Hoover at 298K and the Parrinello-Rahman barostat at 1.0 atm.[39] In all the equilibration simulations above, the integration time step of 2 fs was using with all the hydrogen-involved covalent bonds constrained using the LINCS algorithm, [40, 41] and the coordinates of the non-hydrogen atoms of both the spike protein trimer,

ACE2, and the tetrapeptides were restrained using a force constant of 1000 kJ/mol/nm^2 to preserve the binding structure. These restraints were then removed in the production simulations. The other parameters were the same as those in the production simulation. Each productions simulation was carried out for a duration of 100 ns using the NPT ensemble. A total of 1,000 snapshots were extracted for every 0.1 ns.

The contact map between all the α-carbon atoms of the spike protein from each extracted snapshot was calculated using *gmx mdmat*, a utility program of GROMACS. The evolution trajectory of the spike protein is represented by a sequence of contact maps. A contact map C is a two-dimensional matrix whose element, $C(i,j)$, is the spatial Euclidean distance between the alpha carbon atoms of the $i\ th$ and $j\ th$ amino acids of the spike protein at a particular moment.

Contact Map Feature Extraction using Contrastive Learning

As shown in **Fig. 6**, the contrastive learning algorithm learns the feature representations of contact maps by maximizing the agreement between a positive pair $(\tilde{x}_i, \tilde{x}_j)$ via a loss function, in which $\tilde{x}_i$ and $\tilde{x}_j$ are correlated views of the same contact map x, generated by stochastic data augmentations τ~T and τ'~T, respectively. The loss function between a positive pair is defined in Eq. (1).

$$l_{i,j} = -log \frac{\exp\left(\frac{sim(y_i, y_j)}{\tau}\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\frac{sim(y_i, y_k)}{\tau}\right)} \qquad (1)$$

in which $\mathbb{1}_{[k \neq i]} \in \{0,1\}$ is an indicator function, $sim(u,v) = \frac{u^T v}{\|u\|\|v\|}$ is cosine similarity of (u,v). τ is a temperature parameter, which is empirically determined. $f(\cdot)$ is the backbone representation encoder. Resnet50 is used for this purpose. $g(\cdot)$ is a small projection head, which in this project is a multilayer perceptron (MLP) with one hidden layer. Both $f(\cdot)$ and $g(\cdot)$ are trained to maximize the agreement between the positive pairs of augmented views of contact maps using the loss function. The dimension of the extracted contact map representation is 128×1 in our work. The following augmentations are sequentially and randomly (with a probability of 0.5) applied: random cropping followed by resizing back to the original size, Sobel filtering, random horizontal flipping, and Gaussian blurring. After the contrastive learning model is trained, the projection head is thrown away. The output of the backbone representation encoder is the feature representation of the corresponding contact map. The feature representation vectors of contact maps obtained by the all-atom explicit solvent molecular dynamics simulations are then grouped via the k-means clustering algorithm to reveal the evolution stages of SARS-CoV-2 spike protein structures in the process of binding to the human cell receptor ACE2.

Data availability:

All relevant data are available from authors upon request.

Code Availability:

All relevant codes are available from authors upon request.

Acknowledgments:

Y. W. and A. C. are grateful for the computational resources offered by the National Science Foundation (#1548562 to Y.W.) through its Extreme Science and Engineering Discovery Environment (XSEDE) at the Texas Advanced Computing Center (TACC) and Pittsburgh Supercomputing Center (PSC). T. W. and B.Q. thank the support from the National Science Foundation Award **(#:** 2118099 to T. W. and # 2152853 to B.Q.).

Author contributions:

Y. W. developed machine learning software, carried out machine learning simulations, analyzed data and machine learning results, and wrote the manuscript. A. C. performed atomistic simulations, analyzed data, and wrote the manuscript. Y. L. designed the machine learning approach and analyzed machine learning results. T.W. and B.Q. designed the project, contributed to the atomistic simulations and data analysis and wrote the manuscript.

Competing Interests:

The authors declare no competing interests.

Tables:

Table 1. RMSD (nm) of the spike protein and its 7 active motifs of interest between different frames in the presence of three EELE tetrapeptides. [a]

Transition	*48→164*	*48→330*	*48→463*	*48→639*	*48→832*
PCS-A	0.48	**1.35**	1.36	1.15	**0.63**
PCS-B	**1.37**	1.31	1.27	1.23	1.50
PCS-C	0.56	**0.91**	**0.50**	**1.33**	1.18
NTD-A	0.25	**1.09**	0.95	1.21	1.25
NTD-B	0.87	**1.33**	1.38	1.53	**2.29**
NTD-C	0.54	0.72	0.50	**1.06**	1.08
RBD-C	0.72	**1.19**	1.23	**0.79**	0.83
Entire	0.45	0.56	0.74	0.84	0.80

[a.] The frames were identified by contrastive machine learning. RMSDs with significant variations upon consecutive transitions are highlighted in red and blue for increasing and decreasing values, respectively.

Table 2. RMSD (nm) of the spike protein and its 7 active motifs of interest between different frames in the presence of only one tetrapeptide EELE. [a]

Transition	*26→91*	*26→222*	*26→351*	*26→467*	*26→567*
PCS-A	**0.75**	**0.24**	0.49	0.45	**0.91**
PCS-B	0.54	0.42	0.74	**1.15**	1.15
PCS-C	0.63	0.51	0.46	0.28	0.21
NTD-A	0.40	0.49	0.60	0.61	0.63
NTD-B	0.50	0.41	**0.79**	**1.34**	**0.73**
NTD-C	0.62	0.77	0.90	1.11	1.05
RBD-C	0.67	**1.00**	**1.46**	**0.99**	1.20
Entire	0.39	0.47	0.52	0.64	0.68

Transition	*26→644*	*26→729*	*26→860*	*26→954*
PCS-A	**0.58**	**1.07**	1.21	1.04
PCS-B	1.13	1.14	1.07	1.11
PCS-C	0.35	0.45	0.31	0.37
NTD-A	0.65	0.80	0.84	0.73
NTD-B	**1.17**	**1.52**	**1.10**	**1.40**
NTD-C	1.27	1.11	1.16	1.03
RBD-C	**1.70**	**2.15**	2.10	2.13
Entire	0.74	0.82	0.85	0.87

[a] The frames were identified by contrastive machine learning. RMSDs with significant variations upon consecutive transitions are highlighted in red and blue for increasing and decreasing values, respectively.

Figures:

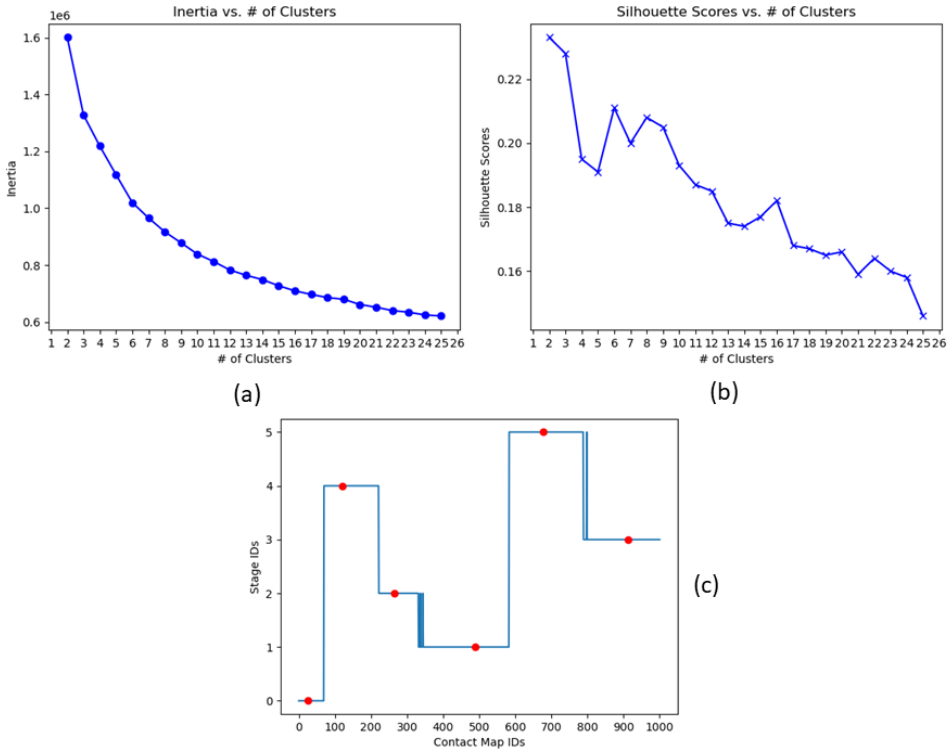

Figure 1. SARS-CoV-2 spike protein structure evolution trajectory analysis using contrastive learning and kmeans clustering. (a) Elbow method using inertia (b) The average silhouette score with different numbers of clusters. Both criteria indicate that k=6 is the optimal number of clusters. (c) Six clustered stages of spike protein molecular structure evolution in the process of the protein-ACE2 binding, in chronological order. The red dots are the positions of contact map IDs that are closest to the centroid of each cluster, respectively. These contact map IDs are 26, 120, 266, 490, 677, and 913.

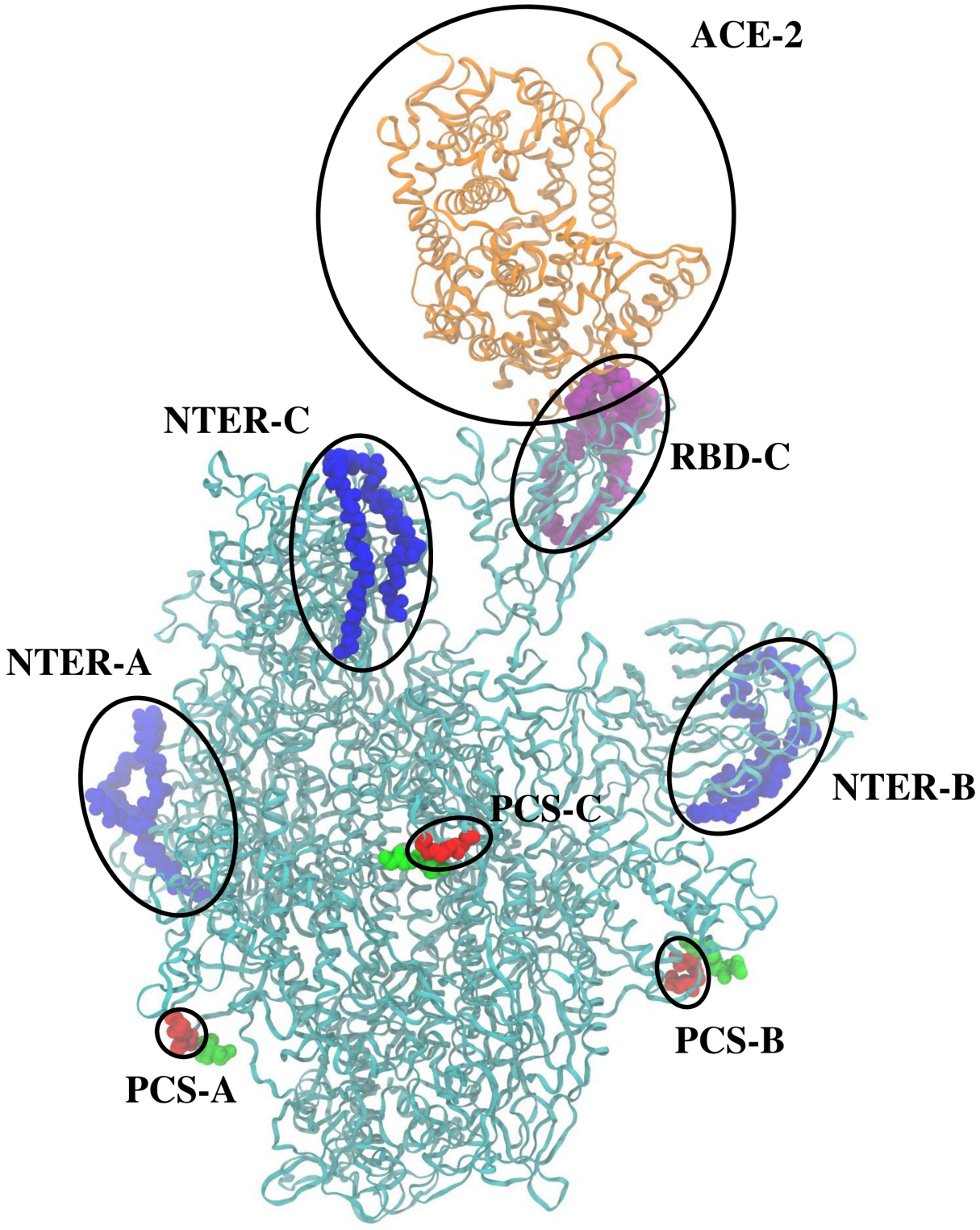

Figure 2. Molecular structure of the spike-ACE2 complex at an MD snapshot. The spike protein's PSC, RBD, and N-terminal moieties are colored in red, purple, and blue, respectively. Moreover, the ACE2 receptor and the three tetrapeptides are colored orange and red, respectively.

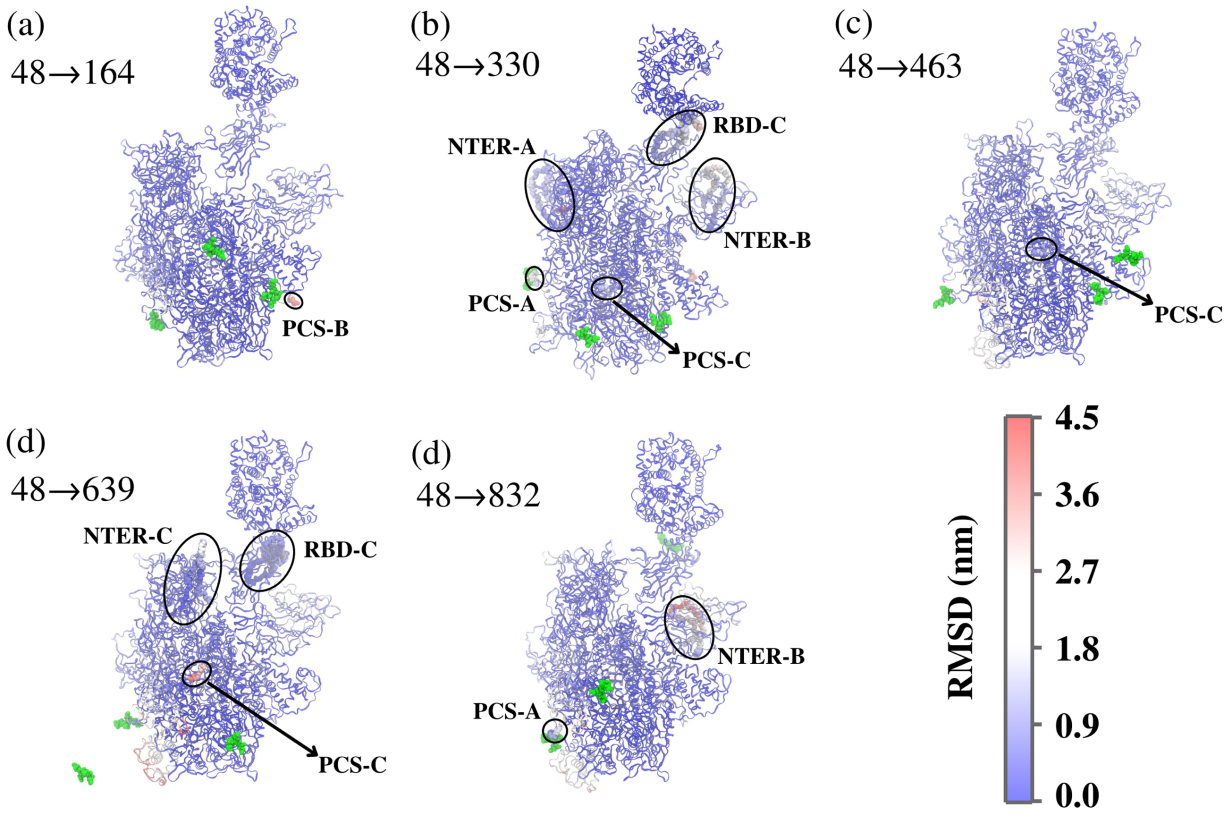

Figure 3. Atomistic simulation snapshots showing the propagation of fluctuations of the spike protein in the presence of three tetrapeptides EELE. Snapshots of the protein at the center of each stage when three EELE tetrapeptides initially bind to the PCS-A, PCS-B, and PCS-C motifs simultaneously. Every residue is colored by the deviation of its α-carbon atom upon the labeled transition, while our motifs of interest with notable changes of RMSD upon consecutive transitions are highlighted and circled. In addition, a color scale bar is shown to represent the variance of RMSD from 0.0 nm to 4.5 nm.

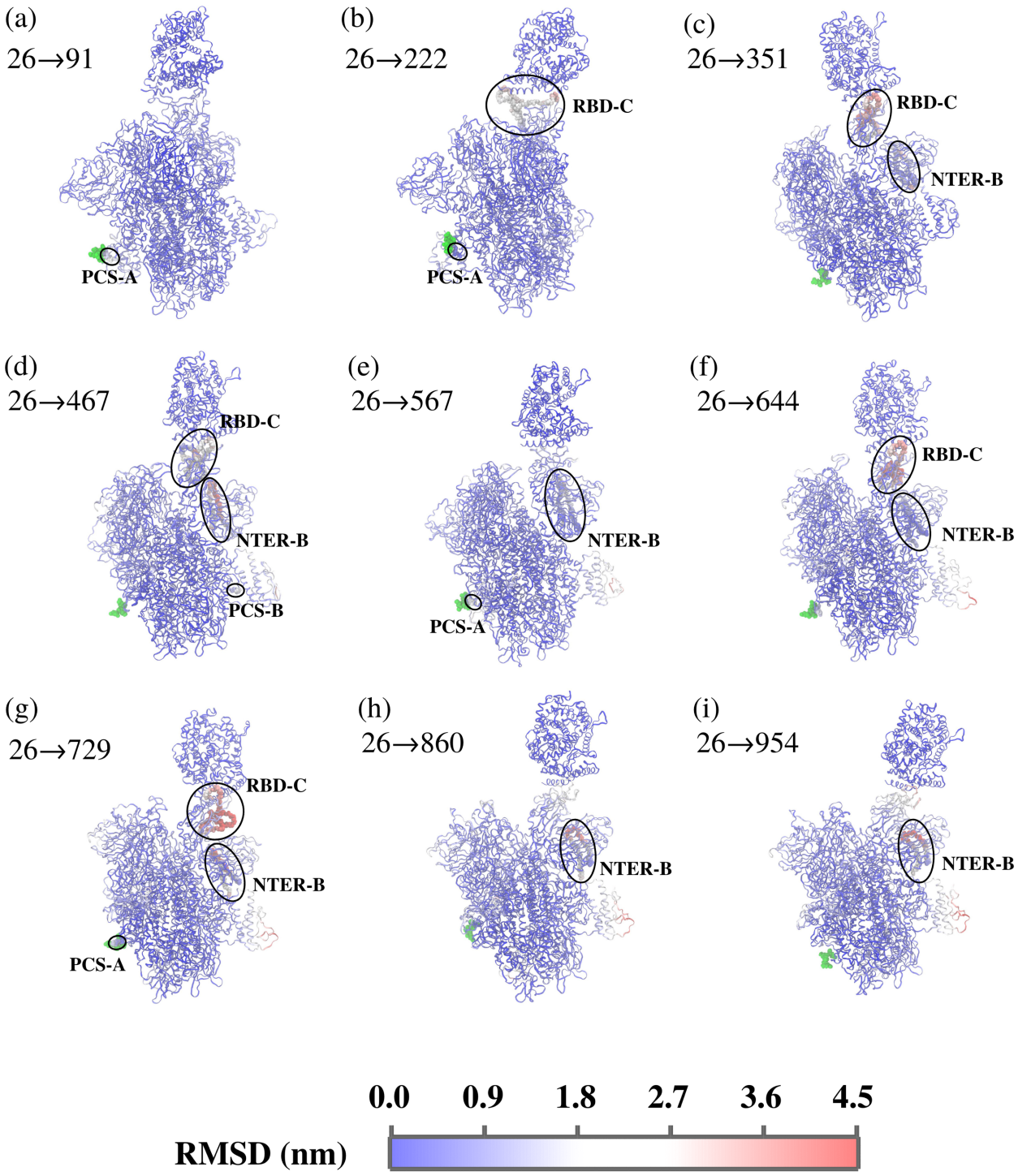

Figure 4. Atomistic simulation snapshots showing the propagation of fluctuations of the spike protein in the presence of only one tetrapeptide EELE. Snapshots of the protein at the center of each stage. Every residue is colored by the deviation of its α-carbon atom upon the labeled transition, while our motifs of interest with notable change of RMSD upon consecutive transitions are highlighted and circled. In addition, a color scale bar is shown to represent the variance of RMSD from 0.0 nm to 4.5 nm.

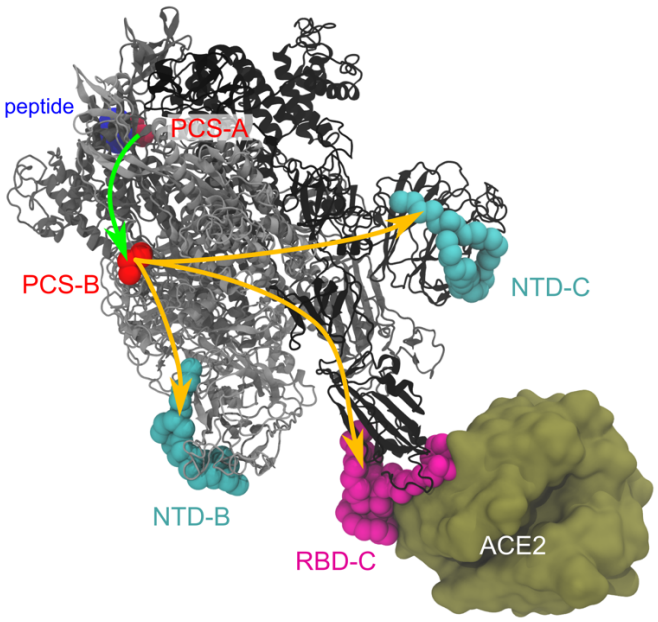

Figure. 5. Schematic representation of the pathway of the allosteric regulation in SARS-CoV-2 spike protein. The pathway is indicated by the arrows. The three monomers of the trimeric spike protein are colored in silver/gray/black for the monomer A/B/C, respectively. The EELE peptide was illustrated by blue balls, which is located at the back of the spike protein in this view angle. PCSs are colored in red, NTDs in cyan. RBD-C is colored in magenta, which is in direct contact with ACE2 (in tan).

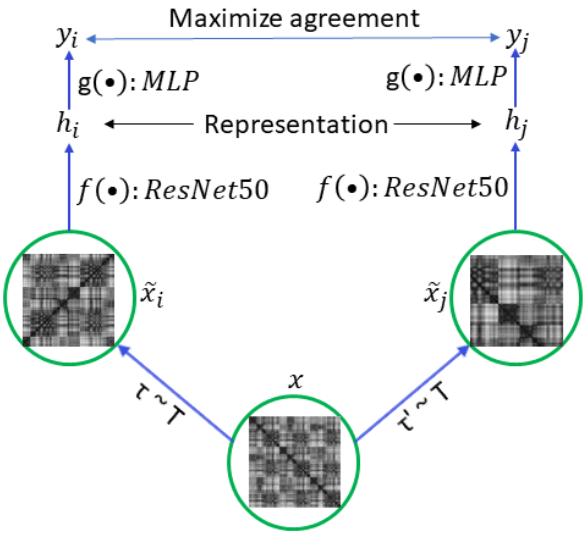

Figure 6. Contrastive learning contact map feature extraction framework. $(\tilde{x}_i, \tilde{x}_j)$ is a positive pair if $\tilde{x}_i$ and $\tilde{x}_j$ are correlated views of the same contact map x, generated by stochastic data augmentations τ~T and τ'~T, respectively. $f(\cdot)$ is the backbone representation encoder. Resnet50 is used for this purpose. $g(\cdot)$ is a small projection head, which is a multilayer perceptron (MLP) with one hidden layer. Both $f(\cdot)$ and $g(\cdot)$ are trained to maximize the agreement between the positive pairs of augmented views of contact maps using the loss function defined in equation (1).

References:

1. Motlagh HN, Wrabl JO, Li J, Hilser VJ. The ensemble nature of allostery. *Nature* **508**, 331-339 (2014).

2. Goodey NM, Benkovic SJ. Allosteric regulation and catalysis emerge via a common route. *Nature Chemical Biology* **4**, 474-482 (2008).

3. Kuriyan J, Eisenberg D. The origin of protein interactions and allostery in colocalization. *Nature* **450**, 983-990 (2007).

4. Xie J, Lai L. Protein topology and allostery. *Current Opinion in Structural Biology* **62**, 158-165 (2020).

5. Arkin Michelle R, Tang Y, Wells James A. Small-Molecule Inhibitors of Protein-Protein Interactions: Progressing toward the Reality. *Chemistry & Biology* **21**, 1102-1114 (2014).

6. Zhang T*, et al.* Protein–ligand interaction detection with a novel method of transient induced molecular electronic spectroscopy (TIMES): experimental and theoretical studies. *ACS central science* **2**, 834-842 (2016).

7. Guarnera E, Berezovsky IN. Allosteric drugs and mutations: chances, challenges, and necessity. *Current Opinion in Structural Biology* **62**, 149-157 (2020).

8. Abdel-Magid AF. Allosteric modulators: an emerging concept in drug discovery. *ACS Med Chem Lett* **6**, 104-107 (2015).

9. Nussinov R, Tsai C-J. Allostery in Disease and in Drug Discovery. *Cell* **153**, 293-305 (2013).

10. Faure AJ, Domingo J, Schmiedel JM, Hidalgo-Carcedo C, Diss G, Lehner B. Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175-183 (2022).

11. Machhi J*, et al.* The natural history, pathobiology, and clinical manifestations of SARS-CoV-2 infections. *Journal of Neuroimmune Pharmacology* **15**, 359-386 (2020).

12. Zuo YY, Uspal WE, Wei T. Airborne transmission of COVID-19: aerosol dispersion, lung deposition, and virus-receptor interactions. *ACS nano* **14**, 16502-16524 (2020).

13. Hou YJ*, et al.* SARS-CoV-2 reverse genetics reveals a variable infection gradient in the respiratory tract. *Cell* **182**, 429-446. e414 (2020).

14. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. Protein structure and sequence reanalysis of 2019-nCoV genome refutes snakes as its intermediate host and the unique similarity between its spike protein insertions and HIV-1. *Journal of proteome research* **19**, 1351-1360 (2020).

15. Wrapp D*, et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260-1263 (2020).

16. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281-292. e286 (2020).

17. Chi X*, et al.* A Neutralizing Human Antibody Binds to the N-Terminal Domain of the Spike Protein of SARS-CoV-2. *Science*, eabc6952 (2020).

18. Yuan M*, et al.* A Highly Conserved Cryptic Epitope in the Receptor Binding Domains of SARS-CoV-2 and SARS-CoV. *Science* **368**, 630-633 (2020).

19. Wang C*, et al.* A Human Monoclonal Antibody Blocking SARS-CoV-2 Infection. *Nature Communications* **11**, 2251 (2020).

20. Qiao B, Olvera de la Cruz M. Enhanced Binding of SARS-CoV-2 Spike Protein to Receptor by Distal Polybasic Cleavage Sites. *ACS Nano* **14**, 10616-10623 (2020).

21. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The Spike Glycoprotein of the new Coronavirus 2019-nCoV Contains a Furin-Like Cleavage Site Absent in CoV of the Same Clade. *Antiviral Research* **176**, 104742 (2020).

22. Wang Q, Qiu Y, Li J-Y, Zhou Z-J, Liao C-H, Ge X-Y. A Unique Protease Cleavage Site Predicted in the Spike Protein of the Novel Pneumonia Coronavirus (2019-nCoV) Potentially Related to Viral Transmissibility. *Virol Sin*, 1-3 (2020).

23. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e286 (2020).

24. Bhowmik D, Gao S, Young MT, Ramanathan A. Deep clustering of protein folding simulations. *BMC bioinformatics* **19**, 47-58 (2018).

25. Chen J, Xu E, Wei Y, Chen M, Wei T, Zheng S. Graph Clustering Analyses of Discontinuous Molecular Dynamics Simulations: Study of Lysozyme Adsorption on a Graphene Surface. *Langmuir* **38**, 10817-10825 (2022).

26. Kolesnikov A, Zhai X, Beyer L. Revisiting self-supervised visual representation learning. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*) (2019).

27. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:180307728*, (2018).

28. Doersch C, Gupta A, Efros AA. Unsupervised visual representation learning by context prediction. In: *Proceedings of the IEEE international conference on computer vision*) (2015).

29. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA. Context encoders: Feature learning by inpainting. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*) (2016).

30. Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. In: *European conference on computer vision*). Springer (2016).

31. Zhang R, Isola P, Efros AA. Colorful image colorization. In: *European conference on computer vision*). Springer (2016).

32. Souza PCT, Thallmair S, Marrink SJ, Mera-Adasme R. An Allosteric Pathway in Copper, Zinc Superoxide Dismutase Unravels the Molecular Mechanism of the G93A Amyotrophic Lateral Sclerosis-Linked Mutation. *The Journal of Physical Chemistry Letters* **10**, 7740-7744 (2019).

33. Hess B, Kutzner C, van der Spoel D, Lindahl E. GROMACS 4: Algorithms for Highly Efficient, Load Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput* **4**, 435-447 (2008).

34. Huang J*, et al.* CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat Meth* **14**, 71-73 (2017).

35. MacKerell AD*, et al.* All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J Phys Chem B* **102**, 3586-3616 (1998).

36. Miyamoto S, Kollman PA. SETTLE: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J Comput Chem* **13**, 952-962 (1992).

37. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-Learning Contact-Map Guided Protein Structure Prediction in CASP13. *Proteins* **87**, 1149-1164 (2019).

38. Wu F*, et al.* A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* **579**, 265-269 (2020).

39. Parrinello M, Rahman A. Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. *J Appl Phys* **52**, 7182-7190 (1981).

40. Hess B. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *J Chem Theory Comput* **4**, 116-122 (2008).

41. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A Linear Constraint Solver for Molecular Simulations. *J Comput Chem* **18**, 1463-1472 (1997).