


## RESEARCH ARTICLE

# Prefusion spike protein stabilization through computational mutagenesis

Dong Yan Zhang<sup>1</sup> | Jian Wang<sup>1</sup> | Nikolay V. Dokholyan<sup>1,2</sup> 

<sup>1</sup>Department of Pharmacology, Penn State College of Medicine, Hershey, Pennsylvania

<sup>2</sup>Departments of Biochemistry & Molecular Biology, Penn State College of Medicine, Hershey, Pennsylvania

## Correspondence

Nikolay V. Dokholyan, Department of Pharmacology, Penn State College of Medicine, Hershey, PA 17033-0850, USA. Email: dokh@psu.edu

## Funding information

National Center for Advancing Translational Sciences, Grant/Award Number: UL1TR002014; National Institute of General Medical Sciences, Grant/Award Number: 1R35GM134864; Passan Foundation; The Huck Institutes of the Life Sciences

## Abstract

A novel severe acute respiratory syndrome (SARS)-like coronavirus (SARS-CoV-2) has emerged as a human pathogen, causing global pandemic and resulting in over 400 000 deaths worldwide. The surface spike protein of SARS-CoV-2 mediates the process of coronavirus entry into human cells by binding angiotensin-converting enzyme 2 (ACE2). Due to the critical role in viral-host interaction and the exposure of spike protein, it has been a focus of most vaccines' developments. However, the structural and biochemical studies of the spike protein are challenging because it is thermodynamically metastable. Here, we develop a new pipeline that automatically identifies mutants that thermodynamically stabilize the spike protein. Our pipeline integrates bioinformatics analysis of conserved residues, motion dynamics from molecular dynamics simulations, and other structural analysis to identify residues that significantly contribute to the thermodynamic stability of the spike protein. We then utilize our previously developed protein design tool, Eris, to predict thermodynamically stabilizing mutations in proteins. We validate the ability of our pipeline to identify protein stabilization mutants through known prefusion spike protein mutants. We finally utilize the pipeline to identify new prefusion spike protein stabilization mutants.

## KEYWORDS

computational mutagenesis, coronavirus, protein stabilization, spike protein

## 1 | INTRODUCTION

The ongoing outbreak of the novel coronavirus,<sup>1-3</sup> which causes fever, severe respiratory illness, and pneumonia, poses a major public health and governance challenges. The emerging pathogen has been characterized as a new member of the betacoronavirus genus (SARS-CoV-2),<sup>4,5</sup> closely related to several bat coronaviruses and to severe acute respiratory syndrome coronavirus (SARS-CoV).<sup>6,7</sup> Compared with SARS-COV, SARS-CoV-2 appears to be more readily transmitted from human to human, spreading to multiple continents and leading to the World Health Organization (WHO)'s declaration of a pandemic on March 11, 2020.<sup>8</sup> According to the WHO, as of June 9, 2020, there had been >7 000 000 confirmed cases globally, leading to >400 000 deaths.

In the initial step of the infection, the coronavirus enters the host cell by binding to a cellular receptor and fusing the viral membrane with the target cell membrane.<sup>9</sup> For SARS-CoV-2, this process is mediated by spike glycoprotein which is a homo-trimer.<sup>10</sup> In the process of viruses fusing to host cells, the spike protein undergoes structural rearrangement and transits from a metastable prefusion conformational state to a highly stable post-fusion conformational state.<sup>11,12</sup> The spike protein comprises of two functional units, S1 and S2 subunits; when fused to the host cell, the two subunits are cleaved. The S1 subunit is responsible for binding to the angiotensin-converting enzyme 2 (ACE2)<sup>13-15</sup> receptor on the host cell membrane and it contains the N-terminal domain (NTD), the receptor-binding domain (RBD) and the C-terminal domain (CTD). NTD in the S1 subunit assists recognize sugar receptors. RBD in the S1 subunit is critical

for the binding of coronavirus to the ACE-2 receptor.<sup>16-19</sup> CTD in the S1 subunit could recognize other receptors.<sup>20</sup> The binding of RBD to ACE2 facilitates the cleavage of the spike protein and promotes the dissociation of the S1 subunit from the S2 subunit.<sup>21</sup> S2 contains two heptad repeats (HR1 and HR2), a fusion peptide, and a protease cleavage site (S2'). The dissociation of S1 induces S2 to undergo a dramatic structural change to fuse the host and viral membranes. Thus, the spike protein serves as a target for development of antibodies, entry inhibitors and vaccines.<sup>22</sup> Coronavirus transits from a metastable prefusion state to a highly stable post-fusion state as part of the spike protein's role in membrane fusion. The instability of the prefusion state presents a significant challenge for the production of protein antigens for antigenic presentation of the prefusion antibody epitopes that are most likely to lead to neutralizing responses. Thus, since the prefusion spike protein exists in a thermodynamically metastable state,<sup>23</sup> a stabilized mutant conformation is critical for the development of vaccines and drugs.

Computational mutagenesis is an effective approach to finding mutations that are able to stabilize proteins. We have previously developed a protein design platform, Eris,<sup>24,25</sup> which utilizes a physical force field<sup>26</sup> for modeling inter-atomic interactions, as well as fast side-chain packing and backbone relaxation algorithms to enable efficient and transferrable protein molecular design. Originally, Eris has been validated on 595 mutants from five proteins, corroborating the unbiased force field, side-chain packing and backbone relaxation algorithms. In many later studies, Eris has been validated through prediction of thermodynamically stabilizing or destabilizing mutations,<sup>27-32</sup> and direct protein design efforts.<sup>33-36</sup>

In this work, we propose a pipeline to automatically stabilize spike proteins through computational mutagenesis. Within the pipeline, we first analyze the conservation score and solvent accessible surface area (SASA) of residues in the protein. We then perform discrete molecular dynamics (DMD)<sup>37-40</sup> simulations to calculate the root mean square fluctuation (RMSF) of residues to analyze their flexibility. Based on this information, we select appropriate residues as mutation sites. We subject the selected residues to computational redesign using Eris to find the stabilizing mutations by calculating the change in free energy  $\Delta\Delta G = \Delta G_{\text{mut}} - \Delta G_{\text{WT}}$ , where  $\Delta G_{\text{mut}}$  and  $\Delta G_{\text{WT}}$  are the free energies of the mutant protein and wild type proteins correspondingly. We utilize this pipeline to identify stabilization mutants of the spike protein. Next, we describe our methods in detail and provide a list of stabilizing mutations for spike protein.

## 2 | METHODS

### 2.1 | Remodeling of the spike protein in different states

We utilize the crystal structure of the prefusion conformation<sup>10</sup> of the spike protein (6VSB) as the template to model the spike protein through homology modeling by using MODELER.<sup>41</sup> We generate five models and select the model with the lowest Discrete Optimized Protein Energy (DOPE),<sup>42</sup> which is a statistical potential used to assess

homology models in protein structure prediction, as the representative conformation of the spike protein. Since steric clashes are common in modeled and low-resolution structures, we employ Chiron<sup>43</sup> to optimize the structure of the spike protein. Chiron resolves atomic clashes by performing short-DMD<sup>37-40,44,45</sup> simulations on protein structures with minimal or no perturbation to the backbone. Both the template structure and the modeled structure have one RBD in up conformation and two RBDs in down conformation (1-up-2-down). We utilize PyMol<sup>46</sup> to create a trimeric spike protein structure with all three RBDs in up conformation (3up) and another trimeric structure with all three RBDs in down conformation (3-down), respectively. Based on the modeled structure with all missing atoms and residues completed, we first duplicate the monomer structure with the RBD in down conformation (down-monomer) and then align it to the monomer structure with the RBD in the up conformation (up-monomer). Then, we delete the up-monomer to get the 3-down trimeric structure. Similarly, we duplicate two up-monomers in the modeled 1-up-2-down structure and then align them to the two down-monomers, respectively. Finally, we delete the two original down-monomers to get a 3-up trimeric structure of the spike protein.

### 2.2 | Multiple sequence alignment and conservation score

We use ConSurf<sup>47</sup> to investigate the conservation score of the spike protein. The ConSurf server is a bioinformatics tool for estimating the evolutionary conservation of amino/nucleic acid positions in a protein/DNA/RNA molecule based on the phylogenetic relations between homologous sequences.

### 2.3 | Molecular dynamics simulation

We use DMD,<sup>37-40</sup> an event-driven simulation which employs a discrete potential energy that relies on the calculation of atomic collisions. In DMD, the all-atom protein model interacts through the implicit solvent force-field, Medusa, which includes the van der Waals interaction, hydrogen bonds, electrostatic, dihedral and angular interactions, and those due to the effective solvent. We perform 1 000 000 step DMD simulation for the cryo-EM structure of the prefusion spike protein, with the temperature set as 0.4 using the ANDERSON thermostat. The heat exchange rate is 0.1. A cubic box (500 Å × 500 Å × 500 Å) with periodic boundary condition has been employed. For the analysis, we only consider equilibrated trajectories, discarding the first part of the trajectories that are not equilibrated due to the initial high energy. The analyses of RMSF and RMSD are performed using MDAnalysis.<sup>48</sup>

### 2.4 | Computational mutagenesis

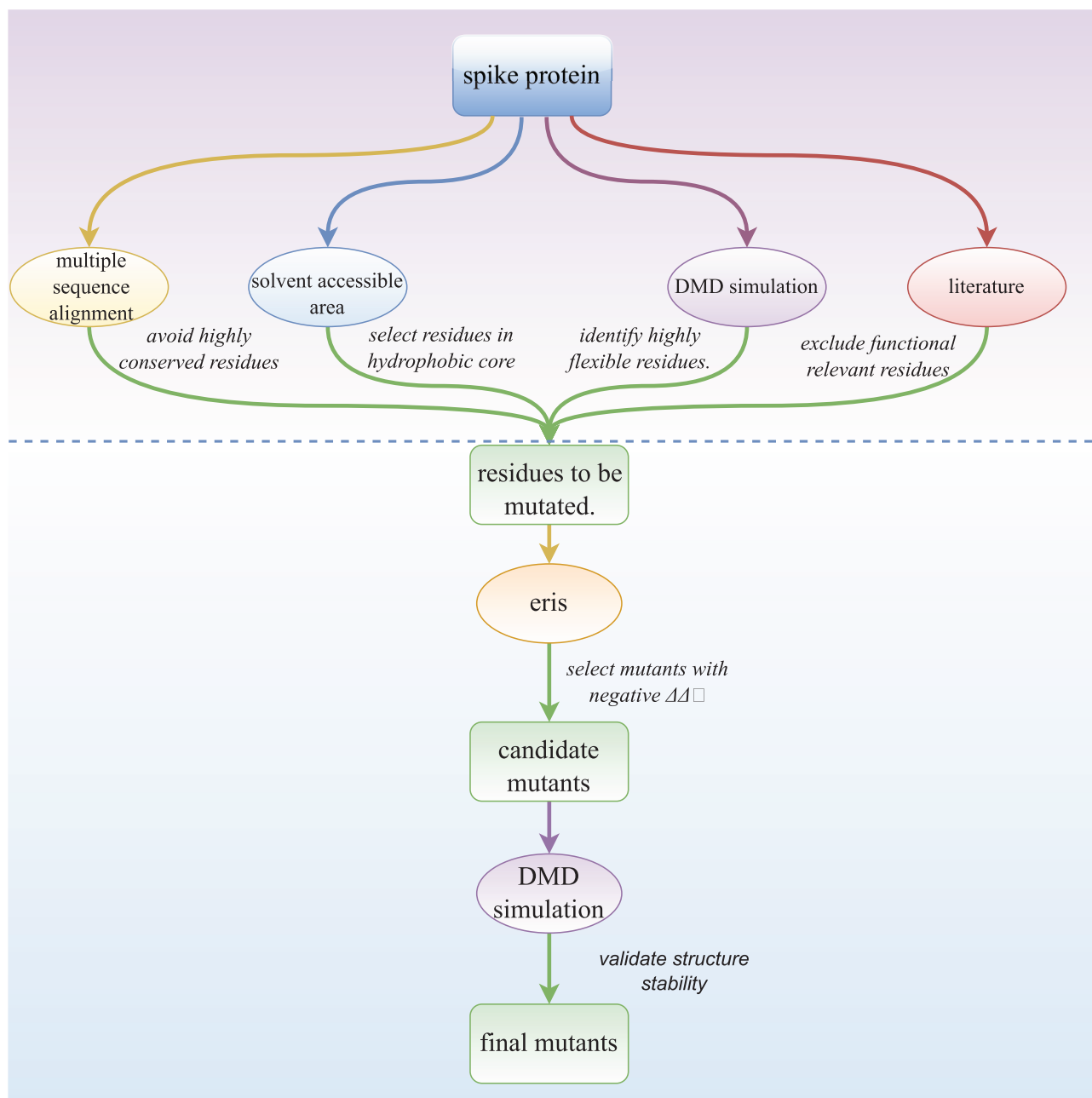
The spike protein structure is subject to *in silico* mutagenesis studies using Eris molecular suite. Eris's protocol induces mutation in proteins

and estimates free energies of mutant ( $\Delta G_{\text{mut}}$ ) and wild type ( $\Delta G_{\text{wt}}$ ) conformations. Eris<sup>24</sup> performs rapid side-chain repacking and backbone relaxation around the mutation site using the Monte-Carlo algorithm and subsequently evaluates  $\Delta G_{\text{wt}}$  and  $\Delta G_{\text{mut}}$  using Medusa force field. Then, Eris's algorithm computes the change in free energy of the protein upon mutation by employing the following formula:  $\Delta\Delta G_{\text{mut}} = \Delta G_{\text{mut}} - \Delta G_{\text{wt}}$ . Finally, Eris evaluates  $\Delta\Delta G_{\text{mut}}$  values to estimate the stabilizing ( $\Delta\Delta G_{\text{mut}} < 0$ ) or destabilizing ( $\Delta\Delta G_{\text{mut}} > 0$ ) effect of the mutations.

### 3 | RESULTS

#### 3.1 | Protein stabilization pipeline

We propose a pipeline to automatically find stabilized mutants of proteins (Figure 1). The pipeline can be divided into two stages. In the first stage, users designate the protein of interest, and then the pipeline will analyze the 3D structure of the protein by using different metrics. In the second stage, users designate the mutation sites



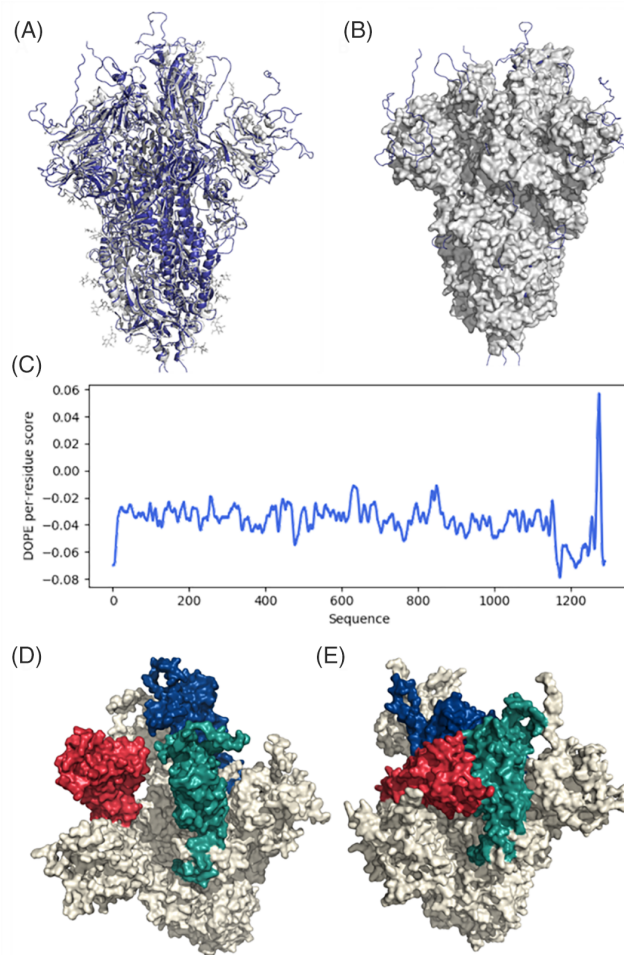
**FIGURE 1** The pipeline of the stabilization of spike protein. The pipeline is roughly divided into two stages. In the first stage, users designate the protein of interest through either the 3D structure of the PDB ID. The pipeline will then analyze the conservation score, solvent accessible surface area (SASA), and root mean square fluctuation (RMSF) of each residue in the protein. In the second stage, users designate the mutation sites for stabilization mutagenesis. The pipeline then utilizes Eris to identify the stabilizing mutantions. Finally, the stabilization capability of these mutants is validated by discrete molecular dynamics (DMD) simulations

according to the analysis of the 3D structure of the protein, and finally the pipeline will determine the stabilizing mutations of the protein. Users can either upload the 3D structure of the protein or input the PDB ID of the protein to designate the protein of interest. In the first stage, the first step is to remodel the 3D structure of the protein of interest to complete the missing atoms and residues. We integrate MODELER into our pipeline to remodel proteins. Next, the pipeline utilizes ConSurf<sup>47</sup> to calculate the conservation score of each residue in the protein of interest. The conservation score indicates the importance of the residue in maintaining protein structure and/or function. Subsequently, the pipeline utilizes DMD to analyze the flexibility of each residue in the spike protein through RMSF. The technique has already been used to efficiently study the protein folding thermodynamics and protein oligomerization and allows for a good equilibration of the structures. Then, the pipeline will calculate SASA of residues in the protein.

In the second stage, users designate the mutation sites according to the conservation score, RMSF, and SASA. A high conservation score ( $\geq 7$ ) indicates the residue may play important roles in the function or the stability of the structure of the protein; residues of high RMSF ( $> 3.5 \text{ \AA}$ ) are likely the culprit to undermine the stability of the structure of the protein, hence we select residues that have a low conservation score or high RMSF; residues with SASA  $< 0.5$  are considered buried and residues with SASA  $\geq 0.5$  are considered exposed to solvent. After the designation of the mutation sites, the pipeline utilizes Eris to determine the changes in free energies of the mutants. For each residue in the mutation sites, Eris will mutate the amino acid type of the residue to the other 19 amino acid types, iteratively. For each mutant, Eris will calculate the  $\Delta\Delta G$ , which is the difference of the free energy of the mutant relative to the wild type. Positive  $\Delta\Delta G$  means the mutant decreases the stability of the protein, and negative  $\Delta\Delta G$  means the mutant can stabilize the protein. The mutants that have the lowest  $\Delta\Delta G$  will be selected as the stabilization mutants.

### 3.2 | Remodeling of the spike protein structure

We utilize MODELER<sup>41</sup> to remodel the 3D structure of the spike protein by using a structure deposited to Protein DataBank (PDB), PDBID:6VSB,<sup>11</sup> the cryo-EM structure of the prefusion state of the spike protein, as the template structure to complete the missing atoms and residues (Figure 2A,B). Next, we use Chiron,<sup>43</sup> a protein energy minimization tool based on DMD, to optimize the modeled structure of the spike protein. We evaluate the modeled structure by calculating the DOPE score of each residue. DOPE of residues in the region (1147-1288) near the C-terminal are higher than that of other regions because this region is not experimentally solved in the template structure 6VSB (Figure 2C). The deposited structure 6VSB is a spike protein trimer, where one RBD is in up conformation and the other two RBDs are in down conformation. RBD domain can only bind to the ACE2 receptor when it is in the up conformation. Based on the modeled structure, we prepared next model two spike protein structures that have all three RBDs in up conformation (Figure 2D) and



**FIGURE 2** Remodeling results of the spike protein structure. A, Comparison of the crystal structure and the remodeled structure of the spike protein. B, The surface representation of the spike protein. The extra blue loops are the completed loops. C, The Discrete Optimized Protein ENergy (DOPE) score of each residue in the structure remodeled by MODELER. D, The remodeled structure of the spike protein with three receptor-binding domain (RBD) in up conformations. E, The remodeled structure of the spike protein with three RBDs in down conformations [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

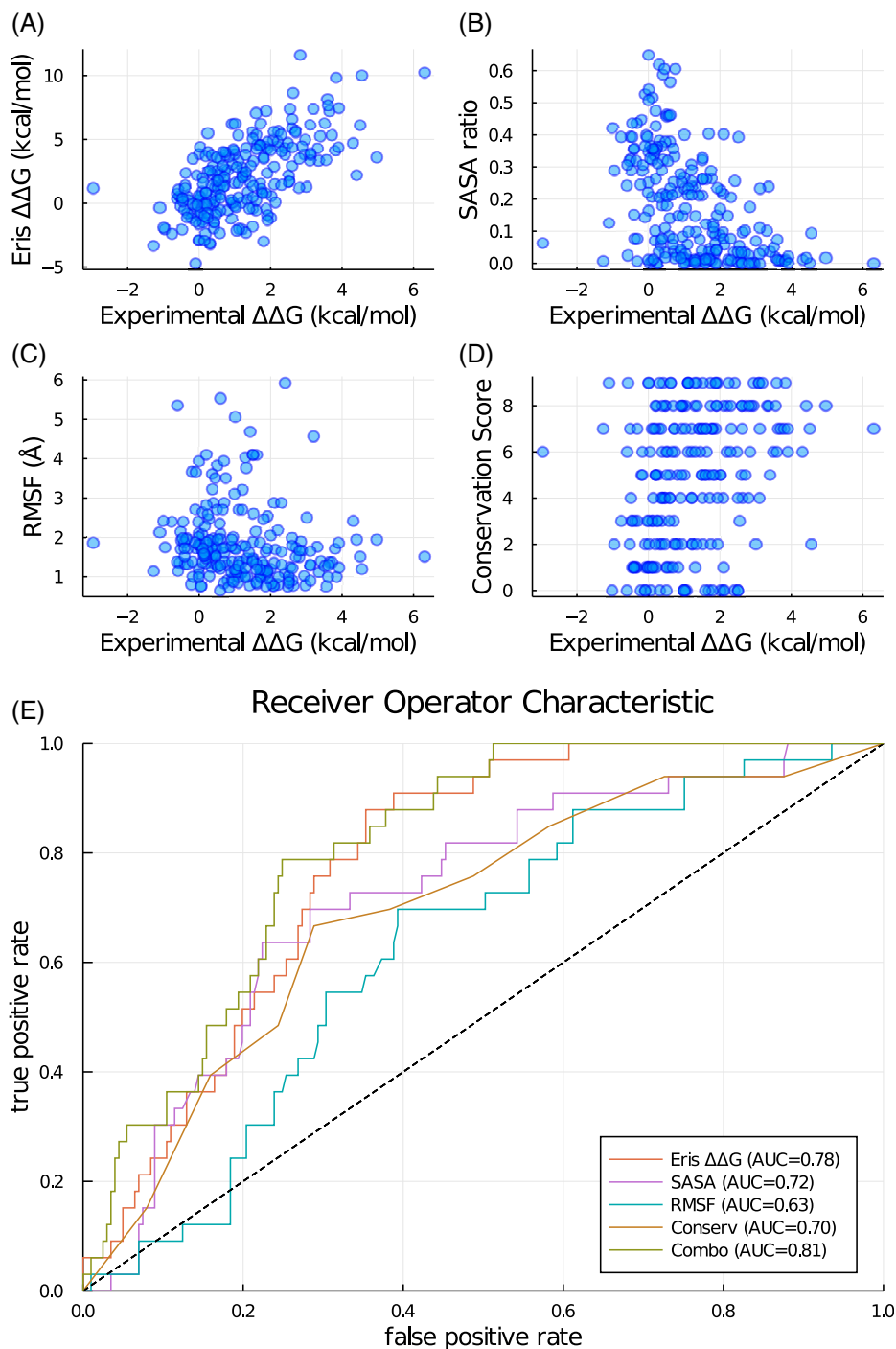
have all three RBDs in down conformation (Figure 2E), respectively. All following computational mutagenesis study are performed using the structure with the three RBDs in down conformation.

### 3.3 | Validation of the protein stabilization capability of the pipeline through known mutants

To validate the ability of the pipeline to identify stabilization mutants, we use the pipeline to calculate the free energy changes of several known prefusion spike protein stabilization mutants. The 2P mutation strategy (K986P and V987P) has been proved effective for the stabilization of spike protein of SARS-COV-2 and other betacoronavirus.<sup>10,18,49</sup> Hsieh et al<sup>50</sup> have tested a large amount of mutants and found the best mutant,

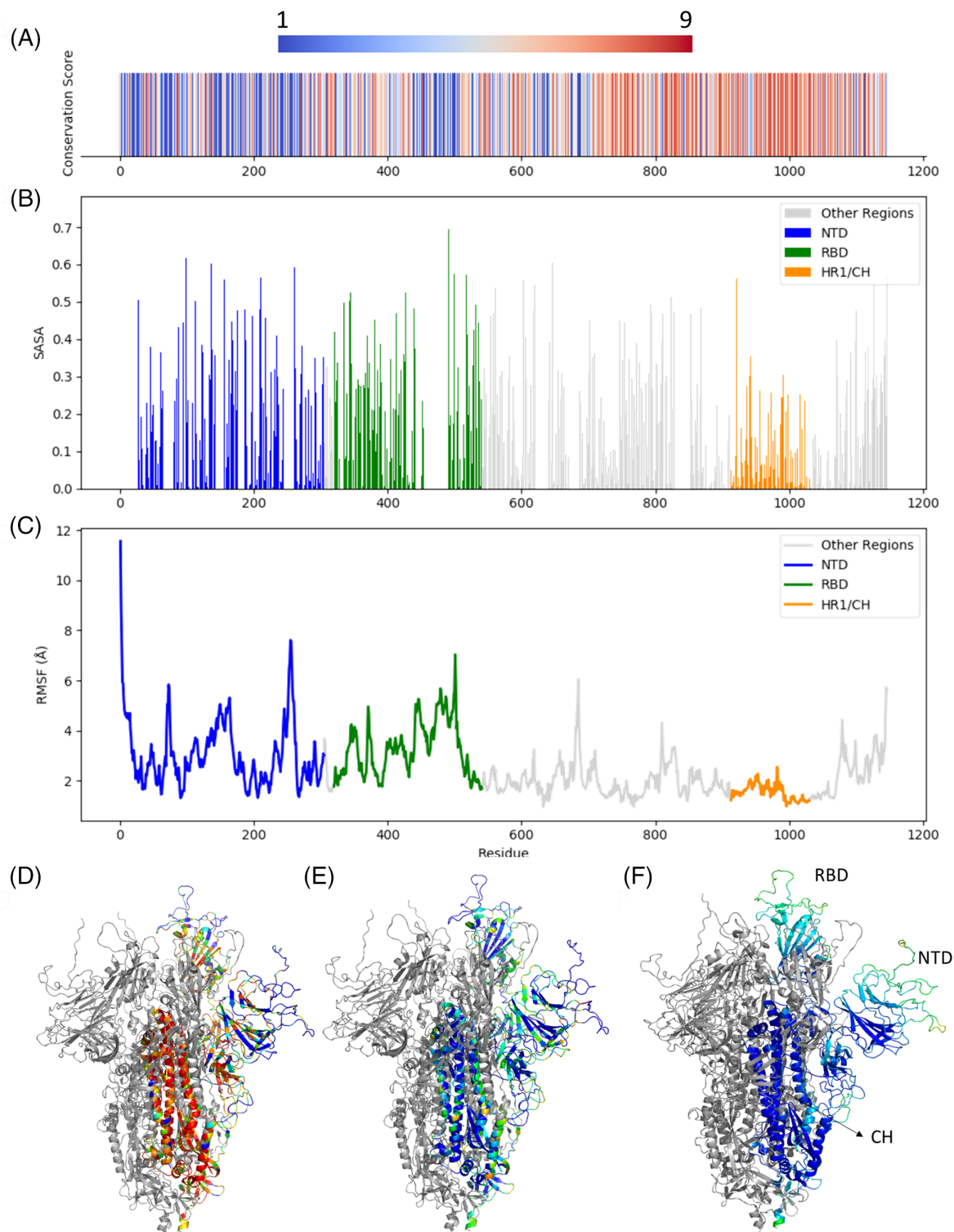
HexaPro, which has six beneficial prolines substitutions (F817P, A892P, A899P, A942P, K986P, and V987P) leading to ~10-fold higher expression. We calculate the free energy change of the 2P mutant and the HexaPro mutant through Eris. The free energy change of the 2P mutant is  $-6.024$  kcal/mol, indicative of the more stable property of the mutant than the wild type structure. The free energy change of the HexaPro mutant is  $-16.143$  kcal/mol, suggesting that it is even more stable than the 2P mutant. Thus, these computational calculations of the stabilities of the 2P mutant and the HexaPro mutant are in agreement with the experimental demonstration of their stability.

We further evaluate the performance of the pipeline on a dataset composed of 28 proteins and 625 mutations. The dataset is compiled by Guerois et al<sup>51</sup> from single and multiple-residue mutation analyses. We can observe a significant positive correlation between Eris  $\Delta\Delta G$  and experimental  $\Delta\Delta G$  (Figure 3A) and a slight positive correlation between conservation score and experimental  $\Delta\Delta G$  (Figure 3D). No obvious correlation is observed between SASA/RMSF and the experimental  $\Delta\Delta G$  (Figure 3B,C). Since our objective is to identify the stabilization mutants (experimental  $\Delta\Delta G < 0$ ), we plot the receiver operator characteristic curve of the four metrics. The true positive



**FIGURE 3** Evaluation of the stabilization mutation identification ability of the pipeline. A, The  $\Delta\Delta G$  predicted by Eris vs experimental  $\Delta\Delta G$  for all mutations in the dataset. B, The solvent accessible surface area (SASA) vs experimental  $\Delta\Delta G$  for all mutations in the dataset. C, The root mean square fluctuation (RMSF) vs experimental  $\Delta\Delta G$  for all mutations in the dataset. D, The conservation score vs experimental  $\Delta\Delta G$  for all mutations in the dataset. E, The receiver operator characteristic (ROC) curve of Eris  $\Delta\Delta G$ , SASA, RMSF, conservation score, and the combination of the four metrics, respectively [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]





**FIGURE 4** Conservation score, solvent accessible surface area (SASA), and root mean square fluctuation (RMSF) of the spike protein. A, The conservation score of residues in the spike protein. Conservation score of 9 means highly conserved, while conservation score of 1 means a highly variable position. B, SASA of residues in the spike protein. C, The RMSF of residues in the spike protein. D, The 3D structures of the spike protein colored by the conservation score. The red/blue colors indicate highly conserved/highly variable residues. E, The 3D structure of the spike protein colored SASA. The red/blue colors indicate exposed/buried residues. F, The 3D structure of the spike protein colored by RMSF. Red means flexible and blue means frozen [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

rate and the false positive rate are calculated by labelling experimental  $\Delta\Delta G < 0$  as positive samples and experimental  $\Delta\Delta G > 0$  as negative

samples. The AUC of Eris  $\Delta\Delta G$ , SASA, RMSF, and conservation score are 0.78, 0.72, 0.63, and 0.70, respectively. Finally, we perform a

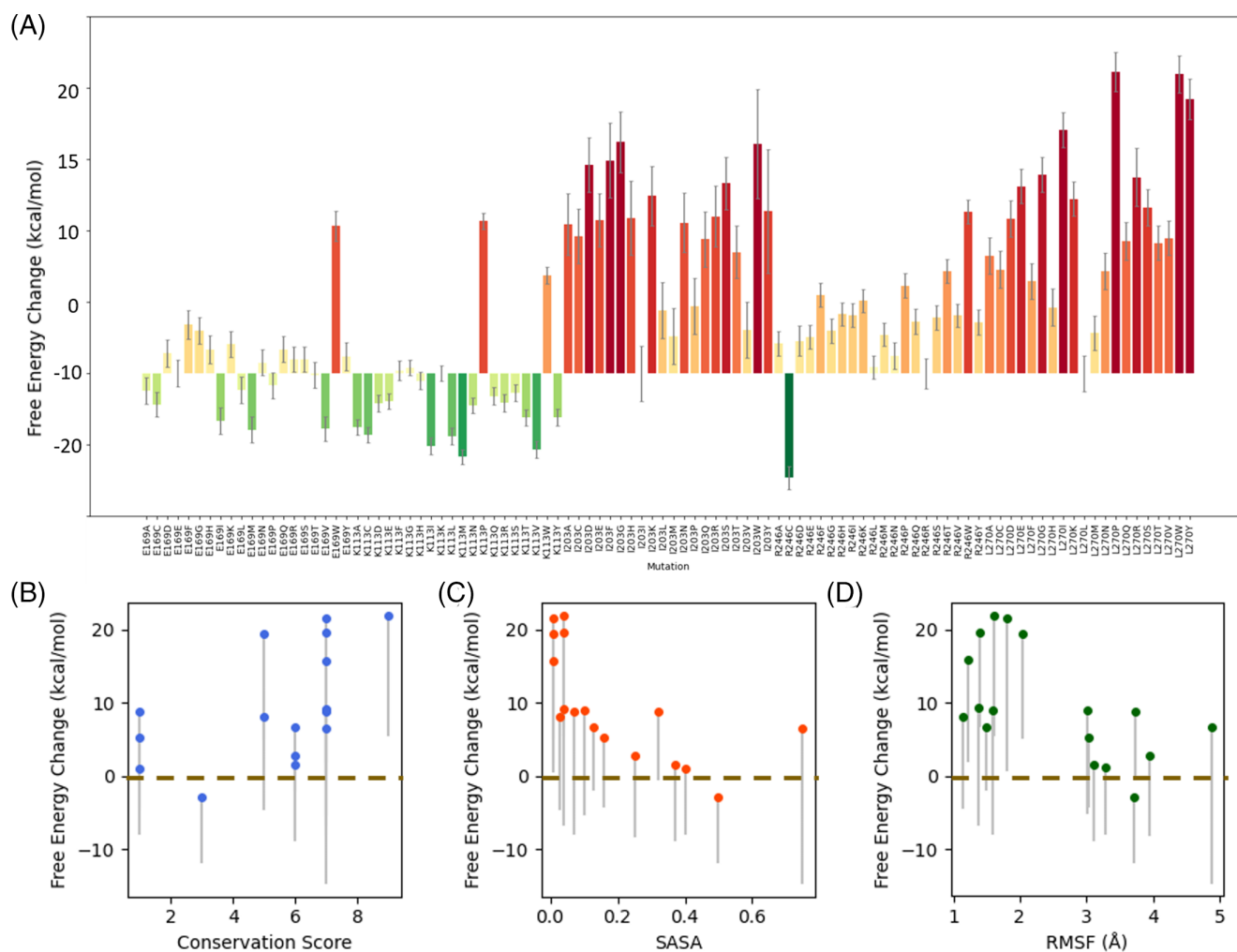
linear regression for the relationship between the four metrics and the experimental  $\Delta\Delta G$  to obtain a best combination of the four metrics ( $0.25*\Delta\Delta G - 0.78*SASA - 0.02*RMSF - 0.23*Normalized Conservation Score + 0.88$ ). The AUC of the combination is improved to 0.81.

### 3.4 | Identification of new mutation sites in the spike protein

The spike protein is mainly composed of S1 and S2 subunits (Figure S1). We select residues for mutation from the NTD and RBD domains in S1, and we also select residues from the HR1 (heptad repeat 1) and CH (central helix) domains in S2. We do not select residues from HR2 domain in S2 because the structure of the HR2 domain has not been solved in the cryo-EM structure 6VSB. At the

outset, we calculate the conservation score (Figure 4A,D) of all residues by using ConSurf. Based on the conservation score, most residues in HR1/CH are conservative, while residues in NTD and RBD are prone to mutation in evolution. Next, we use Pymol<sup>46</sup> to calculate SASA of all residues in the spike protein (Figure 4B,E). SASA indicates the level of residues exposed to the solvent in a protein and usually most of the functional residues are located on the protein structure's surface.<sup>52</sup> All four domains have both low SASA residues and high SASA residues. Then, we perform 1 000 000 steps DMD simulation for the spike protein to calculate the RMSF of each residue (Figure 4C,F). The residues in HR1/CH have extremely low RMSF, while the residues in NTD and RBD domains have moderate to high RMSF.

We select residues that have different conservation scores in these four domains for mutagenesis. In the NTD, we select five residues (Table S1) with the conservation score ranging from 1 (highly



**FIGURE 5** Stabilization results of the spike protein. A, The free energy change ( $\Delta\Delta G$ ) of all mutations on the selected residues in N-terminal domain (NTD) of the spike protein. B, The correlation between free energy change and the conservation score for all mutants of the four domains. The blue dots refer to the average free energy change of all 19 mutants of each residue. The bottom end of each gray line refers to the minimum free energy change of all 19 mutants of the residue. C, The correlation between free energy change and solvent accessible surface area (SASA) for all mutants of residues in the four domains. D, The correlation between free energy change and root mean square fluctuation (RMSF) for all mutants of residues in the four domains [Color figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

variable) to 9 (highly conservative). The SASA of these residues range from 0.01 (buried) to 0.75 (exposed). The RMSF range from 1.61 Å (frozen) to 4.88 Å (flexible). Likewise, in the other three domains (Table S2,S3), we select 5 to 7 residues, respectively. These residues also have diverse conservation scores, SASA, and RMSF. Of note, to avoid affecting the function of the spike protein, these residues are all not chosen from the functional sites of the spike protein, such as the ACE2 binding site in RBD.

### 3.5 | Stabilization mutants of the spike protein

We utilize Eris to calculate the free energy changes of mutants relative to the wild type (Figure 5A, Figure S2,S3, and Table S4-S6). In the NTD, five residues, E169, K113, I203, R246, and L270 are selected for mutagenesis. Among them, the free energy changes of nearly all mutations of residues I203, R246, and L270 are positive, indicating that they are destabilizing the structure. In contrast, most mutations on residues E169 and K113 have negative free energy changes, suggesting that they are stabilizing the structure. The mutant that has the most negative free energy change is R246C. However, cysteine is prone to forming disulfide bond with other cysteine, which may affect the correct folding of the protein structure, so we recommend K113M to be a better choice as the stabilization mutant.

In the RBD, five residues (A411, T415, Y505, N439, and D428) (Table S2) are selected for mutagenesis. The free energy changes calculated by Eris (Figure S2, Table S5) show that most residues have stabilization mutants, which have negative free energy changes, except for Y505. The most stable mutants are T415V and D428M.

In the HR1/CH domain, seven residues (L948, I1018, A1026, Y1007, S1003, T961, and V976) are selected for mutagenesis as shown in Table S3. In stark contrast to NTD and RBD, most residues have extremely high free energy changes (> 30 kcal/mol), suggesting that these residues are not very good choices for mutagenesis. The high free energy changes also implicate that they may play important roles in stabilizing the structure so that they are irreplaceable to some extent. This finding is also in concert with the high conservation scores of these residues. That said, we can still find stabilization mutants for these residues, such as T961A and S1003M (Figure S3).

## 4 | DISCUSSION

Compared to experimental mutagenesis, such as random mutagenesis<sup>53,54</sup> and site-directed mutagenesis,<sup>55,56</sup> computational mutagenesis<sup>24</sup> is an efficient alternative that lays the foundation of large-scale mutation screening. However, performing computational mutation screening for all residues in the spike protein trimeric structure, which consists of 1288 residues in each monomer, is still time-consuming and inefficient, so we only select a subset of residues to perform mutagenesis. Although we only select residues in NTD, RBD, and HR1/CH domains to perform mutagenesis, residues in other regions can also be used as mutation sites. For example, the known 2P

mutation strategy (K986P and V987P) has been proved effective for the stabilization of spike protein of SARS-COV-2 and other betacoronavirus.<sup>10,18,49</sup>

In addition, we exclude residues that play important roles in the function of the spike protein, such as the N-linked glycosylation sites<sup>57</sup> and the ACE2-binding surface in the RBD domain. It has been observed that surfaces with high density of glycans can enable immune recognition,<sup>58</sup> and the ACE2-binding surface in the RBD domain is essential for the binding between the spike protein and ACE2. Thus, we avoid mutating residues in these sites, because such substitutions may disrupt the function of the spike protein. In addition, glycans attached in the spike protein may significantly affect the SASA, RMSF, and free energy of residues in the glycosylation sites. Although we avoid mutating residues in the glycosylation sites, residues in other regions may be affected through allosteric interactions<sup>59-61</sup> in the protein. In future, we will also evaluate allosteric effects in identifying stabilization mutation through the pipeline.

In this work, we propose a pipeline to automatically stabilize proteins through computational mutagenesis. We analyze the conservation score, RMSF, and SASA of residues in the spike protein through the pipeline. We propose criteria based on the conservation score, RMSF, and SASA to identify residues for mutation. Finally, we utilize Eris to calculate the free energy change and find stabilizing mutants.

### ACKNOWLEDGEMENTS

We acknowledge support from the National Institutes for Health 1R35 GM134864, The Huck Institutes of the Life Sciences, and the Passan Foundation. The project described was also supported by the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1 TR002014. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

### CONFLICT OF INTERESTS

The authors declare no potential conflict of interest.

### CODE AVAILABILITY

All source codes are deposited in: <https://bitbucket.org/dokhlab/protein-stabilization>.

### PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1002/prot.26025>.

### ORCID

Nikolay V. Dokholyan  <https://orcid.org/0000-0002-8225-4025>

### REFERENCES

1. Zhu N, Zhang D, Wang W, et al. A novel coronavirus from patients with pneumonia in China, 2019. *N Engl J Med*. 2020;382:727-733.
2. Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C. & Di Napoli, R. Features, evaluation and treatment coronavirus (COVID-19). in Statpearls [internet] (StatPearls Publishing), 2020.



3. Repici A, Maselli R, Colombo M, et al. Coronavirus (COVID-19) outbreak: what the department of endoscopy should know. *Gastrointest Endosc.* 2020;92:192-197.
4. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. *Nat Med.* 2020;26:450-452.
5. Tan W, Zhao X, Ma X, et al. A novel coronavirus genome identified in a cluster of pneumonia cases – Wuhan, China 2019–2020. *China CDC Wkly.* 2020;2:61-62.
6. Lau SKP, Woo PCY, Li KSM, et al. Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats. *Proc Natl Acad Sci U S A.* 2005;102:14040-14045.
7. Lai C-C, Shih T-P, Ko W-C, Tang H-J, Hsueh P-R. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and corona virus disease-2019 (COVID-19): the epidemic and the challenges. *Int J Antimicrob Agents.* 2020;55(3):105924.
8. Sohrabi C, Alsafi Z, O'Neill N, et al. World Health Organization declares global emergency: a review of the 2019 novel coronavirus (COVID-19). *Int J Surg.* 2020;76:71-76.
9. Xia S, Zhu Y, Liu M, et al. Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. *Cell Mol Immunol.* 2020;17(7):765-767. <http://dx.doi.org/10.1038/s41423-020-0374-2>.
10. Wrapp D, Wang N, Corbett KS, et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science.* 2020;367:1260-1263.
11. Li F. Structure, function, and evolution of coronavirus spike proteins. *Annu Rev Virol.* 2016;3:237-261.
12. Bosch BJ, van der Zee R, de Haan CAM, Rottier PJM. The coronavirus spike protein is a class I virus fusion protein: structural and functional characterization of the fusion core complex. *J Virol.* 2003;77:8801-8811.
13. Donoghue M, Hsieh F, Baronas E, et al. A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1-9. *Circ Res.* 2000;87:e1-e9.
14. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell.* 2020;181:271-280.
15. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science.* 2020;367(6485):1444-1448. <http://dx.doi.org/10.1126/science.abb2762>.
16. Yuan Y, Cao D, Zhang Y, et al. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat Commun.* 2017;8:15092.
17. Gui M, Song W, Zhou H, et al. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res.* 2017;27:119-129.
18. Pallesen J, Wang N, Corbett KS, et al. Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. *Proc Natl Acad Sci U S A.* 2017;114:1026-1039.
19. Walls AC, Xiong X, Park YJ, et al. Unexpected receptor functional mimicry elucidates activation of coronavirus fusion. *Cell.* 2019;176:1026-1039.
20. Shang J, Wan Y, Liu C, et al. Structure of mouse coronavirus spike protein complexed with receptor reveals mechanism for viral entry. *PLoS Pathog.* 2020;16:e1008392.
21. Walls AC, Tortorici MA, Snijder J, et al. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc Natl Acad Sci U S A.* 2017;114:11157-11162.
22. Dhama K, Sharun K, Tiwari R, et al. COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics. *Hum Vaccin Immunother.* 2020;11:1-7.
23. Ou X, Liu Y, Lei X, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun.* 2020;11:1-12.
24. Yin S, Ding F, Dokholyan NV. Eris: an automated estimator of protein stability. *Nat Methods.* 2007;4:466-467.
25. Yin S, Ding F, Dokholyan NV. Modeling backbone flexibility improves protein stability estimation. *Structure.* 2007;15:1567-1576.
26. Ding F, Dokholyan NV. Emergence of protein fold families through rational design. *PLoS Comput Biol.* 2006;2:e85.
27. Torres MP, Lee MJ, Ding F, et al. G protein mono-ubiquitination by the Rsp5 ubiquitin ligase. *J Biol Chem.* 2009;284:8940-8950.
28. Li B, Tunc-Ozdemir M, Urano D, et al. Tyrosine phosphorylation switching of a G protein. *J Biol Chem.* 2018;293:4752-4766.
29. Zhu C, Han Q, Samoshkin A, et al. Stabilization of  $\mu$ -opioid receptor facilitates its cellular translocation and signaling. *Proteins.* 2019;87:878-884.
30. Zhu C, Beck MV, Griffith JD, Deshmukh M, Dokholyan NV. Large SOD1 aggregates, unlike trimeric SOD1, do not impact cell viability in a model of amyotrophic lateral sclerosis. *Proc Natl Acad Sci U S A.* 2018;115:201800187.
31. Fay JM, Zhu C, Proctor EA, et al. A Phosphomimetic mutation stabilizes SOD1 and rescues cell viability in the context of an ALS-associated mutation. *Structure.* 2016;24:1898-1906.
32. Proctor EA, Fee L, Tao Y, et al. Nonnative SOD1 trimer is toxic to motor neurons in a model of amyotrophic lateral sclerosis. *Proc Natl Acad Sci U S A.* 2016;113:614-619.
33. Dagliyan O, Krokhotin A, Ozkan-Dagliyan I, et al. Computational design of chimeric and optogenetic split proteins. *Nat Commun.* 2018;9:4042.
34. Dagliyan O, Tarnawski M, Chu PH, et al. Engineering extrinsic disorder to control protein activity in living cells. *Science.* 2016;354:6800-6804.
35. Dagliyan O, Shirvanyants D, Karginov AV, et al. Rational design of a ligand-controlled protein conformational switch. *Proc Natl Acad Sci U S A.* 2013;110:6800-6804.
36. Zhu C, Dukhovlina E, Council O, et al. Rationally designed carbohydrate-occluded epitopes elicit HIV-1 Env-specific antibodies. *Nat Commun.* 2019;10:948.
37. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Discrete molecular dynamics studies of the folding of a protein-like model. *Fold Des.* 1998;3:577-587.
38. Proctor EA, Ding F, Dokholyan NV. Discrete molecular dynamics. *Wiley Interdiscip Rev Comput Mol Sci.* 2011;1:80-92.
39. Proctor EA, Dokholyan NV. Applications of discrete molecular dynamics in biology and medicine. *Curr Opin Struct Biol.* 2016;37:9-13.
40. Ding F, Tsao D, Nie H, Dokholyan NV. Ab initio folding of proteins with all-atom discrete molecular dynamics. *Structure.* 2008;16:1010-1018.
41. Ali VS, Blundell TL, Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol.* 1993;234:779-815.
42. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 2006;15:2507-2524.
43. Convertino M, Dokholyan NV. Computational modeling of small molecule ligand binding interactions and affinities. *Methods Mol Biol.* 2016;1414:23-32.
44. Brodie NI, Popov KI, Petrotchenko EV, Dokholyan N, Borchers C. Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Sci Adv.* 2017;3:e1700479.
45. Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE, Dokholyan NV. Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA.* 2008;14:1164-1173.
46. DeLano WL, Ultsch MH, Wells JA. Convergent solutions to binding at a protein-protein interface. *Science.* 2000;287:1279-1283.
47. Glaser F, Pupko T, Paz I, et al. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. *Bioinformatics.* 2003;19:163-164.

48. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem*. 2011;32:2319-2327.
49. Kirchdoerfer RN et al. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci Rep*. 2018;8:1-11.
50. Hsieh CL, Goldsmith JA, Schaub JM, et al. Structure-based Design of Prefusion-stabilized SARS-CoV-2 spikes. *Science*. 2020;369:1501-1505.
51. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*. 2002;320:369-387.
52. Heffernan R, Paliwal K, Lyons J, et al. Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep*. 2015;5:1-11.
53. Wong TS, Tee KL, Hauer B, Schwaneberg U. Sequence saturation mutagenesis (SeSaM): a novel method for directed evolution. *Nucleic Acids Res*. 2004;32:e26-e26.
54. Bloom J, Meyer M, Meinhold P, Otey C, Macmillan D, Arnold F. Evolving strategies for enzyme engineering. *Current Opinion in Structural Biology*. 2005;15(4):447-452. <http://dx.doi.org/10.1016/j.sbi.2005.06.004>.
55. Shortle D, DiMaio D, Nathans D. Directed mutagenesis. *Annu Rev Genet*. 1981;15:265-294.
56. Flavell RA, Sabo DL, Bandle EF, Weissmann C. Site-directed mutagenesis: effect of an extracistronic mutation on the in vitro propagation of bacteriophage Qbeta RNA. *Proc Natl Acad Sci U S A*. 1975;72:367-371.
57. Watanabe Y, Berndsen ZT, Raghvani J, et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat Commun*. 2020;11:1-10.
58. Watanabe Y, Bowden TA, Wilson IA, Crispin M. Exploitation of glycosylation in enveloped virus pathobiology. *Biochim Biophys Acta*. 2019;1863:1480-1497.
59. Dokholyan NV. Controlling allosteric networks in proteins. *Chem Rev*. 2016;116:6463-6487.
60. Wang J, Jain A, McDonald LR, Gambogi C, Lee AL, Dokholyan NV. Mapping allosteric communications within individual proteins. *Nat Commun*. 2020;11(1):1-13. <http://dx.doi.org/10.1038/s41467-020-17618-2>.
61. Dokholyan N, Shakhnovich EI. Understanding hierarchical protein evolution from first principles. *J Mol Biol*. 2001;312:289-307.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Zhang DY, Wang J, Dokholyan NV. Prefusion spike protein stabilization through computational mutagenesis. *Proteins*. 2021;89:399-408. <https://doi.org/10.1002/prot.26025>