Computational design of SARS-CoV-2 spike glycoproteins to increase immunogenicity by T cell epitope engineering

Edison Ong a,1, Xiaoqiang Huang a,1, Robin Pearce a, Yang Zhang a,b,* , Yongqun He a,c,*

a Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA
b Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA
c Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109, USA

A R T I C L E   I N F O

Article history:
Received 21 October 2020
Received in revised form 24 December 2020
Accepted 24 December 2020
Available online 31 December 2020

Keywords:
COVID-19
Vaccine
Spike glycoprotein
Epitope engineering
Structural vaccinology
EvoDesign

A B S T R A C T

The development of effective and safe vaccines is the ultimate way to efficiently stop the ongoing COVID-19 pandemic, which is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Built on the fact that SARS-CoV-2 utilizes the association of its Spike (S) protein with the human angiotensin-converting enzyme 2 (ACE2) receptor to invade host cells, we computationally redesigned the S protein sequence to improve its immunogenicity and antigenicity. Toward this purpose, we extended an evolutionary protein design algorithm, EvoDesign, to create thousands of stable S protein variants that perturb the core protein sequence but keep the surface conformation and B cell epitopes. The T cell epitope content and similarity scores of the perturbed sequences were calculated and evaluated. Out of 22,914 designs with favorable stability energy, 301 candidates contained at least two pre-existing immunity-related epitopes and had promising immunogenic potential. The benchmark tests showed that, although the epitope restraints were not included in the scoring function of EvoDesign, the top S protein design successfully recovered 31 out of the 32 major histocompatibility complex (MHC)-II T cell promiscuous epitopes in the native S protein, where two epitopes were present in all seven human coronaviruses. Moreover, the newly designed S protein introduced nine new MHC-II T cell promiscuous epitopes that do not exist in the wildtype SARS-CoV-2. These results demonstrated a new and effective avenue to enhance a target protein’s immunogenicity using rational protein design, which could be applied for new vaccine design against COVID-19 and other pathogens.

© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The current Coronavirus Disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in over 77 million confirmed cases and over 1.7 million deaths globally as of December 24, 2020, according to the World Health Organization [1]. Tremendous efforts have been made to develop effective and safe vaccines against this viral infection. The Pfizer-BioNTech BNT162b2 mRNA vaccine showed 95% effectiveness in preventing COVID-19 [2], and the Moderna mRNA-1273 induced strong immune responses among recipients between the age of 18 and 55 during phase III clinical trials [3]. The FDA issued an Emergency Use Authorization for the mRNA-1273 and BNT162b2 mRNA vaccines in record time. On the other hand, the Inovio INO-4800 DNA vaccine not only showed protection from the viral infection in rhesus macaques but was also reported to induce long-lasting memory [4]. In addition to these two vaccines, there are over a hundred COVID-19 vaccines currently in clinical trials, including other types of vaccines such as the Oxford-AstraZeneca adenovirus-vectorized vaccine (ChAdOx1 nCoV-19) [5], CanSino’s adenovirus type-5 (Ad5)-vectorized COVID-19 vaccine [6], and Sinovac’s absorbed COVID-19 (inactivated) vaccine (ClinicalTrials.gov Identifier: NCT04456595). Among all the vaccines, a vast majority of them select the spike glycoprotein as their primary target.

The SARS-CoV-2 Spike (S) protein is a promising vaccine target, and many clinical studies reported anti-S protein neutralizing antibodies in COVID-19 recovered patients [7]. After the SARS outbreak in 2003 [8], clinical studies reported neutralizing antibodies targeting the SARS-CoV S protein [9,10], which was selected as the target of vaccine development [11,12]. Since SARS-CoV-2 shares a high sequence identity with SARS-CoV [13], and both viruses utilize
the attachment of the S protein to the human angiotensin-converting enzyme 2 (ACE2) receptor to invade host cells, neutralization of the SARS-CoV-2 S protein could induce protection in COVID-19 vaccine development [14]. Many computational studies utilizing reverse vaccinology and immuno-informatics reported the S protein to be a promising vaccine antigen [15–17], and clinical studies identified anti-S protein neutralizing antibodies in patients that have recovered from COVID-19 [18–20]. The cryo-EM structures of the S protein [21] and the neutralizing antibodies that bind to the S protein [22,23] were determined. Besides neutralizing antibodies, studies have also shown the importance of the CD4 T cell response in the control of SARS-CoV-2 infection and possible pre-existing immunity in healthy individuals without exposure to SARS-CoV-2 [7,24,25]. Kalita et al. has proposed a multi-peptide subunit-based epitope vaccine that is comprised of B cell, helper T cell, and cytotoxic T cell epitopes [26]. Overall, successful vaccination is likely linked to a robust and long-term humoral response to the SARS-CoV-2 S protein, which could be further enhanced by the rational structural design of the protein.

Structural vaccinology has shown success in improving vaccine candidates’ immunogenicity through protein structural modification. The first proof-of-concept was achieved by fixing the conformation-dependent neutralization-sensitive epitopes on the fusion glycoprotein of the respiratory syncytial virus [27]. A similar strategy has been applied to SARS-CoV-2 to conformationally control the S protein’s receptor-binding domain (RBD) domain between the “up” and “down” configurations [28] and stabilizing the protein in its prefusion form [29] to induce immunogenicity. Besides the structure-based rational design approaches, directed evolution, which is classified as an irrational method, has also been widely used in diverse fields, such as enzyme engineering [29], protein-RNA interaction [30], and COVID-19 therapeutic strategies [31]. The directed evolution usually first applies random mutagenesis to generate a large pool of variants, followed by screening for candidates with the preferred properties using high-throughput strategies. The advantage of directed evolution is that it works well without structural information. However, once a high-quality protein structure can be obtained either from the experimental determination or computational prediction, the structure-based approach is more suitable as it can efficiently explore a much larger sequence/conformational space using computer programs, yielding a few potential candidates for further screening and validation, which is more time-, money-, and labor-saving compared to a typical directed evolution process.

In this study, we extended structural vaccinology to a structure-based computational design of the SARS-CoV-2 S protein. Briefly, we used a protein design approach, EvoDesign [32], to generate multiple stable S protein variants without perturbing the surface amino acids to maintain the same B cell epitope profile. Meanwhile, we introduced mutations to the residues buried inside the S protein so that more major histocompatibility complex (MHC)-II T cell epitopes would be added into the newly designed S protein to potentially induce a stronger immune response. Finally, we evaluated the computationally designed protein candidates and compared them to the native S protein.

2. Materials and methods

2.1. Computational redesign of SARS-CoV-2 S protein

Fig. 1 illustrates the workflow for redesigning the SARS-CoV-2 S protein to improve its immunogenic potential for vaccine design. The full-length structure model (1,273 amino acids for the S monomer) of SARS-CoV-2 S assembled by C-I-TASSER [33] was used as the template for fixed-backbone protein sequence design using EvoDesign [32]. Although the cryo-EM structure for SARS-CoV-2 S is available (PDB ID: 6VSB) [21], it contains a large number of missing residues, and therefore, the full-length C-I-TASSER model was used for S protein design instead. The C-I-TASSER model used the cryo-EM density map to assemble the individual domain models and to refine the structure. The model showed a high similarity to the cryo-EM structure with a TM-score [34] of 0.87 and root-mean-square deviation (RMSD) of 3.4 Å in the commonly aligned regions, indicating a good model quality. The residues in the S protein were categorized into three groups: core, surface, and intermediate [35], according to their solvent accessible surface area ratio (SASAr). Specifically, SASAr is defined as the ratio of the absolute SASA of a residue in the structure to the maximum area of the residue in the GXX state [36], where X is the residue of interest and G is a glycine residue. The most extended GXX conformation measures the maximum exposure degree of the residue X in the solvated environment taking into account the local protein backbone. The SASAr ratios were calculated using the ASA webserver (http://cib.cf.ocha.ac.jp/bitool/ASA/), where the maximum area of each of the 20 canonical amino acid residues is provided. The core and surface residues were defined by us as those with SASAr < 5% and > 25%, respectively, while the other residues were regarded as intermediate. Since the surface residues may be involved in the interactions with other proteins (e.g., the formation of the homotrimeric, S-ACE2 complex, and S-antibody interaction) and may partially constitute the B cell epitopes, these residues were excluded from design, and more rigorously, their side-chain conformations were kept constant as well. Additionally, the residues that may form B cell epitopes reported by Grifoni et al. [16] were also fixed. The remaining core residues were subjected to design, allowing amino acid substitution, whereas the intermediate residues were repacked with conformation substitution. Specifically, 243, 275, and 755 residues were designed, repacked, and fixed, respectively: a list of these residue positions is shown in Supplementary Table S1. The 243 designable core residues were also compared to the global S protein mutations (global frequencies > 0.001) recorded in the GISAID database (as of December 7, 2020) [37,38]. These residues were also evaluated for their intrinsic disorder predisposition based on the reported disorder regions in the DisProt database [39]. The corresponding Jensen-Shannon Divergence (JSD) scores (higher scores indicate greater conservation) of these core residues residing within the disordered regions were reported [40]. During protein design, the evolution term in EvoDesign was turned off as this term would introduce evolutionary constraints on the sequence simulation search, which were not needed for this design [41]; therefore, only the physical energy function, EvoEF2 [35], was used for design scoring to broaden sequence diversity and help to identify more candidates with increased immunogenicity. In previous studies, EvoEF2 has been appropriately utilized to model the binding interactions between the SARS-CoV-2 S-RBD and a large number of ACE2 orthologs to identify the zoonotic origin of this novel coronavirus [42] and to design multiple anti-SARS-CoV-2 peptide therapeutics [41]. We performed 20 independent design simulations and collected all the simulated sequence decoys. A total of 5,963,235 sequences were obtained, and the best-scoring sequence had stability energy of −4100.97 EvoEF2 energy unit (EEU). A set of 22,914 non-redundant sequences that were within a 100 EEU window of the lowest energy and had > 5% of the design residues mutated were retained for further analysis (Fig. 1). We also utilized another popular protein design software, Rosetta [43], to generate 1,000 low-energy S variants using the “flexsb” protocol due to lower computational efficiency. The same surface-intermediate-core criterion was applied to the Rosetta protein design process. The EvoDesign and Rosetta designs were then analyzed and compared to examine the advantages/limitations of EvoDesign designs.
The workflow for designing and screening immunogenicity-enhanced SARS-CoV-2 S proteins started by defining the full-length SARS-CoV-2 native S protein into surface, intermediate, and core residues. This information was then fed into EvoDesign to generate structurally stable designs that introduce mutations to the core residues while keeping the surface conformation unchanged. The output design candidates from EvoDesign were then evaluated based on their immunogenic potential. The top ten candidates were also compared and evaluated in comparison to the native S protein.
2.2. MHC-II T cell epitope prediction and epitope content score calculation

The full-length S protein sequence was divided into 15-mers with ten amino-acid overlaps. For each 15-mer, the T cell MHC-II promiscuous epitopes were predicted using NetMHCIIpan v3.2 [44]. In brief, the percentile ranks of an epitope binding to each of the seven MHC-II alleles (i.e., HLA-DRB1*03:01, HLA-DRB1*07:01, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB4*01:01, and HLA-DRB5*01:01) were calculated, where the percentile rank was generated by comparing the 15-mer predicted binding affinity to the MHC-II molecule against that of a large set of similarly sized peptides randomly selected from the SWISS-PROT database [45]. An epitope was considered a promiscuous epitope if the median percentile rank was ≤ 20% by binding the 15-mer to any of the seven MHC-II alleles [46].

The selection of these seven MHC-II alleles aimed to predict the dominant MHC-II T cell epitopes across different ethnicities and HLA polymorphisms [47]. The MHC-II promiscuous epitopes of the native SARS-CoV-2 S protein (QHD43416) predicted using this method were also validated and compared to the dominant T cell epitopes mapped by Grifoni et al. [16]. In brief, Grifoni et al. mapped the experimentally verified SARS-CoV T cell epitopes reported in the Immune Epitope Database (IEDB) database, which includes experimentally verified T cell MHC-II epitope data, to the SARS-CoV-2 S protein based on sequence homology and reported as the dominant T cell epitopes [46]. The epitope content score (ECS) for a full-length S protein was calculated as the average value of the median percentile ranks for all the 15-mers spanning the whole sequence.

2.3. Human epitope similarity and human similarity score calculation

The human proteome included 20,353 reviewed (Swiss-Prot) human proteins downloaded from Uniprot (as of July 1, 2020) [48]. A total of 261,908 human MHC-II T cell promiscuous epitopes were predicted, as described above. The human epitope similarity between a peptide of interest (e.g., a peptide of the S protein) and a human epitope was then calculated using a normalized peptide similarity metric proposed by Frankild et al. [49]. In brief, the un-normalized peptide similarity score, $A(x,y)$, was first determined by the BLOCKSM35 matrix [50] for all the positions between a target peptide (y) and a human epitope (x), which was subsequently normalized using the minimum and maximum similarity scores for the human epitope (Eq. (1)). The maximum and minimum similarity scores were determined from a range of similarity scores between a human epitope and all of its possible amino acid substitutions. Finally, the normalized similarity score of a 15-mer peptide was calculated by comparing it to all the predicted human MHC-II T cell promiscuous epitopes. The human similarity score (HSS) of the full-length S protein was calculated by averaging the human epitope similarity of all the 15-mers.

$$S(x,y) = \frac{A(x,y) - A_{\min}}{A_{\max} - A_{\min}}$$  

2.4. Pre-existing immunity evaluation of the designed proteins

The pre-existing immunity of the designed proteins was evaluated and compared to that of the native S protein of seven human coronaviruses (HCoVs) (i.e., SARS-CoV-2, SARS-CoV, MERS-CoV, HCoV-229E, HCoV-OC43, HCoV-NL63, and HCoV-HKU1). The sequences of the seven HCoV S proteins were downloaded from Uniprot [49] (Table S2), and the MHC-II T cell epitopes were predicted as described above. The conserved epitopes were determined by the IEDB epitope clustering tool [45] and aligned using SEAVIEW [51].

2.5. Foldability assessment of the designed proteins

Since EvoDesign only produces a panel of mutated sequences, it is important to examine if the designed sequences can fold into the desired structure that the native S protein adopts. To examine their foldability, we used C1-TASSER to model the structure of the designed sequences, where the structural similarity between the native and designed S proteins was assessed by TM-score [52]. The TM-score also recently developed protein structure prediction program, which constructs full-length structure folds by assembling fragments threaded from the PDB, under the guidance of deep neural-network learning-based contact maps [53,54]. The ecotdomain of the S homotrimeric and the functional domains including the N-terminal domain (NTD), receptor-binding domain (RBD), fusion peptide (FP), heptad repeat sequence 1 (HR1), and connector domain (CD) [28,55] were visualized using PyMOL [56]. Sequence logo plots for the top ten and worst ten S protein designs were also generated [57]. The multiple sequence alignment of the top four EvoDesign S protein candidates with balanced ECS and HSS were aligned to the native S protein using SEAVIEW [51].

2.6. Molecular dynamics (MD) simulation

The extracellular domain (amino acids 1-1146) of the trimeric wildtype S and the top design (i.e., Design-10705, see Results) was subjected to MD simulation using GROMACS [58] with the CHARMM36 force field [59]. The initial full-length (amino acids 1-1273) trimers were built by C1-TASSER and residues 1147-1273 were deleted; glycosylation was not considered during structure modeling and the simulation. In each simulation case, a dodecahedron box was constructed with a distance of 10 Å from the solute, and TIP3P [60] water molecules were filled into the box. The system was then neutralized by the addition of an appropriate number of Na+ or Cl-ions. After the system was assembled, energy minimization was carried out using steepest descent minimization with a maximum force of 10 kJ/mol. The system was then equilibrated at 300 K using 100 ps NVT simulations and 100 ps NPT simulations with position restraints (1000 kJ/mol) on the heavy atoms of the protein. After the two equilibration phases, the system was well-equilibrated at the desired temperature and pressure. Unconstrained production MD was then carried out at 300 K for 50 ns as suggested in similar MD simulation studies [61,62]. The LINCS [63] algorithm was used to restrain the hydrogen bonds. Non-bonded interactions were truncated at 12 Å, and the Particle Mesh Ewald [64] method was utilized for long-range electrostatic interactions. The velocity-rescaling thermostat [65] and Parrinello-Rahman barostat [66] were used to couple the temperature and pressure, respectively. About 5,000 snapshots were saved with a time interval of 10 ps and utilized for further analysis using the built-in GROMACS command-line tools.

3. Results

The epitope content score (ECS) and human similarity score (HSS) of the S proteins from seven HCoV strains (severe HCoV: SARS-CoV-2, SARS-CoV, and MERS-CoV; mild HCoV: HCoV-229, HCoV-HKU1, HCoV-NL63, and HCoV-OC43) were computed. The ECS for the severe HCoV S proteins (mean = 49.3, standard deviation (SD) = 24.7) was significantly different (p = 0.0016, Mann-Whitney) from that of the mild ones (mean = 45.8, SD = 24.5). In terms of HSS, the severe HCoV S proteins (mean = 0.640,
SD = 0.03) tended to be less self-like \((p = 0.097, \text{Mann-Whitney})\) than the mild ones \((\text{mean} = 0.642, \text{SD} = 0.03)\). Overall, it was shown that both ECS and HSS might be used as indicators of the immunogenic potential of the designed S proteins.

On the other hand, previous studies suggested the potential role of pre-existing immunity in fighting COVID-19 [7,24,25]. Therefore, the predicted MHC-II T cell promiscuous epitopes of the SARS-CoV-2 S protein were compared to those from the other six HCoVs. There were two SARS-CoV-2 predicted MHC-II T cell promiscuous epitopes, which were also present on all of the seven HCoV S proteins \(\text{(Fig. 2)}\) and could be potentially linked to pre-existing immunity. Therefore, the designs were subsequently filtered based on the availability of these pre-existing immunity-related epitopes \(\text{(Fig. 1)}\). In particular, the SARS-CoV-2 promiscuous epitope S816-D830 overlapped with the dominant B cell epitope F802-E819 reported by Grifoni et al. [16].

EvoDesign generated a total of 22,914 low-energy S protein designs, in which 243 core residues were subjected to substitution \(\text{(see Methods for details)}\). As SARS-CoV-2 has been reported to have substantial mutations in its genome [67], it is important to compare the EvoDesign mutations to the natural mutations \(\text{(global mutation frequency of > 0.001)}\) reported by GISAID \(\text{(Table S3)}\). There were two EvoDesign core residues \(\text{(D80 and S98)}\) that also had natural mutations, and these two core residues had different mutation rates in EvoDesign in comparison to the natural infection. Specifically, for the D80 core residue, the natural mutation

\[
\text{Table 1}
\]

| Design ID | PEC | REC | ECS | HSS | EE (EEU) | RMSD (Å) | TM-score | SI (%) |
|-----------|-----|-----|-----|-----|---------|----------|----------|--------|
| 10,705    | 40  | 31  | 48.78 | 0.6394 | -4051.21  | 3.45     | 0.931   | 94.9   |
| 10,763    | 40  | 31  | 48.80 | 0.6394 | -4051.04  | 3.06     | 0.944   | 95.0   |
| 12,865    | 40  | 31  | 48.76 | 0.6396 | -4044.99  | 3.14     | 0.939   | 95.0   |
| 19,356    | 41  | 30  | 48.44 | 0.6399 | -4020.14  | 3.12     | 0.929   | 94.7   |
| 20,348    | 38  | 30  | 48.99 | 0.6390 | -4014.74  | 3.33     | 0.929   | 95.4   |
| 20,467    | 38  | 30  | 48.97 | 0.6391 | -4014.10  | 4.32     | 0.901   | 95.4   |
| 20,671    | 37  | 28  | 48.83 | 0.6395 | -4013.03  | 3.36     | 0.940   | 94.7   |
| 22,676    | 36  | 28  | 48.37 | 0.6399 | -4001.70  | 3.35     | 0.919   | 95.0   |
| 22,769    | 38  | 28  | 48.51 | 0.6398 | -4001.11  | 3.27     | 0.917   | 95.0   |
| 22,869    | 38  | 28  | 48.55 | 0.6398 | -4000.23  | 3.24     | 0.919   | 94.7   |
| Native    | 32  | 28  | 49.61 | 0.6401 | -         | -        | -        | -      |

PEC: Promiscuous Epitope Count; REC: Recovered Epitope Count; ECS: Epitope Content Score; HSS: Human Similarity Score; EE: EvoDesign Energy \(\text{(in EEU, EvoEF2 energy unit)}\); RMSD: Root-Mean-Square Deviation; TM: TM-score; SI: Sequence identity.

* The number of predicted promiscuous epitopes in designs that overlap with those in the native S protein.

* The RMSD and TM-score compared to the C-I-TASSER model of the native S protein.
frequency (D80Y) was 0.005, but the EvoDesign mutations, D80N, D80A, and D80S, had frequencies of 0.149, 0.106, and 0.003, respectively. For S98, the natural mutation (S98F) frequency was 0.007, but EvoDesign mutations S98T and S98A had frequencies of 0.837 and 0.008, respectively. To further investigate whether the EvoDesign candidates’ mutations were related to the intrinsic disorder predisposition, the 243 core residues were aligned to the DisProt database [39] and the aligned residues’ sequence conservations were evaluated. Twenty core residues were annotated as intrinsically disordered in DisProt, but these residues showed relatively high levels of conservation, with JSD scores ranging from 0.76 to 0.85 (Table S4), in the top 10 designs.

Among the 22,914 designs with relatively low (favorable) stability energy, 19,063 candidates that contained the two pre-existing immunity-related epitopes were ranked based on ECS and HSS (Fig. 3A). Using the ECS and HSS of the native SARS-CoV-2 S as the cutoff, we obtained 301 candidates with better immunogenic potential (i.e., lower ECS and HSS) (Fig. 3B). Ten candidates with balanced ECS and HSS were selected and evaluated (Table 1, full-length sequences in Table S5). The EvoDesign energy

Fig. 4. The multiple sequence alignment of the top four designed S proteins in comparison to the native SARS-CoV-2 S protein. The four EvoDesign S proteins (Design-10705, 10763, 12865, and 19356) were selected based on their high structural similarity to the native S protein and promising immunogenic potential (in terms of promiscuous epitope count, ECS, and HSS scores). The solid red boxes highlight the core residues that were subjected to mutations by EvoDesign. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
and sequence identity of all designs were plotted, and the top 10
designs were highlighted (Fig S1). The S protein variants generated
by EvoDesign had consistently better ECS in comparison to the
Rosetta designs, although the latter had a better HSS score (Fig
S2). All 1,000 Rosetta designs had higher ECS (thus worse immuno-
genic potential) than the native S, whereas EvoDesign was able to
produce a few designs with both lower ECS and HSS, affirming Evo-
Design’s ability to design vaccine candidates with better T cell
promiscuous epitopes.

The multiple sequence alignment of the top four candidates
showed that 88 of the 243 core residues were mutated at least
once (Fig. 4). There were 32 core residues substituted to the same
amino acids (R34T, V62I, I100M, R102Q, C136T, V143T, Y145S,
E191V, T250A, Y279F, R328K, V341I, V350S, W353A, D420A,
Y423M, C432V, S438V, V512I, T523N, T599L, S673T, N777T,
S875A, T881I, L916Y, C1043A, F1052H, S1055A, C1241G, S1242G,
and C1248T) in all top four designs. Additionally, the top ten and
worst ten designs were also plotted to infer functionally important
mutations to enhance immunogenicity (Fig S3). Specifically, there
were 12 core residues in both the top-scoring and worst-scoring
designs that were substituted to the same amino acids in compar-
ison to the native S protein (V62I, C136T, V143T, Y145S, E191V,
R328K, V341I, D420A, Y423M, C432V, S438V, V512I, T523N, T599L,
S673T, N777T, S875A, T881I, L916Y, C1043A, F1052H, S1055A, C1241G,
and C1248T) in all top four designs. Additionally, the top ten and
worst ten designs were also plotted to infer functionally important
mutations to enhance immunogenicity (Fig S3). Specifically, there
were 12 core residues in both the top-scoring and worst-scoring
designs that were substituted to the same amino acids in compar-
ison to the native S protein (V62I, C136T, V143T, Y145S, E191V,
R328K, V341I, D420A, C432V, S438V, T599L, S1055A, and C1241G). In par-
ticular, two remained unmutated in the top-scoring designs but

Fig. 5. The 3D structures of C-I-TASSER S protein trimer (A), cryo-EM trimer (B), Design-10705 trimer (C), and Design-10705 monomer (D). The ectodomain of Design-10705
was modeled using C-I-TASSER. Both the homo-trimer and monomer of Design-10705 are rendered. The NTD, RBD, HR1, FP, and CD domains are also highlighted in the
Design-10705 monomer. The mutations introduced in Design-10705 are shown in red spheres. (For interpretation of the references to colour in this figure legend, the reader
is referred to the web version of this article.)

524
were mutated in the worst-scoring designs (Y265W and V267T), suggesting mutations of these two residues might result in reduced immunogenicity.

Design-10705 was overall the best candidate with high structural similarity to the native S protein and good immunogenic potential (in terms of promiscuous epitope count, ECS, and HSS scores) amongst the top ten candidates. The candidate Design-10705 had a 93.9% sequence identity to the native S protein with a TM-score of 0.931 and an RMSD of 3.45 Å to the C-I-TASSER model of the native S protein. The homo-trimer 3D structure of Design-10705 was visualized and compared to the S protein C-I-TASSER and cryo-EM structural models (Fig. 5). In terms of immunogenicity, it had the second-highest number of promiscuous epitopes. Table 2 shows the complete MHC-II T cell epitope profile of Design-10705. There were 32 predicted promiscuous epitopes in the native S protein (Table S6), and 31 of them were recovered in Design-10705. The two pre-existing immunity-related epitopes, V991-Q1005 and S816-D830, were both recovered in the new design. Besides these two epitopes, there were 19 epitopes identical to the native S protein epitopes, while ten epitopes had at least one mutation in Design-10705. Compared with the native S protein, the only missing MHC-II epitope in Design-10705 was V911-N926, which was predicted to have reduced binding affinity to HLA-DRB4*01:01 and HLA-DRB4*01:02. Critically, this design introduced nine new MHC-II T cell promiscuous epitopes, which could potentially induce a stronger immune response with minimal perturbation compared to the native S protein.

One concern is that the top design, Design-10705, might lose the desired structure and protein function due to reduced stability caused by redesigning the core regions. To examine this concern, 50-ns MD simulations were carried out to compare the stability and flexibility of Design-10705 and the native S. As shown in Fig. 6A, Design-10705 and the wildtype showed similar RMSD shifts for both the backbones and side-chains after convergence at about 30 ns. The root-mean-square fluctuation (RMSF) measurement showed that the two proteins exhibited similar fluctuating profiles and thus comparable flexibility (Fig. 6B). Moreover, the radius of gyration and solvent-accessible surface area (SASA) as a function of simulation time were also analyzed for the two proteins. Design-10705 showed a slightly smaller radius of gyration (Fig. 6C) and a smaller SASA than the wildtype S (Fig. 6D), indicating that Design-10705 had a slightly more compact structure. Taken together, Design-10705 is expected to be sufficiently stable with the desired biological function (e.g., increased immunogenicity).

### Table 2
The predicted promiscuous MHC-II T cell epitopes of Design-17050.

| Epitope Start | End | Median | Percentile Rank | Comment |
|--------------|-----|--------|-----------------|---------|
| VQLDLITGRQLS | 991 | 1005 | 17 | Pre-existing immunity-related epitopes |
| SFIEDLLFKVTLD | 816 | 830 | 16 | |
| VYYPDKVFRSSVILH | 36 | 50 | 11 | |
| KVFRSSVILHSTQDLF | 41 | 55 | 17 | |
| SLLIVNATNVKIVK | 116 | 130 | 6.5 | |
| EFYRSSANNTFHEY | 156 | 170 | 18 | |
| FROYSHKTPINVRD | 201 | 215 | 14 | |
| SVLYNSASFSTFKCY | 366 | 380 | 18 | |
| YLYRIFFRSKNLKPE | 451 | 465 | 5.7 | |
| SIAAYTMSLAGENSV | 691 | 705 | 4.7 | |
| YGSCFTQIQLRALTIG | 756 | 770 | 19 | |
| LLPNVTLADAGFK | 821 | 835 | 17 | |
| CAQKFNGLTVLPIL | 851 | 865 | 19 | |
| GAALQIPFAMQMAYR | 891 | 905 | 18 | |
| IPFAMQMYARFNGIG | 896 | 910 | 3.7 | |
| QMAYRFGCIGNCVQNV | 901 | 915 | 19 | |
| TLVQKLSLIFGAISS | 961 | 975 | 14 | |
| TVYTOQJRAEAIR | 1006 | 1020 | 20 | |
| QJRAEJAERASANLA | 1011 | 1025 | 12 | |
| AEIRASANLAATKMS | 1016 | 1030 | 7.9 | |
| REGIVPSGNTHFWVT | 1091 | 1105 | 9.4 | |
| LPFTPENITWFXAHIV | 56 | 70 | 7.1 | |
| VFYVKINICYPFIVSY | 191 | 205 | 13 | |
| IGNNITRMTRASS | 231 | 245 | 6.2 | |
| TRFMTIRASSSRYLA | 236 | 250 | 1.2 | |
| YVGCLQTPFTELKFN | 266 | 280 | 12 | |
| SNRVSPTETIVKRP | 316 | 330 | 14 | |
| IFNATRFASSYAANR | 341 | 355 | 13 | |
| BSASSYAANRRISN | 346 | 360 | 17 | |
| VLSFELLLHAPANC | 511 | 525 | 14 | |
| KLAQNFNSAICKLQ | 921 | 935 | 17 | |
| NITWFHARHVSNTNG | 61 | 75 | 20 | |
| FNDGVYFAAATLTNTM | 86 | 100 | 14 | |
| GKCQCFNLRVEVFYK | 181 | 195 | 13 | |
| LVIOLPGNTRFM | 236 | 240 | 20 | |
| GVIAIWNNLALAVK | 431 | 445 | 11 | |
| TDEMIAQYTAALLAG | 866 | 880 | 19 | |
| VNQFLAQALNTLVIQK | 951 | 965 | 19 | |
| GAISSVMDILILSRDL | 971 | 985 | 20 | |
| VFLHVNLVPMQXSNF | 1061 | 1075 | 16 | |

4. Discussion

The subunit, DNA, and mRNA vaccines are typically considered to be safer but often induce weaker immune responses than the
live-attenuated and inactivated vaccines. Although the addition of adjuvant or better vaccination strategies can compensate for the immunogenicity, the addition of new epitopes to the antigen provides an alternative way to induce stronger immune responses [68,69]. During the protein design process, we applied design constraints so that the surface conformation, and in particular, B cell epitopes of the designed S protein variants were unchanged. For the designed S proteins with at least 5% of the core residues mutated, the immunogenicity potential of these candidates was evaluated and was structurally compared to the native S protein. The top candidate (Design-10705) recovered 31 out of 32 MHC-II promiscuous epitopes, and the two pre-existing immunity-related epitopes (V991-Q1005 and S816-D830) were present in the design. In addition to the 31 recovered epitopes, Design-10705 also introduced nine new MHC-II promiscuous epitopes with the potential to induce stronger CD4 T cell response. MD analysis of Design-10705 and the native S protein showed that the two proteins shared similar stability and flexibility (Fig. 6). Overall, the newly designed S protein should preserve the native S protein’s structure and function with enhanced immunogenicity.

The addition of epitopes to decrease the immunogenicity has been applied to therapeutic proteins. King et al. disrupted the MHC-II T cell epitopes in GFP and Pseudomonas exotoxin A using the Rosetta protein design protocol [70,71]. The EpiSweep program was also applied to structurally redesign bacteriolytic enzyme lysostaphin as an anti-staphylococcal agent with reduced immunogenicity to the host [72,73]. In this study, a similar strategy, but to improve immunogenicity, was applied to redesign the SARS-CoV-2 S protein as an enhanced vaccine candidate; specifically, we aimed to increase immunogenicity by introducing more MHC-II T cell promiscuous epitopes to the protein without reducing the number of B cell epitopes.

The addition of epitopes to induce stronger immune responses has been previously applied to develop H7N9 vaccines. The H7N9 hemagglutinin (HA) vaccine-elicited non-neutralizing antibody responses in clinical trials [74,75]. Rudenko et al. reported that there were fewer CD4 T cell epitopes found in H7N9 HA in comparison to the seasonal H1 and H3 HA proteins [76]. Based on this finding, Wada et al. improved the H7N9 vaccine by introducing a known H3 immunogenic epitope to the H7 HA protein without perturbing its conformation, which resulted in an over 4-fold increase in the HA-binding antibody response [68]. However, the number of epitopes is not the only factor that influences protective immunity. Studies have reported that CD8 T cell epitopes might induce regulatory T cell responses [49,77], and pathogens adapted to include CD4 and CD8 epitopes with high similarity to human peptides as a means to suppress host immunity for its survival [78]. Therefore, we examined the significance of ECS and HSS in...
The context of mild versus severe forms of HCoV infection and then utilized these two scores to evaluate the designed S protein candidates.

The computational design of the SARS-CoV-2 S protein could be coupled with some other structural modifications for a more rational structure-based vaccine design. The present study aims to introduce new epitopes to the S protein while keeping the surface residues unchanged to minimize the structural change of the designed proteins, and according to the protein structure prediction results, the designed candidates were predicted to be structurally similar to the native S protein (Table 1 & Fig. 5). The structural modifications performed on the native S protein, such as stabilizing the protein in its prefusion form[79], or fixing the RBD in the “up” or “down” state [28], could still be applied to the final candidate in this study. The combination of these structural vaccination techniques into the current pipeline could further enhance the immunogenicity of the S protein as a vaccine target. However, a major limitation of the present study is the wet-lab experimental validation of the designed proteins. First, the newly designed protein sequences need to be folded properly with a structure comparable to that of the native S protein. Second, the capability of the newly added epitopes for binding MHC-II molecules and subsequently inducing immune responses needs to be validated. Finally, these candidates should be tested for their protective efficacy and safety in animal models.

Overall, this study presents a strategy to improve the immunogenicity and antigenicity of a vaccine candidate by manipulating the MHC-II T cell epitopes through computational protein design. In the current settings, the immunogenicity evaluation was carried out after the standard protein design simulations with EvoDesign. In the future, the assessment of the immunogenic potential could be incorporated into the protein design process so that the sequence decoy generated at each step will be guided by balancing both the protein stability and immunogenicity. Moreover, with proper prior knowledge of known epitopes (e.g., both MHC-I and MHC-II from the pathogen proteome), it is also possible to create a chimeric protein, which integrates epitopes from antigens other than the target protein.

Financial Disclosure Statement

YH is supported by the National Institute of Allergy and Infectious Diseases (AI081062). YZ is supported by the National Institute of Allergy and Infectious Diseases (AI134678), the National Institute of General Medical Sciences (GM136422, S10OD026825), and the National Science Foundation (IIS1901119, DBI2030790). The funders had no role in study design, data collection, and analysis, decision to publish, or preparation of the manuscript.

Author contributions

YZ and YH conceived and designed the project. EO and XH performed studies on protein design and analyses. RP participated in formed studies on protein design and analyses. EO and XH performed the computational design of the SARS-CoV-2 S protein. RP participated in the discussions. EO drafted the manuscript. All authors performed result interpretation, edited, and approved the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2020.12.039.

References

[1] World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. 2020 [cited 24 Dec 2020]. Available: https://covid19.who.int.

[2] Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtler A, Lockhart S, Perez JL, Pérez Marc C, Moreira ERD, Zerbini C, Bailey R, Swanson KA, Roychoudhury S, Koury K, Li P, Kalina WV, Cooper D, French RW, Hammitt LL, Tureci O, Nell H, Schaefer A, Unal T, Trensin DB, Mather D, Sormitzer PR, Salih JU, Jansen KV, Grubert HC. Safety and efficacy of the BNT162b2 mRNA covid-19 vaccine: interim analysis of the phase 2/3, randomised controlled trial. Lancet 2020;396(10249):476–83. https://doi.org/10.1016/S0140-6736(20)31604-4.

[3] Zheng XL, Zhu Z, Wang J, Wang Y, Wu W, Zhang Z, Jiang D, Li L, Li N, Yuan J, Chen JY, Li Y, Wang X, Sun Y, Xu Y, Wang Z, Wang X, Wang W, Chen W. Safety, tolerability, and immunogenicity of a recombinant adenovirus type-5 vectored COVID-19 vaccine: a dose-escalation, open-label, non-randomised, first-in-human trial. Lancet 2020;395(10232):1845–54.

[4] Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Modderbacher CR, Rawlings SA, Sutherland AL, Premkumar L, Jadi RS, Marrama D, de Silva AM, Frazier A,
Therapeutic Strategies for COVID-19. bioRxiv [Preprint] 2020. Available: https://doi.org/10.1101/2020.05.19.104042.

(32) Pearce R, Huang X, Setawaran D, Zhang Y. EvoDesign: designing protein-protein binding interfaces using computational protein design. Bioinformatics 2020;36(17):6711–2.

(33) Zhang C, Zheng W, Huang X, Bell EW, Zhu 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. J Proteome Res 2020;19(4):1531–60.

(34) Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the target-to-template score. Nucleic Acids Res 2004;32(11):3805–16.

(35) Huang X, Pearce R, Zhang Y. EvoF2: Accurate and fast energy function for computational protein design. Bioinformatics 2020;36:1135–42.

(36) Tian Y, Huang X, Zhu Y. Computational design of enzyme–ligand binding using a deep-learned energy function and deterministic sequence optimization algorithm. J Mol Model 2015;21:191.

(37) Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Ablarerter W, et al. Tracking changes in SARS-CoV-2 spike: evidence that D614G increases infectivity of the virus. Cell 2020;182(4):812–827.e19.

(38) Elbe S, Buckingham-Merrett G. Data, disease and diplomacy: GISAID’s innovative data platform and synergies with other global platforms. Nature 2020;583(7816):264–7.

(39) E. Ong, X. Huang, R. Pearce et al. Computational and Structural Biotechnology Journal 19 (2021) 518–529.

(40) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(41) Huang X, Pearce R, Zhang Y. EvoEF2: Accurate and fast energy function for analyzing and searching for evolutionary coevolutionary features with deep residual neural networks for contact-structure template quality. Proteins Struct Funct Genet 2004;57(4):702–10.

(42) Huang X, Pearce R, Zhang Y. EvoF2: Accurate and fast energy function for simulating liquid water. J Chem Phys 1983;79:2273–40.

(43) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(44) Huang X, Pearce R, Zhang Y. EvoEF2: Accurate and fast energy function for analyzing and searching for evolutionary coevolutionary features with deep residual neural networks for contact-structure template quality. Proteins Struct Funct Genet 2004;57(4):702–10.

(45) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(46) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(47) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(48) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(49) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(50) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(51) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(52) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(53) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(54) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(55) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(56) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(57) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(58) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(59) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.

(60) Huang X, Pearce R, Zhang Y. De novo design of protein peptides to block coronavirus S protein with high antibody affinity. Chem Sci 2020;11:1581–9.
Li Q, Huang X, Zhu Y. Evaluation of active designs of cephalosporin C acylase by molecular dynamics simulation and molecular docking. J Mol Model 2014;20:2314.

Xue J, Huang X, Zhu Y. Using molecular dynamics simulations to evaluate active designs of cephradine hydrolase by molecular mechanics/Poisson-Boltzmann surface area and molecular mechanics/generalized Born surface area methods. RSC Adv 2019;9(24):13868–77.

Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: a linear constraint solver for molecular simulations. J Comput Chem 1997;18(12):1463–72.

Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LC. A smooth particle mesh Ewald method. J Chem Phys 1995;103(19):8577–93.

Bussi G, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. J Chem Phys 2007;126(1):014101.

Parrinello M, Rahman A. Polymorphic transitions in single crystals: a new molecular dynamics method. J Appl Phys 1981;52(12):7182–90.

Padhi AK, Tripathi T. Can SARS-CoV-2 accumulate mutations in the S-protein to increase pathogenicity?. ACS Pharmacol Transl Sci 2020;3(5):1023–6.

Wada Y, Nithichanon A, Nobusawa E, Moise L, Martin WD, Yamamoto N, et al. A humanized mouse model identifies key amino acids for low immunogenicity of H7N9 vaccines. Sci Rep 2017;7(1).

Hewitt JS, Karuppannan AK, Tan S, Gauger P, Halbur PG, Gerber PF, De Groot AS, Moise L, Opriessing T. A prime-boost concept using a T-cell epitope-driven DNA vaccine followed by a whole virus vaccine effectively protected pigs in the pandemic H1N1 pig challenge model. Vaccine 2019;37(31):4302–9.

King C, Garza EN, Mazor R, Linehan JL, Pastan I, Pepper M, Baker D. Removing T-cell epitopes with computational protein design. Proc Natl Acad Sci 2014;111(23):8577–82.

Fleishman SJ, Leaver-Fay A, Corn JE, Strauch E-M, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G, Meiler J, Baker D, Uversky VN. Rosettascripts: a scripting language interface to the rosetta macromolecular modeling suite. PLoS ONE 2011;6(6):e20161.