

1 **Rational Design of SARS-CoV-2 Spike Glycoproteins To Increase**
2 **Immunogenicity By T Cell Epitope Engineering**

3

4 **Edison Ong^{1,*}, Xiaoqiang Huang^{1,*}, Robin Pearce¹, Yang Zhang^{1,2†}, Yongqun He^{1,3,†}**

5

6 **1** Department of Computational Medicine and Bioinformatics, University of Michigan, Ann
7 Arbor, MI 48109, USA

8 **2** Department of Biological Chemistry, University of Michigan, Ann Arbor, MI 48109, USA

9 **3** Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology,
10 University of Michigan, Ann Arbor, MI 48109, USA

11

12

13 * These authors contributed equally

14 † Corresponding authors: Yang Zhang (zhng@umich.edu); Yongqun He

15 (yongqunh@med.umich.edu)

16 Abstract

17 The current COVID-19 pandemic caused by SARS-CoV-2 has resulted in millions of
18 confirmed cases and thousands of deaths globally. Extensive efforts and progress have been
19 made to develop effective and safe vaccines against COVID-19. A primary target of these
20 vaccines is the SARS-CoV-2 spike (S) protein, and many studies utilized structural vaccinology
21 techniques to either stabilize the protein or fix the receptor-binding domain at certain states. In
22 this study, we extended an evolutionary protein design algorithm, EvoDesign, to create
23 thousands of stable S protein variants without perturbing the surface conformation and B cell
24 epitopes of the S protein. We then evaluated the mutated S protein candidates based on predicted
25 MHC-II T cell promiscuous epitopes as well as the epitopes' similarity to human peptides. The
26 presented strategy aims to improve the S protein's immunogenicity and antigenicity by inducing
27 stronger CD4 T cell response while maintaining the protein's native structure and function. The
28 top EvoDesign S protein candidate (Design-10705) recovered 31 out of 32 MHC-II T cell
29 promiscuous epitopes in the native S protein, in which two epitopes were present in all seven
30 human coronaviruses. This newly designed S protein also introduced nine new MHC-II T cell
31 promiscuous epitopes and showed high structural similarity to its native conformation. The
32 proposed structural vaccinology method provides an avenue to rationally design the antigen's
33 structure with increased immunogenicity, which could be applied to the rational design of new
34 COVID-19 vaccine candidates.

35 **Introduction**

36 The current Coronavirus Disease 2019 (COVID-19) pandemic caused by severe acute
37 respiratory syndrome coronavirus 2 (SARS-CoV-2) has resulted in over 18 million confirmed
38 cases and 702,642 deaths globally as of August 6 2020 according to the World Health
39 Organization [1]. Tremendous efforts have been made to develop effective and safe vaccines
40 against this viral infection. The Moderna mRNA-1273 induced vaccine-induced anti-SARS-
41 CoV-2 immune responses in all 45 participants of phase I clinical trial [2], and advanced to
42 phase III clinical trial in record time. On the other hand, the Inovio INO-4800 DNA vaccine not
43 only showed protection from the viral infection in rhesus macaques, but was also reported to
44 induce long-lasting memory [3]. In addition to these two vaccines, there are over a hundred
45 COVID-19 vaccines currently in clinical trials including other types of vaccines such as the
46 Oxford-AstraZeneca adenovirus-vectored vaccine (ChAdOx1 nCoV-19) [4], CanSino's
47 adenovirus type-5 (Ad5)-vectored COVID-19 vaccine [5], and Sinovac's absorbed COVID-19
48 (inactivated) vaccine (ClinicalTrials.gov Identifier: NCT04456595). Among all the vaccines, a
49 vast majority of them select the spike glycoprotein (S) as their primary target.

50 The SARS-CoV-2 S protein is a promising vaccine target and many clinical studies
51 reported anti-S protein neutralizing antibodies in COVID-19 recovered patients [6]. After the
52 SARS outbreak in 2003 [7], clinical studies reported neutralizing antibodies targeting the SARS-
53 CoV S protein [8,9], which was selected as the target of vaccine development [10,11]. Since
54 SARS-CoV-2 shares high sequence identity with SARS-CoV [12], it is presumed that
55 neutralization of the SARS-CoV-2 S protein could be an important correlate of protection in
56 COVID-19 vaccine development [13]. Many computational studies utilizing reverse vaccinology
57 and immuno-informatics reported the S protein to be a promising vaccine antigen [14–16], and
58 clinical studies identified anti-S protein neutralizing antibodies in COVID-19 recovered patients
59 [17–19]. The cryo-EM structure of the S protein [20] and the neutralizing antibodies binding to
60 the S protein [21,22] were determined. Besides neutralizing antibodies, studies have also shown
61 the importance of CD4 T cell response in the control of SARS-CoV-2 infection and possible pre-
62 existing immunity in healthy individuals without exposure to SARS-CoV-2 [6,23,24]. Overall,
63 successful vaccination is likely linked to a robust and long-term humoral response to the SARS-
64 CoV-2 S protein, which could be further enhanced by the rational structural design of the
65 protein.

66 Structural vaccinology has shown successes to improve vaccine candidates'
67 immunogenicity through protein structural modification. The first proof-of-concept was achieved
68 by fixing the conformation-dependent neutralization-sensitive epitopes on the fusion
69 glycoprotein of respiratory syncytial virus [25]. A similar strategy has been applied to SARS-
70 CoV-2 to conformationally control the S protein's receptor-binding domain (RBD) domain
71 between the "up" and "down" configurations to induce immunogenicity [26]. In this study, we
72 extended structural vaccinology to rationally design the SARS-CoV-2 S protein by generating
73 thousands of stable S protein variants without perturbing the surface conformation of the protein
74 to maintain the same B cell epitope profile. In the meantime, mutations were introduced to the
75 residues buried inside the S protein so that more MHC-II T cell epitopes would be added into the
76 newly designed S protein to potentially induce a stronger immune response. Finally, we
77 evaluated the computationally designed protein candidates and compared them to the native S
78 protein.

79

80 **Materials and methods**

81 **Computational redesign of SARS-CoV-2 S protein**

82 Fig 1 illustrates the workflow for redesigning the SARS-CoV-2 S protein to improve its
83 immunogenic potential toward vaccine design. The full-length structure model (1,273 amino acids
84 for an S monomer) of SARS-CoV-2 S assembled by C-I-TASSER [27] was used as the template
85 for fixed-backbone protein sequence design using EvoDesign [28]. Although the cryo-EM
86 structure for SARS-CoV-2 S is available (PDB ID: 6VSB) [20], it contains a large number of
87 missing residues, and therefore, the full-length C-I-TASSER model was used for S protein design
88 instead. The C-I-TASSER model of the S protein showed a high similarity to the cryo-EM structure
89 with a TM-score [29] of 0.87 and RMSD of 3.4 Å in the common aligned regions, indicating a
90 good model quality. The residues in the S protein were categorized into three groups: core, surface,
91 and intermediate [30], according to their solvent accessible surface area ratio (SASAr).
92 Specifically, SASAr is defined as the ratio of the absolute SASA of a residue in the structure to
93 the maximum area of the residue in the GXG state [31], where X is the residue of interest; the
94 SASAr ratios were calculated using the ASA web-server (<http://cib.cf.ocha.ac.jp/bitool/ASA/>).
95 The core and surface residues were defined as those with SASAr <5% and >25%, respectively,
96 while the other residues were regarded as intermediate. Since the surface residues may be involved

97 in the interactions with other proteins (e.g., the formation of the S homotrimer, S-ACE2 complex,
98 and S-antibody interaction) and may partially constitute the B cell epitopes, these residues were
99 excluded from design, and more rigorously, their side-chain conformations were kept constant as
100 well. Besides, the residues that may form B cell epitopes reported by Grifoni et al. [15] were also
101 fixed. The remaining core residues were subjected to design, allowing amino acid substitution,
102 whereas the intermediate residues were repacked with conformation substitution. Specifically, 243,
103 275, and 755 residues were designed, repacked, and fixed, respectively; a list of these residue
104 positions is shown in Supplementary Table S1.

105

106 During protein design, the evolution term in EvoDesign was turned off as this term would
107 introduce evolutionary constraints on the sequence simulation search which were not needed for
108 this design [32]; therefore, only the physical energy function, EvoEF2 [30], was used for design
109 scoring to broaden sequence diversity and help to identify more candidates with increased
110 immunogenicity. We performed 20 independent design simulations and collected all the simulated
111 sequence decoys. A total of 5,963,235 sequences were obtained, and the best-scoring sequence
112 had stability energy of -4100.97 EvoEF2 energy unit (EEU). A set of 22,914 non-redundant
113 sequences that were within a 100 EEU window of the lowest energy and had >5% of the design
114 residues mutated were retained for further analysis (Fig. 1).

115

116 **MHC-II T cell epitope prediction and epitope content score calculation**

117 The full-length S protein sequence was divided into 15-mers with 10 amino-acid overlaps. For
118 each 15-mer, the T cell MHC-II promiscuous epitopes were predicted using NetMHCIIpan v3.2
119 [33], and an epitope was counted if the median percentile rank was $\leq 20.0\%$ by binding the 15-
120 mer to any of the seven MHC-II alleles [34] (i.e., HLA-DRB1*03:01, HLA-DRB1*07:01, HLA-
121 DRB1*15:01, HLA-DRB3*01:01, HLA-DRB3*02:02, HLA-DRB4*01:01, and HLA-
122 DRB5*01:01). The selection of these seven MHC-II alleles aimed to predict the dominant MHC-
123 II T cell epitopes across different ethnicity and HLA polymorphism. The MHC-II promiscuous
124 epitopes of the native SARS-CoV-2 S protein (QHD43416) predicted using this method were also
125 validated and compared to the dominant T cell epitopes mapped by Grifoni et al. [15]. In brief,
126 Grifoni et al. mapped the experimentally verified SARS-CoV T cell epitopes reported in the IEDB
127 database to the SARS-CoV-2 S protein based on sequence homology and reported as the dominant

128 T cell epitopes. The epitope content score (ECS) for a full-length S protein was defined as the
129 average value of the median percentile ranks for all the 15-mers spanning the whole sequence.

130

131 **Human epitope similarity and human similarity score calculation**

132 The human proteome included 20,353 reviewed (Swiss-Prot) human proteins downloaded from
133 Uniprot (as of July 1, 2020) [35]. A total of 261,908 human MHC-II T cell promiscuous epitopes
134 were predicted, as described above. The human epitope similarity between a peptide of interest
135 (e.g., a peptide of the S protein) and a human epitope was then calculated using a normalized
136 peptide similarity metric proposed by Frankild et al. [36]. In brief, the un-normalized peptide
137 similarity score, $A(x, y)$, was first determined by the BLOSUM35 matrix [37] for all the positions
138 between a target peptide (y) and a human epitope (x), which was subsequently normalized using
139 the minimum and maximum similarity scores for the human epitope (Eq. 1). Finally, the maximum
140 normalized similarity score of a 15-mer peptide was calculated by comparing to all the predicted
141 human MHC-II T cell promiscuous epitopes. The human similarity score (HSS) of the full-length
142 S protein was calculated by averaging the human epitope similarity of all the 15-mers.

143

$$144 S(x, y) = \frac{A(x, y) - A_{min}^x}{A_{max}^x - A_{min}^x} \quad (1)$$

145

146 **Pre-existing immunity evaluation of the designed proteins**

147 The pre-existing immunity of the designed proteins was evaluated and compared to that of the
148 native S protein of seven human CoVs (i.e., SARS-CoV-2, SARS-CoV, MERS-CoV, HCoV-229E,
149 HCoV-OC43, HCoV-NL63, and HCoV-HKU1). The sequences of the seven HCoV S proteins
150 were downloaded from Uniprot [35] (Table S2), and the MHC-II T cell epitopes were predicted as
151 described above. The conserved epitopes were determined by the IEDB epitope clustering tool [38]
152 and aligned using SEAVIEW [39].

153

154 **Foldability assessment of the designed proteins**

155 Since EvoDesign only produces a panel of mutated sequences, it is important to examine if the
156 designed sequences can fold into the desired structure that the native S protein adopts. To examine
157 their foldability, we used C-I-TASSER to model the structure of the designed sequences, where

158 the structural similarity between the native and designed S proteins was assessed by TM-score
159 [40]. Here, C-I-TASSER is a recently developed protein structure prediction program, which
160 constructs full-length structure folds by assembling fragments threaded from the PDB, under the
161 guidance of deep neural-network learning-based contact maps [41,42]. The ectodomain of the S
162 homotrimers was visualized via PyMOL [43].

163

164 **Results**

165 The epitope content score (ECS) and human similarity score (HSS) of the S proteins from
166 seven HCoV strains (severe HCoV: SARS-CoV-2, SARS-CoV, and MERS-CoV; mild HCoV:
167 HCoV-229, HCoV-HKU1, HCoV-NL63, and HCoV-OC43) were computed. The ECS for the
168 severe HCoV S proteins was significantly different from that for the mild ones ($p = 0.0016$, Mann-
169 Whitney). In terms of HSS, the severe HCoV S proteins tended to be less self-like compared to
170 the mild ones ($p = 0.097$, Mann-Whitney). Overall, it was shown that both ECS and HSS might be
171 used as indicators of the immunogenic potential of the designed S proteins.

172 On the other hand, previous studies suggested the potential role of pre-existing immunity
173 in fighting COVID-19 [6,23,24]. Therefore, the predicted MHC-II T cell promiscuous epitopes of
174 the SARS-CoV-2 S protein were compared to those from the other six HCoVs. There were two
175 SARS-CoV-2 predicted MHC-II T cell promiscuous epitopes, which were also present on all of
176 the seven HCoV S proteins (Fig 2), which could be potentially linked to pre-existing immunity.
177 Therefore, the designs were subsequently filtered based on the availability of these pre-existing
178 immunity-related epitopes (Fig 1). In particular, the SARS-CoV-2 promiscuous epitope S816-
179 D830 overlapped with the dominant B cell epitope F802-E819 reported by Grifoni et al. [15].

180 Among the 22,914 designs with relatively low stability energy, 19,063 candidates that
181 contained the two pre-existing immunity-related epitopes were ranked based on ECS and HSS (Fig
182 3A). Using the ECS and HSS of the native SARS-CoV-2 S as the cutoff, we obtained 301
183 candidates with a better immunogenic potential (i.e., lower ECS and HSS) (Fig 3B). Ten
184 candidates with balanced ECS and HSS were selected and evaluated (Table 1, full-length
185 sequences in Table S3).

186

187 Design-10705 was overall the best candidate with high structural similarity to the native S
188 protein and good immunogenic potential (in terms of promiscuous epitope count, ECS and HSS

189 scores) amongst the top ten candidates. The candidate Design-10705 had a 93.9% sequence
190 identity to the native S protein with TM-score (0.931) and RMSD (3.45 Å) to the C-I-TASSER
191 model of the native S protein. The homo-trimer 3D structure of Design-10705 was visualized and
192 compared to the S protein C-I-TASSER and cryo-EM structural models (Fig 4). In terms of
193 immunogenicity, it had the second-highest number of promiscuous epitopes. Table 2 showed the
194 complete MHC-II T cell epitope profile of Design-10705. There were 32 predicted promiscuous
195 epitopes in the native S protein (Table S4), and 31 of them were recovered in Design-10705. The
196 two pre-existing immunity-related epitopes, V991-Q1005 and S816-D830, were both recovered in
197 the new design. Besides these two epitopes, there were 19 epitopes identical to the native S protein
198 epitopes, while 10 epitopes had at least one mutation in Design-10705. Compared with the native
199 S protein, the only missing MHC-II epitope in design 10705 was V911-N926, which was predicted
200 to have reduced binding affinity to HLA-DRB1*03:01 and HLA-DRB4*01:01. Critically, this
201 design introduced nine new MHC-II T cell promiscuous epitopes, which could potentially induce
202 a stronger immune response with minimal perturbation compared with the native S protein.

203 **Discussion**

204 The subunit, DNA, and mRNA vaccines are typically considered to be safer but often
205 induce weaker immune responses than the live-attenuated and inactivated vaccines. Although the
206 addition of adjuvant or better vaccination strategies can compensate for the immunogenicity, the
207 addition of new epitopes to the antigen provides an alternative way to induce stronger immune
208 responses [44,45]. During the protein design process, we applied design constraints so that the
209 surface conformation, and in particular, B cell epitopes of the designed S protein variants were
210 unchanged. For the designed S proteins with at least 5% of the core residues mutated, the
211 immunogenicity potential of these candidates was evaluated and was structurally compared to
212 the native S protein. The top candidate (Design-10705) recovered 31 out of 32 MHC-II
213 promiscuous epitopes, and, the two pre-existing immunity-related epitopes (V991-Q1005 and
214 S816-D830) were present in the design. In addition to the 31 recovered epitopes, Design-10705
215 also introduced nine new MHC-II promiscuous epitopes with the potential to induce stronger
216 CD4 T cell response.

217 The concept of manipulating epitopes to decrease the immunogenicity has been applied
218 to therapeutic proteins. King et al. disrupted the MHC-II T cell epitopes in GFP and
219 *Pseudomonas* exotoxin A using the Rosetta protein design protocol [46,47]. The EpiSweep

220 program was also applied to structurally redesign bacteriolytic enzyme lysostaphin as an anti-
221 staphylococcal agent with reduced immunogenicity to the host [48,49]. In this study, a similar
222 strategy, but to improve immunogenicity, was applied to redesign the SARS-CoV-2 S protein as
223 an enhanced vaccine candidate; specifically, we aimed to increase immunogenicity by
224 introducing more MHC-II T cell promiscuous epitopes to the protein without reducing the
225 number of B cell epitopes.

226 The addition of epitopes to induce stronger immune responses has been previously
227 applied to develop H7N9 vaccines. The H7N9 hemagglutinin (HA) vaccine elicits non-
228 neutralizing antibody responses in clinical trials [50,51]. Rudenko et al. reported that there were
229 fewer CD4 T cell epitopes found in H7N9 HA in comparison to the seasonal H1 and H3 HA
230 proteins [52]. Based on this finding, Wada et al. improved the H7N9 vaccine by introducing a
231 known H3 immunogenic epitope to the H7 HA protein without perturbing its conformation,
232 which resulted in an over 4-fold increase of HA-binding antibody response [44]. However, the
233 number of epitopes is not the only factor that influences the protective immunity. Studies have
234 reported that CD8 T cell epitopes might induce regulatory T cell responses [36,53], and
235 pathogens adapted to include CD4 and CD8 epitopes with high similarity to human peptides as a
236 means to suppress host immunity for its survival [54]. Therefore, we examined the significance
237 of ECS and HSS in the context of mild versus severe forms of HCoV infection and then utilized
238 these two scores to evaluate the designed S protein candidates.

239 The computational design of the SARS-CoV-2 S protein could be coupled with some
240 other structural modifications for a more rational structure-based vaccine design. The present
241 study aims to introduce new epitopes to the S protein while keeping the surface residues
242 unchanged to minimize the structural change of the designed proteins, and according to protein
243 structure prediction, the designed candidates were structurally similar to the native S protein
244 (Table 1 & Fig 4). The structural modifications performed on the native S protein, such as
245 stabilizing the protein in its prefusion form [55], or fixing the RBD in the “up” or “down” state,
246 could still be applied to the final candidate in this study. The combination of these structural
247 vaccinology technologies into the current pipeline could further enhance the immunogenicity of
248 the S protein as a vaccine target. However, a major limitation of the present study is the wet-lab
249 experimental validation of the designed proteins. First, the newly designed protein sequences
250 need to be folded properly with a structure comparable to that of the native S protein. Second,

251 the capability of the newly added epitopes for binding MHC-II molecules and subsequently
252 inducing immune responses need to be validated. Finally, these candidates should be tested for
253 their protectiveness and safety in animal models.

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

263

264 **References**

- 265 1. World Health Organization. WHO Coronavirus Disease (COVID-19) Dashboard. 2020
266 [cited 6 Aug 2020]. Available: <https://covid19.who.int/>
- 267 2. Jackson LA, Anderson EJ, Roushphael NG, Roberts PC, Makhene M, Coler RN, et al. An
268 mRNA Vaccine against SARS-CoV-2 — Preliminary Report. *N Engl J Med.* 2020;
269 NEJMoa2022483. doi:10.1056/NEJMoa2022483
- 270 3. Patel A, Walters J, Reuschel EL, Schultheis K, Parzych E, Gary EN, et al. Intradermal-
271 delivered DNA vaccine provides anamnestic protection in a rhesus macaque SARS-CoV-2
272 challenge model. *bioRxiv* [Preprint]. 2020 [cited 1 Aug 2020]. Available:
273 <https://www.biorxiv.org/content/10.1101/2020.07.28.225649v1>
- 274 4. Folegatti PM, Ewer KJ, Aley PK, Angus B, Becker S, Belij-Rammerstorfer S, et al. Safety
275 and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2: a
276 preliminary report of a phase 1/2, single-blind, randomised controlled trial. *Lancet.* 2020.
277 doi:10.1016/S0140-6736(20)31604-4
- 278 5. Zhu FC, Li YH, Guan XH, Hou LH, Wang WWJ, Li JX, et al. Safety, tolerability, and
279 immunogenicity of a recombinant adenovirus type-5 vectored COVID-19 vaccine: a dose-
280 escalation, open-label, non-randomised, first-in-human trial. *Lancet.* 2020;395: 1845–
281 1854.

282 6. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Rydzynski Moderbacher C, et al.
283 Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19
284 disease and unexposed individuals. *Cell*. 2020;181: 1489–1501.

285 7. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and
286 epidemiology of 2019 novel coronavirus: implications for virus origins and receptor
287 binding. *Lancet*. 2020;395: 565–574.

288 8. Temperton NJ, Chan PK, Simmons G, Zambon MC, Tedder RS, Takeuchi Y, et al.
289 Longitudinally profiling neutralizing antibody response to SARS coronavirus with
290 pseudotypes. *Emerg Infect Dis*. 2005;11: 411–416.

291 9. Chan JFW, Lau SKP, To KKW, Cheng VCC, Woo PCY, Yue KY. Middle East
292 Respiratory syndrome coronavirus: Another zoonotic betacoronavirus causing SARS-like
293 disease. *Clin Microbiol Rev*. 2015;28: 465–522.

294 10. Shim BS, Park SM, Quan JS, Jere D, Chu H, Song MK, et al. Intranasal immunization
295 with plasmid DNA encoding spike protein of SARS-coronavirus/polyethylenimine
296 nanoparticles elicits antigen-specific humoral and cellular immune responses. *BMC*
297 *Immunol*. 2010;11: 65.

298 11. Yang ZY, Kong WP, Huang Y, Roberts A, Murphy BR, Subbarao K, et al. A DNA
299 vaccine induces SARS coronavirus neutralization and protective immunity in mice.
300 *Nature*. 2004;428: 561–564.

301 12. Zhou P, Yang X Lou, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak
302 associated with a new coronavirus of probable bat origin. *Nature*. 2020;579: 270–273.

303 13. Tay MZ, Poh CM, Rénia L, MacAry PA, Ng LFP. The trinity of COVID-19: immunity,
304 inflammation and intervention. *Nat Rev Immunol*. 2020;20: 363–374.

305 14. Ong E, Wong MU, Huffman A, He Y. COVID-19 coronavirus vaccine design using
306 reverse vaccinology and machine learning. *Front Immunol*. 2020;11: 1581.

307 15. Grifoni A, Sidney J, Zhang Y, Scheuermann RH, Peters B, Sette A. A Sequence
308 Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune
309 Responses to SARS-CoV-2. *Cell Host Microbe*. 2020;27: 671-680.e2.

310 16. Enayatkhani M, Hasaniazad M, Faezi S, Guklani H, Davoodian P, Ahmadi N, et al.
311 Reverse vaccinology approach to design a novel multi-epitope vaccine candidate against
312 COVID-19: an in silico study. *J Biomol Struct Dyn*. 2020; 1–16.

313 17. Wu F, Wang A, Liu M, Wang Q, Chen J, Xia S, et al. Neutralizing Antibody Responses to
314 SARS-CoV-2 in a COVID-19 Recovered Patient Cohort and Their Implications. medRxiv
315 [Preprint]. 2020 [cited 1 Aug 2020]. doi:10.2139/ssrn.3566211

316 18. Ni L, Ye F, Cheng M-L, Feng Y, Deng Y-Q, Zhao H, et al. Detection of SARS-CoV-2-
317 specific humoral and cellular immunity in COVID-19 convalescent individuals.
318 *Immunity*. 2020;52: 971–977.

319 19. Cao Y, Su B, Guo X, Sun W, Deng Y, Bao L, et al. Potent Neutralizing Antibodies against
320 SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent
321 Patients' B Cells. *Cell*. 2020;182: 73–84.

322 20. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh C-L, Abiona O, et al. Cryo-EM
323 structure of the 2019-nCoV spike in the prefusion conformation. *Science*. 2020;367:
324 1260–1263.

325 21. Barnes CO, West AP, Huey-Tubman KE, Hoffmann MAG, Sharaf NG, Hoffman PR, et
326 al. Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common
327 Epitopes and Recurrent Features of Antibodies. *Cell*. 2020. doi:10.1016/j.cell.2020.06.025

328 22. Wrapp D, De Vlieger D, Corbett KS, Torres GM, Wang N, Van Breedam W, et al.
329 Structural Basis for Potent Neutralization of Betacoronaviruses by Single-Domain
330 Camelid Antibodies. *Cell*. 2020;181: 1004–1015.

331 23. Bert N Le, Tan AT, Kunasegaran K, Tham CYL, Hafezi M, Chia A, et al. SARS-CoV-2-
332 specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls.
333 *Nature*. 2020. doi:10.1038/s41586-020-2550-z

334 24. Braun J, Loyal L, Frentsche M, Wendisch D, Georg P, Kurth F, et al. SARS-CoV-2-
335 reactive T cells in healthy donors and patients with COVID-19. *Nature*. 2020.
336 doi:10.1038/s41586-020-2598-9

337 25. McLellan JS, Chen M, Joyce MG, Sastry M, Stewart-Jones GBE, Yang Y, et al. Structure-
338 Based Design of a Fusion Glycoprotein Vaccine for Respiratory Syncytial Virus. *Science*.
339 2013;342: 592–598.

340 26. Henderson R, Edwards RJ, Mansouri K, Janowska K, Stalls V, Gobeil SMC, et al.
341 Controlling the SARS-CoV-2 spike glycoprotein conformation. *Nat Struct Mol Biol*.
342 2020. doi:10.1038/s41594-020-0479-4

343 27. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. Protein Structure and

344 Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and
345 the Unique Similarity between Its Spike Protein Insertions and HIV-1. *J Proteome Res.*
346 2020;19: 1351–1360.

347 28. Pearce R, Huang X, Setiawan D, Zhang Y. *EvoDesign: Designing Protein–Protein*
348 *Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an*
349 *Optimized Physical Energy Function.* *J Mol Biol.* 2019;431: 2467–2476.

350 29. Zhang Y, Skolnick J. *TM-align: A protein structure alignment algorithm based on the*
351 *TM-score.* *Nucleic Acids Res.* 2005;33: 2302–2309.

352 30. Huang X, Pearce R, Zhang Y. *EvoEF2: Accurate and fast energy function for*
353 *computational protein design.* *Bioinformatics.* 2020;36: 1135–1142.

354 31. Tian Y, Huang X, Zhu Y. *Computational design of enzyme–ligand binding using a*
355 *combined energy function and deterministic sequence optimization algorithm.* *J Mol*
356 *Model.* 2015;21: 191.

357 32. Huang X, Pearce R, Zhang Y. *De novo design of protein peptides to block association of*
358 *the SARS-CoV-2 spike protein with human ACE2.* *Aging.* 2020;12: 11263–11276.

359 33. Jensen KK, Andreatta M, Marcatili P, Buus S, Greenbaum JA, Yan Z, et al. *Improved*
360 *methods for predicting peptide binding affinity to MHC class II molecules.* *Immunology.*
361 2018;154: 394–406.

362 34. Paul S, Lindestam Arlehamn CS, Scriba TJ, Dillon MBC, Oseroff C, Hinz D, et al.
363 *Development and validation of a broad scheme for prediction of HLA class II restricted T*
364 *cell epitopes.* *J Immunol Methods.* 2015;422: 28–34.

365 35. The UniProt Consortium. *The Universal Protein Resource (UniProt).* *Nucleic Acids Res.*
366 2008;36: D193-197.

367 36. Frankild S, de Boer RJ, Lund O, Nielsen M, Kesmir C. *Amino acid similarity accounts for*
368 *T cell cross-reactivity and for “holes” in the T cell repertoire.* *PLoS One.* 2008;3: e1831.

369 37. Henikoff S, Henikoff JG. *Amino acid substitution matrices from protein blocks.* *Proc Natl*
370 *Acad Sci U S A.* 1992;89: 10915–10919.

371 38. Dhanda SK, Mahajan S, Paul S, Yan Z, Kim H, Jespersen MC, et al. *IEDB-AR: immune*
372 *epitope database—analysis resource in 2019.* *Nucleic Acids Res.* 2019;47: W502–W506.

373 39. Gouy M, Guindon S, Gascuel O. *Sea view version 4: A multiplatform graphical user*
374 *interface for sequence alignment and phylogenetic tree building.* *Mol Biol Evol.* 2010;27:

375 221–224.

376 40. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure
377 template quality. *Proteins Struct Funct Genet.* 2004;57: 702.

378 41. Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y. Ensembling multiple raw coevolutionary
379 features with deep residual neural networks for contact-map prediction in CASP13.
380 *Proteins Struct Funct Bioinforma.* 2019;87: 1082–1091.

381 42. Li Y, Hu J, Zhang C, Yu DJ, Zhang Y. ResPRE: High-accuracy protein contact prediction
382 by coupling precision matrix with deep residual neural networks. *Bioinformatics.* 2019;35:
383 4647–4655.

384 43. Schrödinger L. The PyMol Molecular Graphics System, Version~1.8. 2015 [cited 15 May
385 2020]. Available: <https://pymol.org>

386 44. Wada Y, Nithichanon A, Nobusawa E, Moise L, Martin WD, Yamamoto N, et al. A
387 humanized mouse model identifies key amino acids for low immunogenicity of H7N9
388 vaccines. *Sci Rep.* 2017;7: 1–11.

389 45. Hewitt JS, Karuppannan AK, Tan S, Gauger P, Halbur PG, Gerber PF, et al. A prime-
390 boost concept using a T-cell epitope-driven DNA vaccine followed by a whole virus
391 vaccine effectively protected pigs in the pandemic H1N1 pig challenge model. *Vaccine.*
392 2019;37: 4302–4309.

393 46. King C, Garza EN, Mazor R, Linehan JL, Pastan I, Pepper M, et al. Removing T-cell
394 epitopes with computational protein design. *Proc Natl Acad Sci.* 2014;111: 8577–8582.

395 47. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch EM, Khare SD, Koga N, et al.
396 Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling
397 suite. *PLoS One.* 2011;6: e20161.

398 48. Blazanovic K, Zhao H, Choi Y, Li W, Salvat RS, Osipovitch DC, et al. Structure-based
399 redesign of lysostaphin yields potent antistaphylococcal enzymes that evade immune cell
400 surveillance. *Mol Ther - Methods Clin Dev.* 2015;2: 15021.

401 49. Choi Y, Verma D, Griswold KE, Bailey-Kellogg C. EpiSweep: Computationally Driven
402 Reengineering of Therapeutic Proteins to Reduce Immunogenicity While Maintaining
403 Function. In: Samish I, editor. *Computational Protein Design.* New York, NY: Springer
404 New York; 2017. pp. 375–398.

405 50. Mulligan MJ, Bernstein DI, Winokur P, Rupp R, Anderson E, Roush N, et al.

406 Serological responses to an avian influenza A/H7N9 vaccine mixed at the point-of-use
407 with MF59 adjuvant a randomized clinical trial. *JAMA - J Am Med Assoc.* 2014;312:
408 1409–1419.

409 51. Guo L, Zhang X, Ren L, Yu X, Chen L, Zhou H, et al. Human antibody responses to avian
410 influenza A(H7N9) virus, 2013. *Emerg Infect Dis.* 2014;20: 192–200.

411 52. Rudenko L, Isakova-Sivak I, Naykin A, Kiseleva I, Stukova M, Erofeeva M, et al. H7N9
412 live attenuated influenza vaccine in healthy adults: A randomised, double-blind, placebo-
413 controlled, phase 1 trial. *Lancet Infect Dis.* 2016;16: 303–310.

414 53. Calis JJA, de Boer RJ, Keşmir C. Degenerate T-cell recognition of peptides on MHC
415 molecules creates large holes in the T-cell repertoire. *PLoS Comput Biol.* 2012;8:
416 e1002412.

417 54. Moise L, Gutierrez AH, Bailey-kellogg C, Terry F, Leng Q, Hady KMA, et al. The two-
418 faced T cell epitope: Examining the host-microbe interface with JanusMatrix. *Hum*
419 *Vaccin Immunother.* 2013;9: 1577–1586.

420 55. Bos R, Rutten L, Lubbe JEM van der, Bakkers MJG, Hardenberg G, Wegmann F, et al.
421 Ad26-vector based COVID-19 vaccine encoding a prefusion stabilized SARS-CoV-2
422 Spike immunogen induces potent humoral and cellular immune responses. *bioRxiv*
423 [Preprint]. 2020 [cited 1 Aug 2020]. Available:
424 <https://www.biorxiv.org/content/10.1101/2020.07.30.227470v1>

425

426

427 **Acknowledgments**

428 This work was supported in part by the National Institute of Allergy and Infectious Diseases
429 (AI081062, AI134678), the National Institute of General Medical Sciences (GM136422,
430 S10OD026825), and the National Science Foundation (IIS1901191, DBI2030790).

431

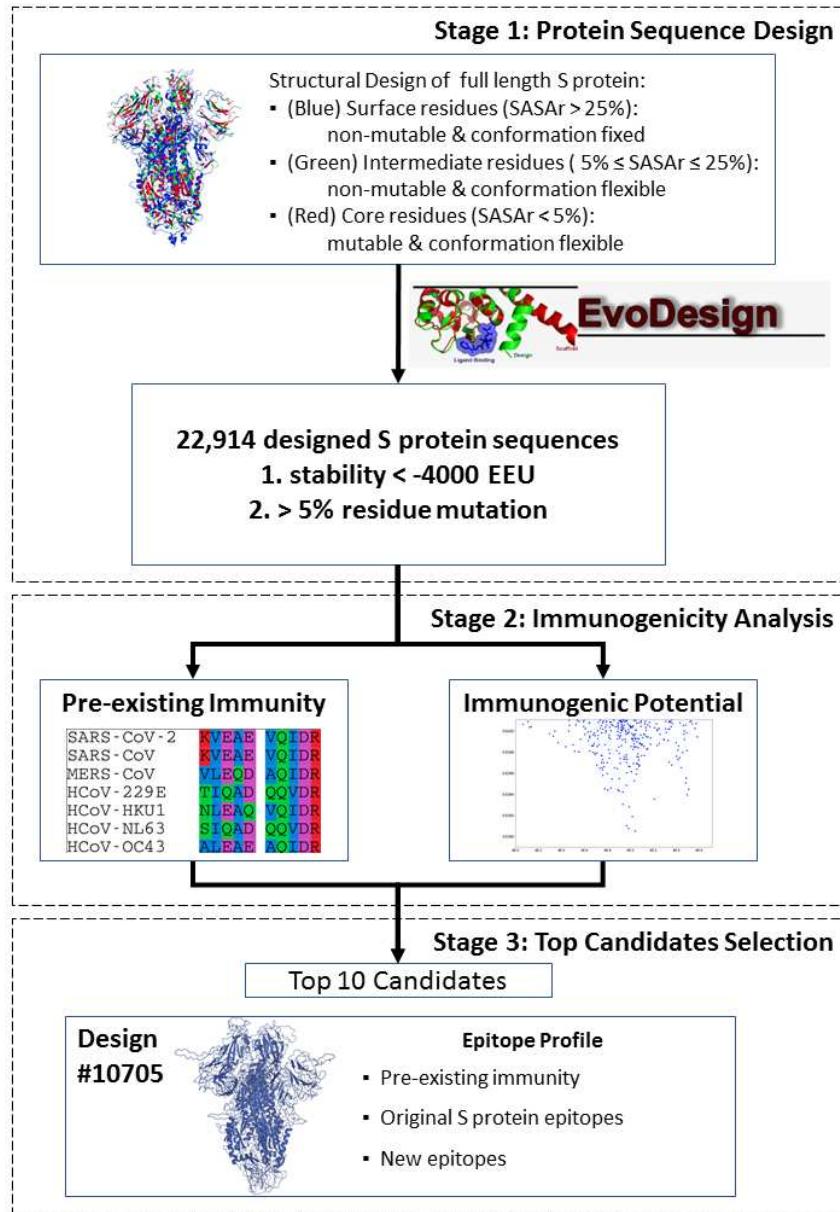
432 **Author contributions**

433 Y.Z. and Y. H. conceived and designed the project. E.O. and X.H. performed the studies on
434 protein sequence design and structural analyses. R.P. participated in the discussion. E.O. drafted
435 the manuscript. All authors performed result interpretation, edited, and approved the manuscript.

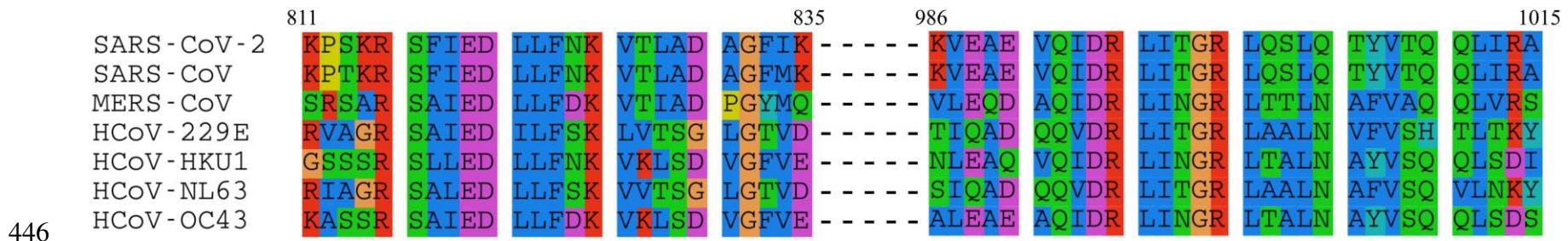
436

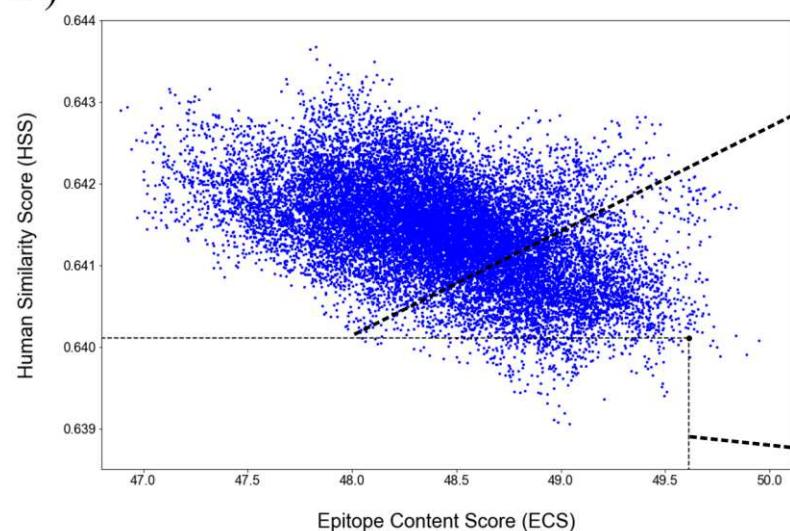
437 **Competing financial interests**

438 The authors declare no competing financial interests.

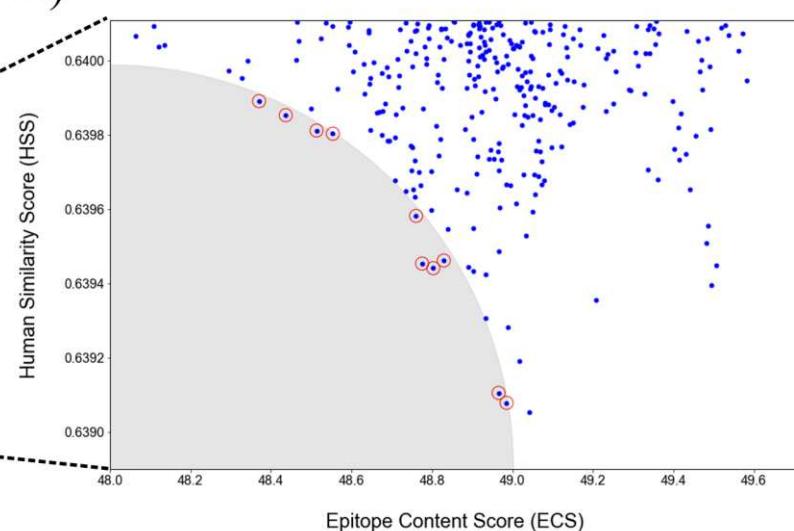


439 **Fig 1. The workflow of designing and screening immunogenicity-enhanced SARS-CoV-2 S**
440 **proteins.** The procedure started from defining the full-length SARS-CoV-2 native S protein into
441 surface, intermediate, and core residues. This information was then fed into EvoDesign to generate
442 structurally stable designs that introduce mutations to the core residues while keeping the surface
443 conformation unchanged. The output design candidates from EvoDesign were then evaluated
444 based on their immunogenic potential. The top ten candidates were also compared and evaluated
445 in comparison to the native S protein.



449
A)

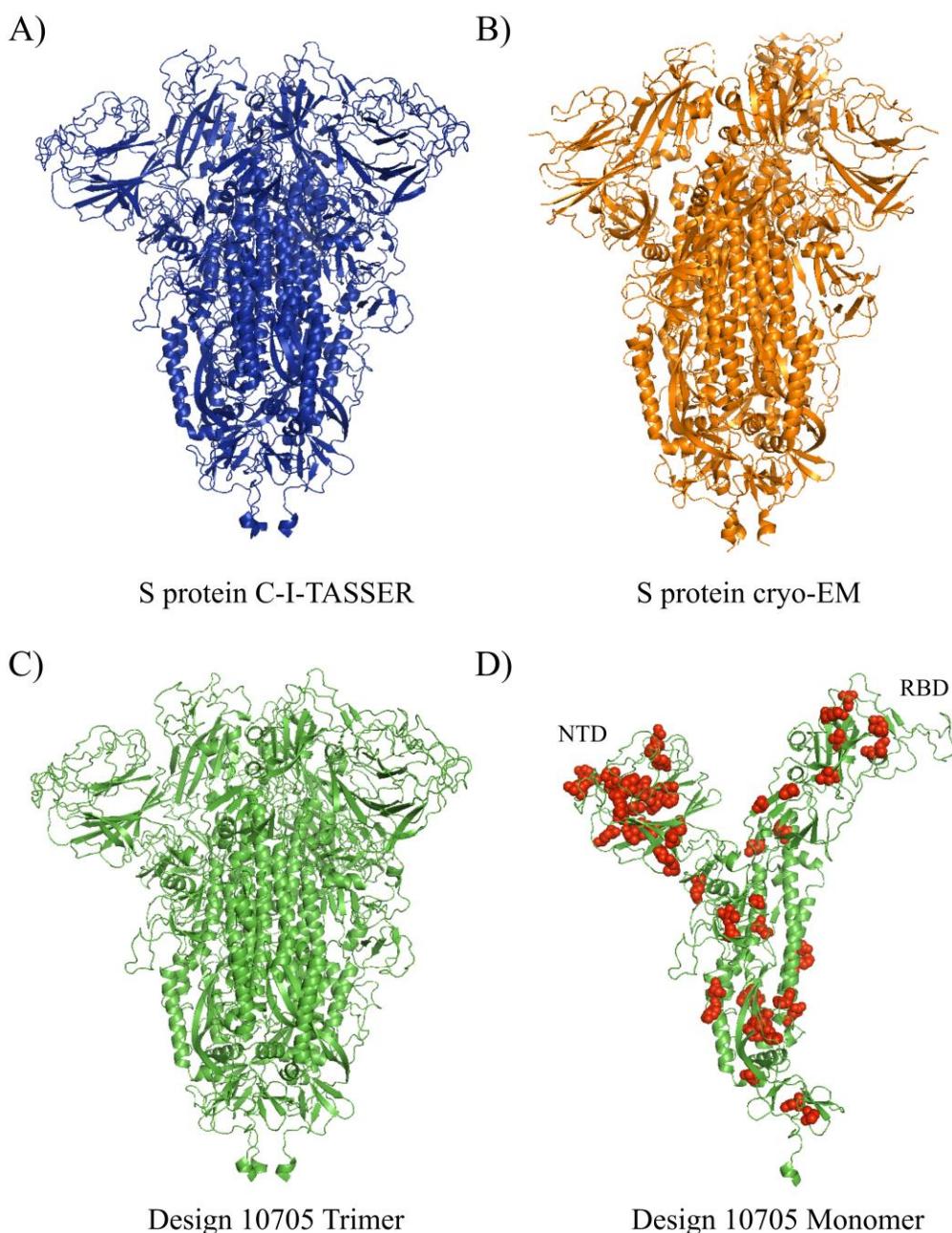
B)



450

451 **Fig 3. The epitope content score (ECS) and human similarity score (HSS) for designed S proteins.** (A) All 22,914 designs. Each
452 design is shown as a blue dot, whereas the native SARS-CoV-2 S was plotted as a black dot. The dashed-line box defines the 301
453 candidates with both lower ECS and HSS scores than the native. (B) The shaded area contains the top ten candidates with balanced
ECS and HSS scores.

454



455

456 **Fig 4. The 3D structures of A) C-I-TASSER S protein trimer, B) cryo-EM trimer, C)**
457 **Design-10705 trimer, and D) Design-10705 monomer.** The ectodomain of Design-10705 was
458 modeled using C-I-TASSER. Both the homo-trimer and monomer of Design-10705 were
459 rendered. The mutations introduced in Design-10705 are shown in red spheres.

460 **Table 1.** Summary of the features for the top 10 designs. The table is ranked based on the
461 designs' free energy scores (from low to high) except the native S protein.

Design ID	PEC	REC ^a	ECS	HSS	FE (EEU)	RMSD (Å) ^b	TM-score ^b	SI (%)
10705	40	31	48.78	0.6394	-4051.21	3.45	0.931	93.9
10763	40	31	48.80	0.6394	-4051.04	3.06	0.944	92
12865	40	31	48.76	0.6396	-4044.99	3.14	0.939	91.9
19356	41	30	48.44	0.6399	-4020.14	3.12	0.929	90.9
20348	38	30	48.99	0.6390	-4014.74	3.33	0.929	94
20467	38	30	48.97	0.6391	-4014.10	4.32	0.901	92
20671	37	28	48.83	0.6395	-4013.03	3.36	0.94	94.7
22676	36	28	48.37	0.6399	-4001.70	3.35	0.939	93.8
22769	38	28	48.51	0.6398	-4001.11	3.27	0.937	93
22869	38	28	48.55	0.6398	-4000.23	3.24	0.919	90.3
Native	32	--	49.61	0.6401	--	--	--	--

462 PEC: Promiscuous Epitope Count; REC: Recovered Epitope Count; ECS: Epitope Content Score; HSS: Human
463 Similarity Score; FE: Free Energy (EvoEF2 energy unit); RMSD: Root Mean Square Deviation; TM: TM-score; SI:
464 Sequence identity.

465 ^a: The number of predicted promiscuous epitopes in designs that overlap with those in the native S protein.

466 ^b: The RMSD and TM-score compared to the C-I-TASSER model of the native S protein.

467

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

Epitope	Start	End	Median Percentile Rank	Comment
VQLDRLITGRLQLSQ	991	1005	17	Pre-existing immunity-related epitopes
SFIEDLLFNKVTLAD	816	830	16	
VYYPDVKFRSSVLHS	36	50	11	
KVFRSSVLHSTQDLF	41	55	17	
SLLIVNNATNVVIKV	116	130	6.5	
EFRVYSSANNCTFEY	156	170	18	
FKIYSKHTPINLVRD	201	215	14	
SVLYNSASFSTFKCY	366	380	18	
YLYRLFRKSNLKPFE	451	465	5.7	
SIIAYTMSLGAENSV	691	705	4.7	
YGSFCTQLNRALTGI	756	770	19	
LLFNKVTLADAGFIK	821	835	17	Identical epitopes to native S protein
CAQKFNGLTVLPPLL	851	865	19	
GAALQIPFAMQMAYR	891	905	18	
IPFAMQMAYRFNGIG	896	910	3.7	
QMA YRFNGIGVTQNV	901	915	19	
TLVKQLSSNFGAIS	961	975	14	
TYVTQQLIRAAEIRA	1006	1020	20	
QLIRAAEIRASANLA	1011	1025	12	
AEIRASANLAATKMS	1016	1030	7.9	
REGVFVSNGTHWFVT	1091	1105	9.4	
LPFFSNITWFHAIHV	56	70	7.1	
VFVYKNIDGYFKIYS	191	205	13	
IGINITRFMTIRASS	231	245	6.2	
TRFMTIRASSRSYLA	236	250	1.2	
YVGYLQPRTFLLKFN	266	280	12	Mutated epitopes
SNFRVQPTETIVKFP	316	330	14	
IFNATRFASSYAANR	341	355	13	
RFASSYAANRKRI SN	346	360	17	
VILSFELLHAPANVC	511	525	14	
KLIANQFNSAIGKLQ	921	935	17	
NITWFHAIHVSGTNG	61	75	20	
FNDGVYFAATLKTNM	86	100	14	
GKQGNFKNLRVFVYK	181	195	13	
LVDLPIGINITRFMT	226	240	20	
GVVIAWNVNNLDAKV	431	445	11	New epitopes
TDEMIAQYTAALLAG	866	880	19	
VVNQLAQALNTLVKQ	951	965	19	
GAISSVMNDILSRD	971	985	20	
VFLHVNLVPAQEKNF	1061	1075	16	

470 **Supporting Information**

471 **S1 Table. SARS-CoV-2 S protein residues' core, intermediate, and surface definition for**
472 **EvoDesign.**

473 **S2 Table. Seven human coronavirus S proteins.**

474 **S3 Table. The full-length sequences of the top ten designs.**

475 **S4 Table. The predicted MHC-II T cell promiscuous epitopes of the native SARS-CoV-2 S**
476 **protein.**