Journal of proteome research

pubs.acs.org/jpr

Article

Identification and Characterization of Species-Specific Severe Acute Respiratory Syndrome Coronavirus 2 Physicochemical Properties

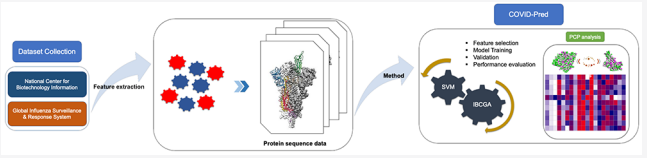
Srinivasulu Yerukala Sathipati and Shinn-Ying Ho*

Cite This: https://doi.org/10.1021/acs.jproteome.1c00156

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: There is an urgent need to elucidate the underlying mechanisms of coronavirus disease (COVID-19) so that vaccines and treatments can be devised. Severe acute respiratory syndrome coronavirus 2 has genetic similarity with bats and pangolin viruses, but a comprehensive understanding of the functions of its proteins at the amino acid sequence level is lacking. A total of 4320 sequences of human and nonhuman coronaviruses was retrieved from the Global Initiative on Sharing All Influenza Data and the National Center for Biotechnology Information. This work proposes an optimization method COVID-Pred with an efficient feature selection algorithm to classify the species-specific coronaviruses based on physicochemical properties (PCPs) of their sequences. COVID-Pred identified a set of 11 PCPs using a support vector machine and achieved 10-fold cross-validation and test accuracies of 99.53% and 97.80%, respectively. These findings could provide key insights into understanding the driving forces during the course of infection and assist in developing effective therapies.



KEYWORDS: *SARS-CoV-2 classification, machine learning, support vector machines, physicochemical properties*

INTRODUCTION

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is responsible for the COVID-19 pandemic that has spread around the globe since its first appearance in Wuhan, Hubei province of China, in early December.[1] As of 21 February 2021, the World Health Organization has reported 110.38 million confirmed cases and 2,446,008 deaths globally, becoming a major health concern. As of 18 February 2021, at least seven different vaccines across three platforms have been rolled out in countries.

Coronaviruses are enveloped single-stranded positive-sense RNA viruses that belong to a large family of viruses that constitute a subfamily *Orthocoronavirinae* in the family of *Coronaviridae*.[2] The genome sequence of SARS-CoV-2 is closely related to severe acute respiratory syndrome coronavirus (SARS-CoV) and bat coronaviruses. It shares 79.6% sequence identity to SARS-CoV, and it is 96% identical to bat coronavirus.[3] Human coronavirus (HCoV) genomes encode four major structural proteins including the spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins.[4] Each protein plays a significant role in the structure of the virus and in other aspects of the replication process.

The S glycoprotein of coronaviruses binds to appropriate receptors to facilitate viral entry into human host cells. SARS-CoV-2 uses the SARS-CoV receptor antigen converting enzyme 2 (ACE2) to enter into the host cell primed by TMPRSS2.[5] The S glycoprotein of SARS-CoV-2 and its entry into the host cell through ACE2 is well characterized.[6,7] The primary role of the N protein is to pack the viral genome into a nucleocapsid,[8] and it is considered to be a multifunctional protein in coronaviruses involved in the host cellular response to viral infection and replication.[9] The N protein is a key molecule in the egress and assembly of SARS-CoV, and transient expression of N is involved in the production of viruslike particles of coronaviruses.[10]

The M protein plays an important role in virus assembly and in the production of viral particles.[11] Homotypic interactions of M proteins are involved in envelope formation,[12] and they contribute to the core stability of coronaviruses.[13] Elongated and compacted M proteins are associated with the flexibility and density of S proteins.[11] The interaction between the M and E proteins is involved in envelope formation and budding of coronavirus particles.[11]

The coronavirus E protein is a minor component of the virus particles, but it plays an important role in virion assembly and virus host−cell interactions.[14] The absence of E protein in gastroenteritis coronaviruses blocks virus trafficking in the secretory pathway and prevents virus maturation.[15] Together, the structural and functional studies of the SARS-CoV-2 proteins can provide invaluable information about the binding potential of viruses to host cells, that is, information necessary for vaccine design.

Received: February 23, 2021

Each protein has distinct properties that allow it to perform its functions, and its interactions and dynamics depend on its physicochemical properties (PCPs). HCoV shares 89.1% nucleotide and 77.2% amino acid sequence similarity with some bat coronaviruses.[16] More importantly, the amino acid sequence comparison of the receptor-binding domain of the S proteins from HCoV and SARS-CoV showed that they shared only 73.8−74.9% sequence identities.[16] In addition, small changes in the amino acid sequence of the S protein are crucial for binding to its host. For instance, the bat SARS-like CoV strain cannot bind to human ACE2[17] due to minor amino acid differences from SARS-CoV. Therefore, knowing the HCoV protein PCPs and understanding the amino acid differences at the sequence level would be crucial for determining the mechanisms behind their species specificity and the functions of the HCoV proteins.

Extensive efforts are being made to eradicate the COVID-19 pandemic; the number of COVID-19 tests is rapidly increasing and it produces a huge dataset, which makes it difficult to derive the key elements that are essential for treatment. Artificial intelligence and machine learning are playing a critical role in COVID-19, especially by decreasing the workload of medical experts using computed tomography scans to detect COVID-19.[18] Machine learning techniques can broaden the screening process and identify potential antiviral agents based on their protein structures and DNA sequences to predict the drug binding sites of SARS-CoV-2.[19] Therefore, machine learning methods are ideal tools for analyzing large volumes of data and for identifying promising candidates for treating COVID-19.

In this study, we retrieved the protein sequences of 4320 coronaviruses from the Global Initiative on Sharing All Influenza Data (GISAID) and the National Center for Biotechnology Information (NCBI) databases. We constructed a dataset with 2225 human−host coronaviruses (HCoV) as positive samples and 2095 nonhuman−host coronaviruses (*n*HCoV) as negative samples. We used a support vector machine (SVM)-based optimization method called COVID-Pred to distinguish HCoV and *n*HCoV using their amino acid sequences. COVID-Pred uses an optimal feature selection algorithm called the inheritable bi-objective combinatorial genetic algorithm (IBCGA)[20] to select informative PCPs that are differentiated between HCoV and *n*HCoV. COVID-Pred identified 11 PCPs that are able to distinguish HCoV and *n*HCoV proteins. The objective of this study was to explore the PCPs and amino acid compositions that are specific to HCoV, which may be helpful in understanding how HCoV proteins function and may provide a guide for vaccine design.

MATERIALS AND METHODS

Dataset

The protein sequences of 2225 HCoV were retrieved from the GISAID database (https://www.gisaid.org) on June 3, 2020, and 2095 *n*HCoV protein sequences were retrieved from the NCBI database. The initial dataset thus consisted of the protein sequences of 4320 coronaviruses. Since each amino acid sequence is crucial for the binding of coronaviruses to their hosts, we reduced the sequence identity to 90%. After removal of redundancy and sequence uncertainties, the final dataset consisted of 141 HCoV structural proteins of coronaviruses as positive samples, whereas 163 *n*HCoV S protein sequences were negative samples. Furthermore, the dataset was divided into training and test sets in a ratio of 7:3. There were 213 sequences (HCoV and *n*HCoV) in the training set and 91 sequences (HCoV and *n*HCoV) in the test set. Additionally, we used seven S protein sequences of HCoV from the NCBI database after sequence identity reduction for an independent test. All of the data set information is summarized in Tables S1,S2 (Supplementary Data 2).

Physicochemical Properties

This study used 531 PCPs retrieved from the AAindex database developed by Kawashima and Kanehisa[21] as candidate features to construct COVID-Pred to distinguish species-specific coronavirus proteins. The original coronavirus' amino acid sequences were converted into numerical indices according to the 531 PCP values. The feature representation of the 531 PCPs is described as follows

a) Collect the HCoV and *n*HCoV protein sequences from the dataset.
b) Calculate the composition $f(a_i)$ of a protein for the i^{th} amino acid a_i of 20 amino acids to encode the protein sequence of variable length into a feature vector of length 531.
c) Calculate the feature value of the n^{th} physicochemical property, PCP(n), of a coronavirus protein, where n = 1, 2, ..., 531.

$$\mathrm{PCP}(n) = \sum_{i=1}^{20} f(a_i) \times \mathrm{PCP}_n(a_i) \tag{1}$$

where $\mathrm{PCP}_n(a_i)$ is the value of the a_i amino acid of the n^{th} physicochemical property.

Proposed COVID-Pred Method

To investigate the properties of the coronavirus proteins, we proposed the COVID-Pred method, which was customized using the SVM incorporating the optimal feature selection algorithm IBCGA.

Inheritable Bi-objective Combinatorial Genetic Algorithm

To construct the COVID-Pred method, IBCGA was used for feature selection. IBCGA is a well-known feature selection algorithm that has been used for solving biological problems such as cancer survival predictions,[22−24] protein function predictions,[25] and modeling gene regulatory networks.[26,27] IBCGA is an efficient global optimization technique with an intelligent evolutionary algorithm (IEA) to select a small set of informative features from a large pool of candidate features while optimizing the prediction performance. COVID-Pred utilized the SVM classifier for distinguishing the HCoV and *n*HCoV. In COVID-Pred, the SVM classifier was implemented in the LIBSVM package.[28] The radial basis function (RBF) kernel was used for the implementation of SVM in the LIBSVM package. The scoring function of the RBF kernel was computed in the feature space between the two data points, x_i and x_j. The RBF kernel function is defined as follows

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|)^2 \tag{2}$$

In IBCGA, the commonly used genetic algorithm (GA) terms such as gene and chromosome, represented as GA-gene and GA-chromosome, were used. The chromosome of IEA consists of m = 531 binary genes for selecting informative PCPs and two 4-bit GA-genes for encoding the parameters C and γ of SVM. The high performance of COVID-Pred arises from the simultaneous optimization of feature selection and

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 1. Performance Comparisons of COVID-Pred

	10-CV (%)	MCC	SN	SP	AUC
Naive Bayes	89.80	0.80	0.98	0.84	0.96
MLP	92.10	0.84	0.96	0.88	0.96
SMO	88.15	0.77	0.98	0.82	0.87
SGD	91.77	0.84	0.98	0.87	0.91
LMT	90.78	0.81	0.93	0.88	0.95
J48	92.43	0.84	0.93	0.91	0.93
decision tree	83.55	0.69	0.96	0.76	0.82
random forest	96.38	0.92	0.97	0.95	0.98
COVID-Pred	99.67	0.99	1.00	0.99	0.99
COVID-Pred (mean)	99.38 ± 0.11	0.98 ± 0.003	0.99 ± 0.003	0.99 ± 0.001	0.99 ± 0.001

fine-tuning of SVM using IBCGA. In COVID-Pred, numerical protein sequences encoded as 531 PCPs in the training dataset were used as the input. The IBCGA can simultaneously provide a set of solutions, X_r, where $r = r_{end}, r_{end+1}, ..., r_{start}$ in a single run. The feature selection algorithm IBCGA used can be described as follows

Step 1: (Initialization) Randomly generate an initial population of *N*pop individuals. In this work, *N*pop = 50, $r_{start} = 50$, $r_{end} = 10$, and $r = r_{start}$.

Step 2: (Evaluation) Evaluate the fitness value of all individuals using the fitness function, that is the prediction ACC in terms of 10-fold cross-validation.

Step 3: (Selection) Use a conventional method of tournament selection that selects the winner from two randomly selected individuals to generate a mating pool.

Step 4: (Crossover) Select two parents from the mating pool to perform an orthogonal array crossover operation of IEA.

Step 5: (Mutation) Apply a conventional bit mutation operator to parameter genes and a swap mutation to the binary genes for keeping r selected features. The best individual was not mutated for the elite strategy.

Step 6: (Termination test) If the stopping condition for obtaining the solution X_r is satisfied, output the best individual as the solution X_r. Otherwise, go to Step 2.

Step 7: (Inheritance) If $r > r_{end}$, randomly change one bit in the binary genes for each individual from 1 to 0; decrease the number r by one and go to Step 2. Otherwise, stop the algorithm.

Step 8: (Output) Obtain a set of m PCPs from the chromosome of the best solution X_m among the solutions X_r, where $r = r_{end}, r_{end} + 1, ..., r_{start}$.

Weka Classifiers

We used eight famous machine learning methods in Weka data mining software[29] to distinguish HCoV and *n*HCoV for performance comparison with COVID-Pred. They were Naive Bayes, multilayer perceptron (MLP), sequential minimal optimization (SMO), stochastic gradient descent (SGD), logistic model tree (LMT), J48, decision tree, and random forest. The classifier subset evaluator and the best first search were used for feature selection to design classifiers for distinguishing HCoV and *n*HCoV.

Evaluation Metrics

We evaluated the predictive performance of COVID-Pred using the following evaluation metrics: sensitivity (SN), specificity (SP), Matthews correlation coefficient (MCC), accuracy (ACC), and area under the ROC curve (AUC).

Amino Acid and Dipeptide Compositions

Amino acid composition (AAC) was measured for the HCoV and *n*HCoV. For the 20 amino acids denoted as $A_1...A_{20}$, the frequency of each amino acid (Af_i) was measured for the protein sequence length (L). AAC is represented as follows

$$\mathrm{AAC} = \left[\frac{Af_1}{L}, \frac{Af_2}{L}, ...\frac{Af_{20}}{L}\right] \quad (3)$$

Dipeptide composition (DPC) is defined as pairs of amino acids denoted as dipeptides, A_iA_j (i.e., AA, AC.... YY), and the frequency of occurrence of dipeptides is defined as df_{ij}. The DPC is computed as

$$\mathrm{DPC} = \left[\frac{df_{1,1}}{n}, \frac{df_{1,2}}{n}, ...\frac{df_{20,20}}{n}\right] \quad (4)$$

where $n = df_{1,1} + df_{2,2} + ... + df_{20,20}$.

RESULTS

Identification of SARS-CoV-2 Proteins

The objective of this study was to identify and analyze the PCPs that are specific to different coronavirus species and to explore the crucial driving forces that are involved in HCoV protein functions. For this purpose, 4320 protein sequences from HCoV and other organisms (*n*HCoV) in FASTA format were extracted. After preprocessing the initial dataset, the final dataset consisting of 141 HCoV and 163 *n*HCoV protein sequences was obtained from the GISAID and NCBI databases. The COVID-Pred method was established using the SVM incorporating the optimal feature selection algorithm IBCGA to identify the PCPs that could distinguish between HCoV and *n*HCoV. COVID-Pred selected 11 PCPs and achieved 10-fold cross-validation (10-CV) ACC, SN, SP, MCC, AUC, test ACC, and test AUCs of 99.53%, 1.00, 0.99, 0.99, 0.996, 97.80%, and 0.991, respectively. COVID-Pred obtained 100% (7/7) accuracy on an independent dataset consisting of seven HCoV S protein sequences. The COVID-Pred performance was evaluated using ROC curves as shown in Figure S1 (Supplementary Data 1).

Next, the prediction performance of COVID-Pred was compared with some machine learning methods of the Weka classifier using the full dataset ($n = 304$). We used the classifier subset evaluator and the best first search for the feature selection and selected 28 features to distinguish HCoV and *n*HCoV. Eight standard classifiers such as Naive Bayes, MLP, SMO, SGD, LMT, J48, decision tree, and random forest were used for the performance comparison. The Naive Bayes classifier achieved 10-CV ACC, MCC, SN, SP, and AUC of

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

89.80%, 0.80, 0.98, 0.84, and 0.96, respectively; MLP achieved 92.10%, 0.84, 0.96, 0.88, and 0.96, respectively; SMO achieved 88.15%, 0.77, 0.98, 0.82, and 0.87, respectively; SGD achieved 91.77%, 0.84, 0.98, 0.87, and 0.91, respectively; LMT achieved 90.78%, 0.81, 0.93, 0.88, and 0.95, respectively; J48 achieved 92.43%, 0.84, 0.93, 0.91, and 0.93, respectively; decision tree achieved 83.55%, 0.69, 0.96, 0.76, and 0.82, respectively; and random forest achieved 96.38%, 0.92, 0.97, 0.95, and 0.98, respectively. COVID-Pred obtained 10-CV ACC, MCC, SN, SP, and AUC of 99.67%, 0.99, 1.00, 0.99, and 0.99, respectively. The prediction performance of COVID-Pred was better than those of the other machine learning methods, as shown in Table 1. The COVID-Pred method achieved mean performance, 10-CV, MCC, SN, SP, and AUC of 99.38 ± 0.11, 0.98 ± 0.003, 0.99 ± 0.003, 0.99 ± 0.001, and 0.99 ± 0.001, respectively.

Informative PCP Characterization

We ranked the identified 11 PCPs based on their prediction performance using the main effect difference (MED). A larger MED score indicates a greater contribution toward prediction accuracy. The identified 11 PCPs and their corresponding ranks and MED scores are listed in Table 2. The identified 11 properties, including FAUJ880103, ONEK900101, PALJ810116, AURR980102, FAUJ880106, TANS770103, FASG760101, MONM990101, AURR980116, DAYM780201, and RICJ880117, were analyzed further to explore their roles in SARS-CoV-2 proteins.

Table 2. MED Analysis

rank	AAindex-ID	AAindex-desc	MED
1	FAUJ880103	normalized van der Waals volume	9.94
2	ONEK900101	delta G values for the peptides extrapolated to 0 M urea	9.33
3	PALJ810116	normalized frequency of turn in α/β class	8.05
4	AURR980102	normalized positional residue frequency at the helix termini N″′	6.83
5	FAUJ880106	STERIMOL maximum width of the side chain	6.56
6	TANS770103	normalized frequency of the extended structure	6.56
7	FASG760101	molecular weight	5.68
8	MONM990101	turn propensity scale for transmembrane helices	4.33
9	AURR980116	normalized positional residue frequency at the helix termini Cc	3.80
10	DAYM780201	relative mutability	1.91
11	RICJ880117	relative preference value at C″	0.50

Normalized van der Waals Volume. The top PCP based on the MED results was normalized by the van der Waals volume (FAUJ880103),[30] with a MED score of 9.94. Fauchère et al. measured the side chain parameters of the 20 amino acids. The relevance of the parameters for hydrophobicity and steric and electric properties of the amino acid side chains was assessed[30] in which the normalized van der Waals volume of the amino acid side chains was measured. There are different mechanisms involved in protein molecule interactions, including electrostatic forces, salvation forces, and van der Waals forces. Van der Waals forces act during interactions of proteins with other molecules.[31] Recently, stronger van der Waals interactions were found between SARS-CoV-2 and ACE2 compared to those between SARS-CoV and ACE2.[32] A molecular docking study on SARS-CoV-2 reported that van der Waals interactions play a major role in the binding process.[33] Yan et al. found that subtle amino acid changes improve the van der Waals interactions between SARS-CoV-2 and ACE2 and might determine the stronger interaction.[34] More amino acids that formed hydrogen bonds and van der Waals interactions were found at the SARS-CoV-2 interaction sites when compared to those at the SARS-CoV interaction sites. Wang et al. identified that the SARS-CoV-2-CTD binding interface has more amino acid residues forming van der Waals interactions than SARS-RBD that directly interacts with ACE2.[35] Stronger electrostatic and van der Waals interactions were observed between SARS-CoV-2 and ACE2 compared to those between SARS-CoV and ACE2.[36]

We thus measured the normalized van der Waals volumes for HCoV and *n*HCoV according to FAUJ880103.[30] We observed that the average normalized van der Waals volumes for HCoV were slightly higher than those for *n*HCoV. The mean normalized van der Waals volumes obtained for HCoV and *n*HCoV were 0.17 ± 0.10 and 0.16 ± 0.09, respectively. Among the 20 amino acids, larger van der Waals volume differences were observed between HCoV and *n*HCoV for L, K, N, R, and V. Additionally, we also observed slightly larger van der Waals volumes for the HCoV S proteins compared to the *n*HCoV S proteins. The amino acids R, Y, K, and P showed larger differences in van der Waals volumes between the HCoV and *n*HCoV proteins, as shown in Figure S2A (Supplementary Data 1).

Delta G Values for the Peptides Extrapolated to 0 M Urea. The conformational preferences of the amino acids influence the secondary and tertiary structures of proteins. Aydin et al. reported a 50% α-helical content in a designed recombinant SARS-CoV S2 domain fusion protein.[37] Subsequent conformational changes at the helices are critical to the fusion of viral and host membranes and the release of the viral genome into the host cells.[38] Karyn et al. measured the free energy difference ($\Delta\Delta G^0$) values of amino acids by substituting them in the guest sites of alpha helices.[39] We further calculated the $\Delta\Delta G^0$ values for HCoV and *n*HCoV according to ONEK900101.[39] The mean $\Delta\Delta G^0$ values did not show much difference between HCoV and *n*HCoV, but among the 20 amino acids, L, N, K, and E showed the largest differences in $\Delta\Delta G^0$ values between HCoV and *n*HCoV. A slight difference in the $\Delta\Delta G^0$ value was observed between the HCoV and *n*HCoV S proteins in which amino acids R, Y, V, and K showed a larger difference in $\Delta\Delta G^0$ value compared to the others.

Normalized Frequency of Turn in α/β Class. The property of PALJ810116 is described as "Normalized frequency of turn in α/β class."[40] Palau et al. calculated the conformational propensities of each amino acid for secondary structural alignments. The utilization of amino acids depends on the amount and topology of different secondary structures, and there are distinct preferences for α/β protein amino acids, such as I and V being the preferred amino acids in the α/β structures.[40] A circular dichroism spectroscopy study reported that a SARS-CoV-2 fusion peptide has an α-helical content.[41] Therefore, we measured the normalized propensities of α/β in HCoV and *n*HCoV. We observed a slight difference in the mean normalized α/β turns between HCoV and *n*HCoV with a mean normalized frequency of α/β turns of 0.048 ± 0.02 and 0.050 ± 0.03, respectively. Larger differences in the amino acids for this property between HCoV and *n*HCoV were observed for N, K, G, S, and Y. There was no mean difference

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

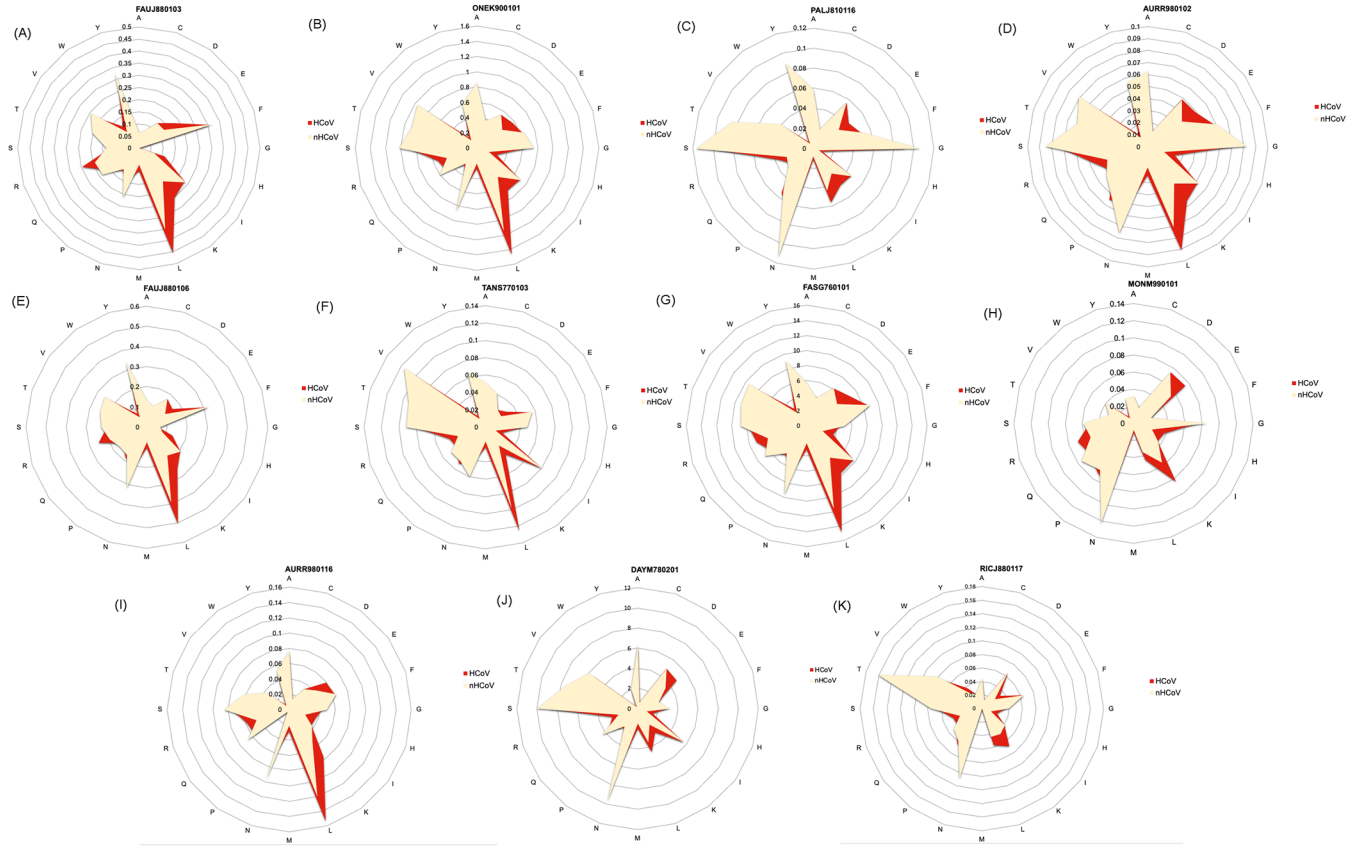


Figure 1. Comparison of PCPs between the HCoV and *n*HCoV proteins. (A) FAUJ880103, (B) ONEK900101, (C) PALJ810116, (D) AURR980102, (E) FAUJ880106, (F) TANS770103, (G) FASG760101, (H) MONM990101, (I) AURR980116, (J) DAYM780201, and (K) RICJ880117.

observed for this property between the HCoV and *n*HCoV S proteins, but amino acids Y, R, P, G, and S showed a difference in the normalized frequency of turns in α/β between the HCoV and *n*HCoV S proteins. This analysis indicates that amino acid propensities at α/β structures of SARS-CoV-2 might play an important role in the ACE2 binding process.

Normalized Positional Residue Frequency at the Helix Termini N″. The property of AURR980102 describes the normalized positional residue frequency at the helix termini N″. Aurora and Rose examined the role of helix capping in the secondary structures of proteins and identified seven distinct capping motifs at the helices C-terminus and N-terminus, where each motif exhibits a pattern of hydrogen bonds with hydrophobic interactions.[42] Various experiments demonstrated that the capping stabilizes the α-helices in proteins[43−45] and mutations of interacting residues in the capping motifs affect protein stability.[46] According to a previous study,[42] the normalized frequency of Pro is higher in N-terminal motifs, and also Pro functions as a hydrophobic residue. The CoV S glycoprotein is characterized by a complex of heptad-repeated regions (HR1 and HR2). The amide groups at the N-terminus of HR2 are capped by Asn, which interacts with the amide group via ordered water molecules, which may be one of the influential factors that stabilize the S glycol protein.[47] To examine the helix capping preferences, we measured the normalized frequencies of amino acids at helix capping in HCoV and *n*HCoV. We observed a slight difference in the mean normalized positional residue frequency at the helix termini N″ between HCoV and *n*HCoV. Although there was no large mean difference between HCoV and *n*HCoV for this property, we observed a larger difference in the normalized positional residue frequency at the helix termini N″ for the amino acids N, L, K, E, and G between HCoV and *n*HCoV as well as for R, P, K, S, and E between the S proteins of HCoV and *n*HCoV.

Relative Mutability. Dayhoff et al. calculated the relative mutability of amino acids, which indicates the probability of amino acid changes in a given small evolutionary interval.[48] The genome analysis of SARS-CoV-2 revealed that nearly 80% of the recurrent mutations produced nonsynonymous changes at the protein level, and these mutations are possible candidates for continuing adaptation of SARS-CoV-2 to its novel human host.[49] The genetic analysis of SARS-CoV-2 discovered various mutations and deletions in coding and noncoding regions.[50] The rapid mutations of SARS-CoV-2 play important roles in the virus spread. Hence, we measured the relative mutability of HCoV and *n*HCoV according to DAYM780201.[48] There was a slight difference in the mean relative mutability between HCoV and *n*HCoV, 3.77 ± 2.24 and 3.9 ± 2.85, respectively. A larger difference in the relative mutability of amino acids between HCoV and *n*HCoV was observed for N, E, S, L, and T, and differences in the relative mutability of amino acids between the S proteins of HCoV and *n*HCoV were observed for R, S, V, N, and P. Furthermore, the mutations in the S proteins were compared with the reference strain SARS-like bat virus, which falls under *n*HCoV (SLCoVZXC21/2015). We observed 203 mutations in the S protein (PDBID:6ACJ) compared with the reference sequence

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

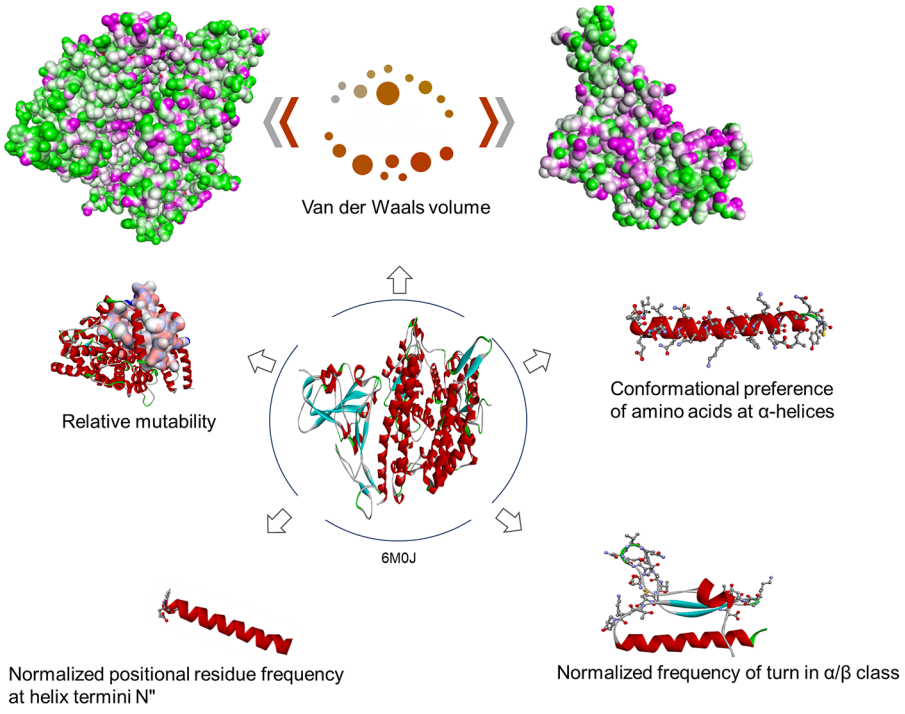
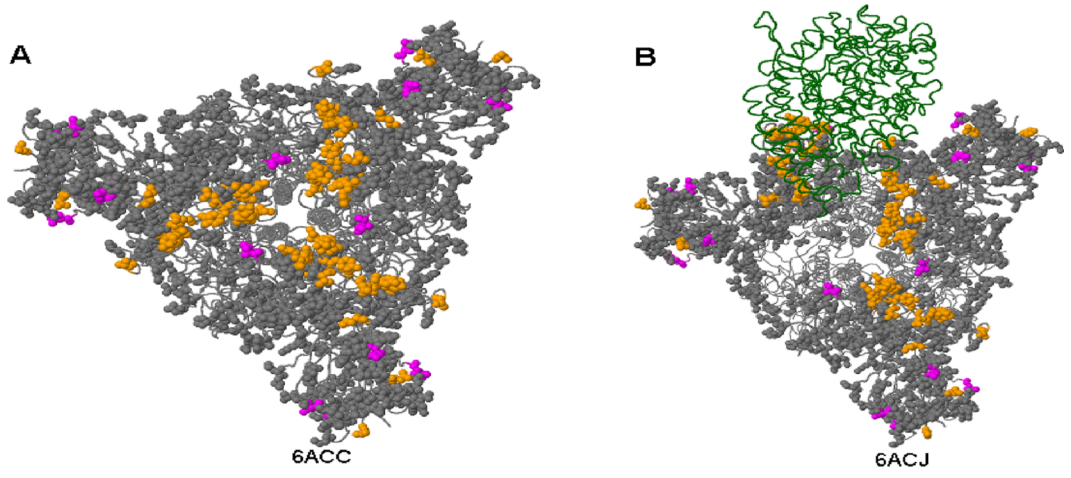


Figure 2. Graphical representation of the analyzed informative PCPs using the secondary structure of 6M0J as a model.



G20T P23Q N25P **N27A S32F** Q33T T41K I42V Y43F D46S T47V V49H L50S S51T
G53D Y54L Y59F S63T Y65F Y66H S67A L68I T69H T70V N71S N72G A73T **ins73N**
A74G T78F I82V D84P K86N I89V A93S H96K V100I **N111S** T112K S113T I126V
N131E D133Q **Y136N** Y139F S141G G142V **ins145K** T149S S151M I152E R153S
A156R F160S Y161A A162N K171Q S172P M174L L175M **N176D** I177L **S178E**
N180K G181Q L183N N185K T186N R193K V195I H198Y F205H V208I N211V
G213D T216Q L218F V220A Q222E E226D V229I S230G K235R R237Q T241A
I242L **ins244SYLTP P247del** M248del N250S N251S F256G S257A F261Y K266Q
M271L L301F S302T Q304E Q319E V330L H334G K335E P343A E349N T351K
K352R D355N I357V T361S F363L **T367A** S380T I383N S389N T394S L396V F399G
S400D V405I V412K T433S **A434N** K435N Q436L T438S **ins439VGGNY G439K**
H440N F442L S445L H446F **S448K T449S** K450N **L458I** S460T D461E
ins462YQAGSTPC E462I **ins466GFNCYF R466E** T467P **S469Q T470S D472G N474Q**
N476T V477N P478G L479V E480G A483P **T484Y** N495H L505K Q508N Q513K
K523T D530E S532N R534K Q536L S537P K543R A545I S546A F548T I549T
S551A S582N E583Q D595E T598V T599A A608T I611V A613S I614T T616S
S617N Q622R A634N H651Q(650) A653Q(650) **ins654NSPR(650,664)** S654T(650)
I655R(650) L656A(650) T659V(664) G660A(664) Q661S(664) K662Q(664)
A663S(664) V665I I677V A680S S691T V698I M699L A704T I719T

Figure 3. Visualization of the S glycoprotein with mutations. (A) Structure of the SARS-CoV S protein (PDB: 6acc, EM 3.6 Angstrom). (B) S glycoprotein (PDB: 6acj, EM 4.2 Angstrom) in complex with the host cell receptor ACE2 (green ribbon); mutations identified in the query sequences are shown as colored balls (based on the nearest residue if in the loop/termini region).

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

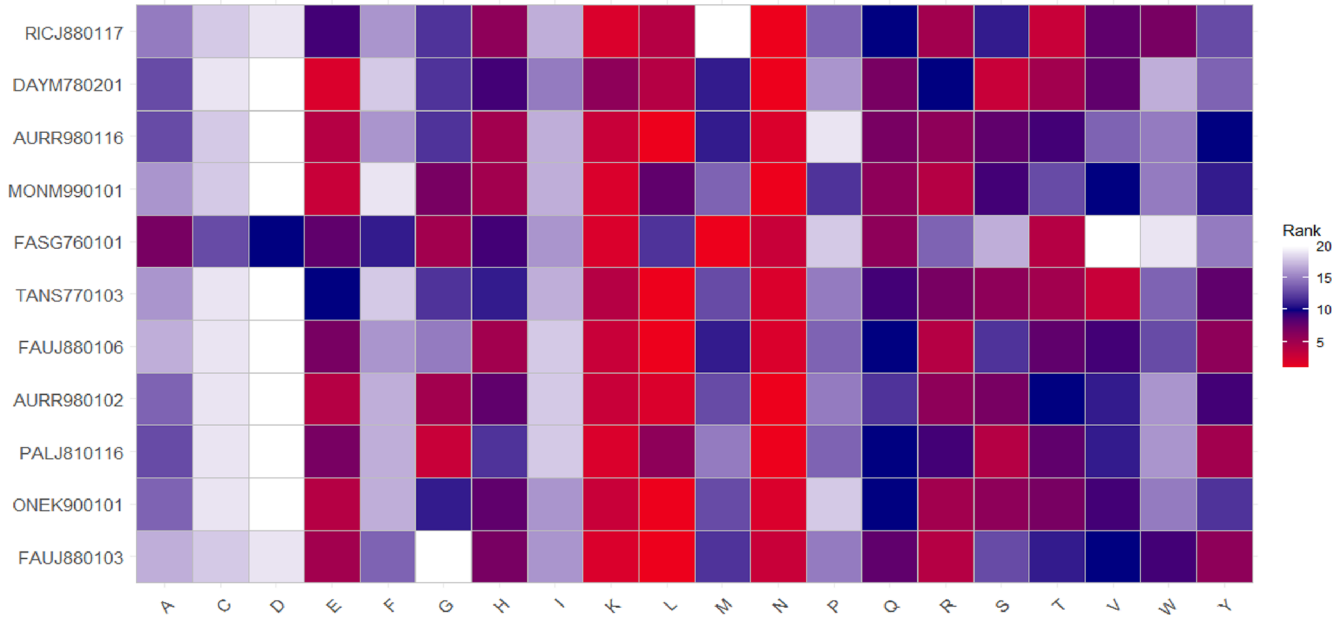


Figure 4. Normalized amino acid compositional preferences showing differences in the 11 PCPs between HCoV and *n*HCoV.

SARS-like/Bat/Nanjing/SL-CoVZXC21/2015 (PDBID:6ACC). A detailed list of the mutations between these two structures is given in Table S3 (Supplementary Data 3).

Furthermore, we performed a statistical analysis using t-test to identify the significant amino acids of the six PCPs across HCoV and *n*HCoV. The p <0.05 was considered as statistical significance in the analysis. A significant difference (p <0.005) in van der Waals volume between HCoV and *n*HCoV was observed for the amino acids K, L, N, R, and V. The amino acids L, N, K, and E showed a significant difference in $\Delta\Delta G^0$ values between HCoV and *n*HCoV. A significant difference in the amino acids for the normalized frequency of turn in α/β class between HCoV and *n*HCoV was observed for N, K, G, S, and Y. A significant difference in the normalized positional residue frequency at the helix termini N″ between HCoV and *n*HCoV was observed for the amino acids N, L, K, E, and G. A significant difference in the relative mutability of amino acids between HCoV and *n*HCoV was observed for the amino acids N, E, S, L, and T.

Additionally, the other six properties identified were the STERIMOL maximum width of the side chain (FAUJ880106), normalized frequency of the extended structure (TANS770103), molecular weight (FASG760101), turn propensity scale for transmembrane helices (MONM990101), normalized positional residue frequency at the helix termini Cc (AURR980116), and the relative preference value at C″ (RICJ880117). Their differences between HCoV and *n*HCoV are shown in Figure 1. The amino acid compositional preferences for the 11 PCPs between the S proteins of HCoV and *n*HCoV are shown in Figure S2 (Supplementary Data 1). Graphical representations of the five analyzed properties, FAUJ880103, ONEK900101, PALJ810116, AURR980102, and DAYM780201, are shown in Figure 2. The comparison of the mutations between the S protein and the reference strain SARS-like bat virus is shown in Figure 3. Furthermore, we ranked the amino acids based on their compositional preference differences between HCoV and *n*HCoV for the 11 PCPs. The amino acid rank is proportional to the compositional preference difference, meaning that the rank one amino acid has the greatest difference between HCoV and *n*HCoV. The amino acids that show compositional preference differences for the 11 PCPs between HCoV and *n*HCoV are shown in Figure 4. The amino acids that show compositional preference differences for the 11 PCPs between the S proteins of HCoV and *n*HCoV are shown in Figure S3 (Supplementary Data 1). The identified 11 PCPs and their amino acid compositional preferences are reported in Table S4 (Supplementary Data).

Analysis of Amino Acid and Dipeptide Compositions

Amino acid differences in different proteins could shed light on how SARS-CoV-2 is functionally and structurally different from humans and other organisms. Hence, the AAC differences were measured between HCoV and *n*HCoV. The maximum amino acid compositional differences between HCoV and *n*HCoV were obtained for L, K, and N with a ±2% composition difference and for E, H, R, M, Y, Q, G, V, S, and T with a ±1% composition difference, while the other amino acids did not show any differences, as shown in Figure 5A. The AAC differences for all of the amino acids are listed in Table 3. When we compared the S proteins of HCoV and *n*HCoV, the maximum AAC difference was observed for the amino acid R with a 2% composition difference and for P, K, G, V, N, and V with a ±1% difference, as shown in Figure S4 (Supplementary Data 1).

Dipeptides play an important role in folding and peptide binding. Therefore, DPCs were measured for the HCoV, *n*HCoV, and HCoV S proteins. The top five DPCs obtained for HCoV were LL, FL, LV, VL, and TL, while for *n*HCoV, they were LL, VN, NG, SV, and SL; for the HCoV S proteins, they were RR, VL, IA, SN, and SV. A heatmap showing the differences in the DPC of HCoV and *n*HCoV is shown in Figure 5B. These AAC and DPC differences may be important factors for functional and pathogenic divergence of SARS-CoV-2. Heatmaps showing the DPCs of HCoV and *n*HCoV are shown in Figure S5A,5B (Supplementary Data 1).

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

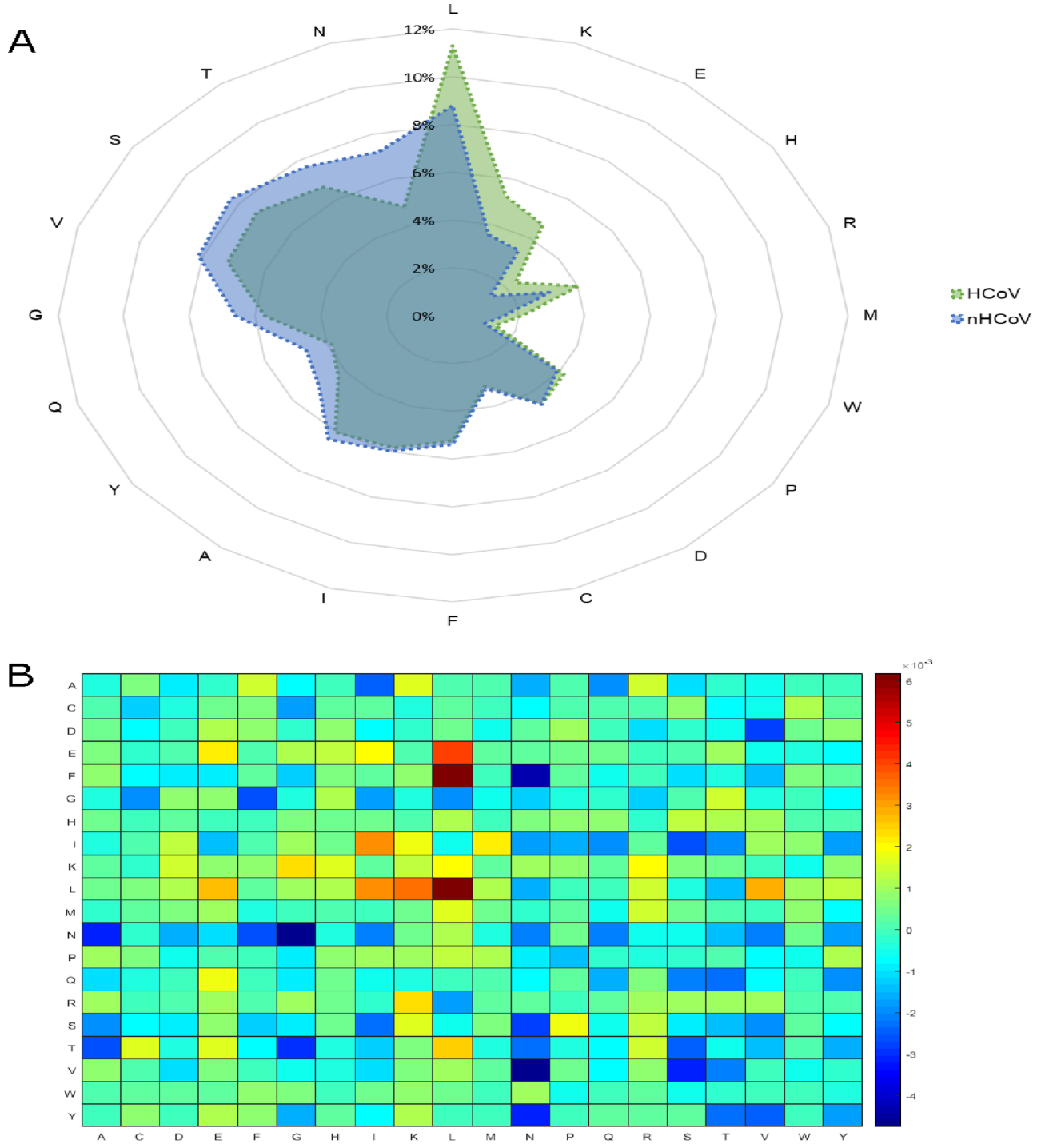


Figure 5. Amino acid and dipeptide compositional analysis. (A) Amino acid compositional differences between HCoV and *n*HCoV and (B) heatmap showing dipeptide compositional differences between HCoV and *n*HCoV.

CONCLUSIONS

Currently, substantial efforts are being made to develop therapeutic strategies[51−53] to eradicate the COVID-19 health crisis. Identifying the informative PCPs of COVID proteins could assist in vaccine design and COVID prevention. Due to the potential role of machine learning in solving many biological issues, it is considered a suitable tool for COVID-19 research. Hence, machine learning-based prediction models for COVID-19 are necessary to identify and analyze the important biomarkers for vaccine design. Here, to explore the PCPs of HCoV, we developed COVID-Pred for identification of valuable information of COVID proteins that could help in understanding their functions. A dataset consisting of protein sequences from 4320 HCoV and *n*HCoV was retrieved from the GISAID and NCBI databases. COVID-Pred was developed for the identification of informative PCPs and for the prediction of species-specific coronavirus proteins. COVID-Pred selected 11 PCPs and achieved 10-CV ACC, AUC, test ACC, and test AUC of 99.53%, 0.996, 97.80%, and 0.991, respectively, and obtained 100% (7/7) accuracy on an

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

Table 3. AAC Difference between the HCoV and *n*HCoV Proteins

amino acids	HCoV	*n*HCoV	composition difference
L	11%	9%	2%
K	5%	4%	2%
E	5%	3%	1%
H	2%	1%	1%
R	4%	3%	1%
M	2%	2%	1%
W	1%	1%	0%
P	4%	4%	0%
D	5%	5%	0%
C	3%	3%	0%
F	5%	5%	0%
I	6%	6%	0%
A	6%	6%	0%
Y	4%	5%	−1%
Q	4%	5%	−1%
G	6%	7%	−1%
V	7%	8%	−1%
S	7%	8%	−1%
T	7%	8%	−1%
N	5%	7%	−2%

independent data set consisting of seven HCoV S protein sequences.

Further analysis of five informative PCPs revealed that van der Waals forces, α-helices, frequencies of amino acids at α/β turns, helix capping, and mutability played some significant roles in differences between HCoV and *n*HCoV proteins. First, the characterization analysis of these informative PCPs revealed that there was a slight difference observed in the van der Waals volume between HCoV and *n*HCoV. Second, a difference in the $\Delta\Delta G^0$ value was observed between the S proteins of HCoV and *n*HCoV in which the amino acids R, Y, V, and K showed a larger difference in $\Delta\Delta G^0$ value compared to the other amino acids. Third, a larger difference in the amino acids for PALJ810116 was observed between HCoV and *n*HCoV for N, K, G, S, and Y. Fourth, we observed a larger difference in the normalized positional residue frequency at the helix termini N″ for the amino acids N, L, K, E, and G between HCoV and *n*HCoV as well as for R, P, K, S, and E between the S proteins of HCoV and *n*HCoV. Fifth, a larger difference in the relative mutability of amino acids between HCoV and *n*HCoV was observed for N, E, S, L, and T, whereas the relative mutability of amino acids between the S proteins of HCoV and *n*HCoV was observed for R, S, V, N, and P. The mutational analysis showed the mutations in the S proteins compared with the reference strain SARS-like bat virus, which falls under *n*HCoV (SLCoVZXC21/2015). Furthermore, we observed a difference in the AACs and DPCs between HCoV and *n*HCoV. The amino acid and dipeptide compositional differences for specific amino acids and dipeptides were also observed between HCoV and *n*HCoV. We believe that these findings could be helpful in understanding the functions of COVID proteins, which will be invaluable in designing vaccines.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00156.

Prediction performance of COVID-Pred evaluated using the ROC curve (Figure S1), comparison of PCPs between the S proteins of HCoV and *n*HCoV (Figure S2), normalized amino acid compositional preferences showing difference in the 11 PCPs between the S proteins of HCoV and *n*HCoV (Figure S3), amino acid compositional differences between the S proteins of HCoV and *n*HCoV (Figure S4), dipeptide compositional differences between HCoV and *n*HCoV (Figure S5), and 11 PCP values for the amino acids (AA index) (Table S4) (Supplementary Data 1); human−host coronaviruses (Table S1) and nonhuman−host coronaviruses (Table S2) (Supplementary Data 2); and mutations between the reference sequences of HCoV-19 and SARS-like bat coronaviruses (Table S3) (Supplementary Data 3) (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Shinn-Ying Ho − *Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan; Institute of Bioinformatics and Systems Biology, Department of Biological Science and Technology, and Center for Intelligent Drug Systems and Smart Bio-devices (IDS²B), National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan*; Email: syho@mail.nctu.edu.tw

Author

Srinivasulu Yerukala Sathipati − *Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, Wisconsin 54449, United States; Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu 300, Taiwan; Institute of Population Health Sciences, National Health Research Institutes, Miaoli 350, Taiwan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.1c00156

Author Contributions

S.Y.S. and S.Y.H. designed the system and carried out the detailed study. S.Y.S. implemented the programs. All authors participated in manuscript preparation and approved the final manuscript.

Funding

This work was funded by the Ministry of Science and Technology ROC under the contract numbers MOST 109−2221-E-009-129-, 109−2740-B-400-002-, and 108−3011-F-075-001-, and was financially supported by the "Center for Intelligent Drug Systems and Smart Bio-devices (IDS²B)" from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Notes

The authors declare no competing financial interest.
All the data used in this analysis can be found at the GISAID (https://www.gisaid.org) and the NCBI databases.

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

REFERENCES

(1) Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K. S. M.; Lau, E. H. Y.; Wong, J. Y.; Xing, X.; Xiang, N.; Wu, Y.; Li, C.; Chen, Q.; Li, D.; Liu, T.; Zhao, J.; Liu, M.; Tu, W.; Chen, C.; Jin, L.; Yang, R.; Wang, Q.; Zhou, S.; Wang, R.; Liu, H.; Luo, Y.; Liu, Y.; Shao, G.; Li, H.; Tao, Z.; Yang, Y.; Deng, Z.; Liu, B.; Ma, Z.; Zhang, Y.; Shi, G.; Lam, T. T. Y.; Wu, J. T.; Gao, G. F.; Cowling, B. J.; Yang, B.; Leung, G. M.; Feng, Z. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus−Infected Pneumonia. *N. Engl. J. Med.* **2020**, *382*, 1199−1207.

(2) Family - Coronaviridae. *In Virus Taxonomy*; King, A. M. Q.; Adams, M. J.; Carstens, E. B.; Lefkowitz, E. J., Eds. Elsevier: San Diego, 2012; 806−828.

(3) Zhou, P.; Yang, X.-L.; Wang, X.-G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H.-R.; Zhu, Y.; Li, B.; Huang, C.-L.; Chen, H.-D.; Chen, J.; Luo, Y.; Guo, H.; Jiang, R.-D.; Liu, M.-Q.; Chen, Y.; Shen, X.-R.; Wang, X.; Zheng, X.-S.; Zhao, K.; Chen, Q.-J.; Deng, F.; Liu, L.-L.; Yan, B.; Zhan, F.-X.; Wang, Y.-Y.; Xiao, G.-F.; Shi, Z.-L. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, *579*, 270−273.

(4) Masters, P. S. The Molecular Biology of Coronaviruses. *In Advances in Virus Research*; Academic Press, 2006; *66*, 193−292.

(5) Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T. S.; Herrler, G.; Wu, N.-H.; Nitsche, A.; Müller, M. A.; Drosten, C.; Pöhlmann, S. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181*, 271−280.

(6) Ou, X.; Liu, Y.; Lei, X.; Li, P.; Mi, D.; Ren, L.; Guo, L.; Guo, R.; Chen, T.; Hu, J.; Xiang, Z.; Mu, Z.; Chen, X.; Chen, J.; Hu, K.; Jin, Q.; Wang, J.; Qian, Z. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat. Commun.* **2020**, *11*, 1620.

(7) Lei, C.; Qian, K.; Li, T.; Zhang, S.; Fu, W.; Ding, M.; Hu, S. Neutralization of SARS-CoV-2 spike pseudotyped virus by recombinant ACE2-Ig. *Nat. Commun.* **2020**, *11*, 2070.

(8) de Haan, C. A. M.; Rottier, P. J. M. Molecular Interactions in the Assembly of Coronaviruses. *In Advances in Virus Research*; Academic Press, 2005, *64*, 165−230.

(9) McBride, R.; van Zyl, M.; Fielding, B. C. The coronavirus nucleocapsid is a multifunctional protein. *Viruses* **2014**, *6*, 2991−3018.

(10) Siu, Y. L.; Teoh, K. T.; Lo, J.; Chan, C. M.; Kien, F.; Escriou, N.; Tsao, S. W.; Nicholls, J. M.; Altmeyer, R.; Peiris, J. S. M.; Bruzzone, R.; Nal, B. The M, E, and N Structural Proteins of the Severe Acute Respiratory Syndrome Coronavirus Are Required for Efficient Assembly, Trafficking, and Release of Virus-Like Particles. *J. Virol.* **2008**, *82*, 11318.

(11) Neuman, B. W.; Kiss, G.; Kunding, A. H.; Bhella, D.; Baksh, M. F.; Connelly, S.; Droese, B.; Klaus, J. P.; Makino, S.; Sawicki, S. G.; Siddell, S. G.; Stamou, D. G.; Wilson, I. A.; Kuhn, P.; Buchmeier, M. J. A structural analysis of M protein in coronavirus assembly and morphology. *J. Struct. Biol.* **2011**, *174*, 11−22.

(12) de Haan, C. A.; Vennema, H.; Rottier, P. J. Assembly of the coronavirus envelope: homotypic interactions between the M proteins. *J. Virol.* **2000**, *74*, 4967−4978.

(13) Escors, D.; Ortego, J.; Laude, H.; Enjuanes, L. The Membrane M Protein Carboxy Terminus Binds to Transmissible Gastroenteritis Coronavirus Core and Contributes to Core Stability. *J. Virol.* **2001**, *75*, 1312.

(14) Liu, D. X.; Yuan, Q.; Liao, Y. Coronavirus envelope protein: A small membrane protein with multiple functions. *Cell. Mol. Life Sci.* **2007**, *64*, 2043−2048.

(15) Ortego, J.; Ceriani, J. E.; Patiño, C.; Plana, J.; Enjuanes, L. Absence of E protein arrests transmissible gastroenteritis coronavirus maturation in the secretory pathway. *Virology* **2007**, *368*, 296−308.

(16) Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; Yuan, M.-L.; Zhang, Y.-L.; Dai, F.-H.; Liu, Y.; Wang, Q.-M.; Zheng, J.-J.; Xu, L.; Holmes, E. C.; Zhang, Y.-Z. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265−269.

(17) Ren, W.; Qu, X.; Li, W.; Han, Z.; Yu, M.; Zhou, P.; Zhang, S.-Y.; Wang, L.-F.; Deng, H.; Shi, Z. Difference in Receptor Usage between Severe Acute Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin. *J. Virol.* **2008**, *82*, 1899.

(18) Ai, T.; Yang, Z.; Hou, H.; Zhan, C.; Chen, C.; Lv, W.; Tao, Q.; Sun, Z.; Xia, L. Correlation of Chest CT and RT-PCR Testing in Coronavirus Disease 2019 (COVID-19) in China: A Report of 1014 Cases. *Radiology* **2020**, *296*, E32−E40.

(19) Liu, X.; Wang, X.-J. Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. *J. Genetics and Genomics* **2020**, *47*, 119.

(20) Shinn-Ying, H.; Jian-Hung, C.; Meng-Hsun, H. Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. *IEEE Trans. Syst. Man Cybern. B Cybern.* **2004**, *34*, 609−620.

(21) Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* **2007**, *36*, D202−D205.

(22) Yerukala Sathipati, S.; Ho, S.-Y. Identifying the miRNA signature associated with survival time in patients with lung adenocarcinoma using miRNA expression profiles. *Sci. Rep.* **2017**, *7*, 7507.

(23) Yerukala Sathipati, S.; Sahu, D.; Huang, H.-C.; Lin, Y.; Ho, S.-Y. Identification and characterization of the lncRNA signature associated with overall survival in patients with neuroblastoma. *Sci. Rep.* **2019**, *9*, 5125.

(24) Yerukala Sathipati, S.; Ho, S.-Y. Novel miRNA signature for predicting the stage of hepatocellular carcinoma. *Sci. Rep.* **2020**, *10*, 14452.

(25) Srinivasulu, Y. S.; Wang, J.-R.; Hsu, K.-T.; Tsai, M.-J.; Charoenkwan, P.; Huang, W.-L.; Huang, H.-L.; Ho, S.-Y. Characterizing informative sequence descriptors and predicting binding affinities of heterodimeric protein complexes. *BMC bioinformatics* **2015**, *16*, S14−S14.

(26) Tsai, M.-J.; Wang, J.-R.; Ho, S.-J.; Shu, L.-S.; Huang, W.-L.; Ho, S.-Y. GREMA: Modelling of emulated gene regulatory networks with confidence levels based on evolutionary intelligence to cope with the underdetermined problem. *Bioinformatics* **2020**, *36*, 3833−3840.

(27) Chen, Y. H.; Yang, C. D.; Tseng, C. P.; Huang, H. D.; Ho, S. Y. GeNOSA: inferring and experimentally supporting quantitative gene regulatory networks in prokaryotes. *Bioinformatics* **2015**, *31*, 2151−2158.

(28) Chang, C.-C.; Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1−27.

(29) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* **2009**, *11*, 10−18.

(30) Fauchère, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; Pliska, V. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res.* **1988**, *32*, 269−278.

(31) Roth, C. M.; Neal, B. L.; Lenhoff, A. M. Van der Waals interactions involving proteins. *Biophys. J.* **1996**, *70*, 977−987.

(32) Amin, M.; Sorour, M. K.; Kasry, A. Comparing the Binding Interactions in the Receptor Binding Domains of SARS-CoV-2 and SARS-CoV. *J. Phys. Chem. Lett.* **2020**, 4897−4900.

(33) Das, S.; Sarmah, S.; Lyndem, S.; Singha Roy, A. An investigation into the identification of potential inhibitors of SARS-CoV-2 main protease using molecular docking study. *J. Biomol. Struct. Dyn.* **2020**, 1−11.

(34) 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*, 1444.

(35) Wang, Q.; Zhang, Y.; Wu, L.; Niu, S.; Song, C.; Zhang, Z.; Lu, G.; Qiao, C.; Hu, Y.; Yuen, K. Y.; Wang, Q.; Zhou, H.; Yan, J.; Qi, J. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* **2020**, *181*, 894−904.

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX

(36) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367*, 1260.
(37) Aydin, H.; Al-Khooly, D.; Lee, J. E. Influence of hydrophobic and electrostatic residues on SARS-coronavirus S2 protein stability: Insights into mechanisms of general viral fusion and inhibitor design. *Protein Sci.* **2014**, *23*, 603−617.
(38) Walls, A. C.; Tortorici, M. A.; Snijder, J.; Xiong, X.; Bosch, B.-J.; Rey, F. A.; Veesler, D. Tectonic conformational changes of a coronavirus spike glycoprotein promote membrane fusion. *Proc. Nat. Acad.Sci.* **2017**, *114*, 11157.
(39) O'Neil, K. T.; DeGrado, W. F. A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science* **1990**, *250*, 646−651.
(40) Palau, J.; Argos, P.; Puigdomenech, P. Protein secondary structure. Studies on the limits of prediction accuracy. *Int. J. Pept. Protein Res.* **1982**, *19*, 394−401.
(41) Madu, I. G.; Roth, S. L.; Belouzard, S.; Whittaker, G. R. Characterization of a highly conserved domain within the severe acute respiratory syndrome coronavirus spike protein S2 domain with characteristics of a viral fusion peptide. *J. Virol.* **2009**, *83*, 7411−7421.
(42) Aurora, R.; Rose, G. D. Helix capping. *Protein Sci.* **1998**, *7*, 21−38.
(43) Bell, J. A.; Becktel, W. J.; Sauer, U.; Baase, W. A.; Matthews, B. W. Dissection of helix capping in T4 lysozyme by structural and thermodynamic analysis of six amino acid substitutions at Thr 59. *Biochemistry* **1992**, *31*, 3590−3596.
(44) Chakrabartty, A.; Doig, A. J.; Baldwin, R. L. Helix capping propensities in peptides parallel those in proteins. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 11332−11336.
(45) Thapar, R.; Nicholson, E. M.; Rajagopal, P.; Waygood, E. B.; Scholtz, J. M.; Klevit, R. E. Influence of N-cap mutations on the structure and stability of Escherichia coli HPr. *Biochemistry* **1996**, *35*, 11268−11277.
(46) Serrano, L.; Kellis, J. T., Jr.; Cann, P.; Matouschek, A.; Fersht, A. R. The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. *J. Mol. Biol.* **1992**, *224*, 783−804.
(47) Duquerroy, S.; Vigouroux, A.; Rottier, P. J. M.; Rey, F. A.; Bosch, B. J. Central ions and lateral asparagine/glutamine zippers stabilize the post-fusion hairpin conformation of the SARS coronavirus spike glycoprotein. *Virology* **2005**, *335*, 276−285.
(48) Dayhoff, M.; Schwartz, R.; Orcutt, B. 22 a model of evolutionary change in proteins. *Atl. protein seq. struct.* **1978**, *5*, 345−352.
(49) van Dorp, L.; Acman, M.; Richard, D.; Shaw, L. P.; Ford, C. E.; Ormond, L.; Owen, C. J.; Pang, J.; Tan, C. C. S.; Boshier, F. A. T.; Ortiz, A. T.; Balloux, F. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *Infect. Gene. Evol.* **2020**, *83*, No. 104351.
(50) Phan, T. Genetic diversity and evolution of SARS-CoV-2. *Infect. gene. evol.* **2020**, *81*, 104260−104260.
(51) Song, J.-W.; Zhang, C.; Fan, X.; Meng, F.-P.; Xu, Z.; Xia, P.; Cao, W.-J.; Yang, T.; Dai, X.-P.; Wang, S.-Y.; Xu, R.-N.; Jiang, T.-J.; Li, W.-G.; Zhang, D.-W.; Zhao, P.; Shi, M.; Agrati, C.; Ippolito, G.; Maeurer, M.; Zumla, A.; Wang, F.-S.; Zhang, J.-Y. Immunological and inflammatory profiles in mild and severe cases of COVID-19. *Nat. Commun.* **2020**, *11*, 3410.
(52) Poh, C. M.; Carissimo, G.; Wang, B.; Amrun, S. N.; Lee, C. Y.-P.; Chee, R. S.-L.; Fong, S.-W.; Yeo, N. K.-W.; Lee, W.-H.; Torres-Ruesta, A.; Leo, Y.-S.; Chen, M. I. C.; Tan, S.-Y.; Chai, L. Y. A.; Kalimuddin, S.; Kheng, S. S. G.; Thien, S.-Y.; Young, B. E.; Lye, D. C.; Hanson, B. J.; Wang, C.-I.; Renia, L.; Ng, L. F. P. Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients. *Nat. Commun.* **2020**, *11*, 2806.
(53) Kneller, D. W.; Phillips, G.; O'Neill, H. M.; Jedrzejczak, R.; Stols, L.; Langan, P.; Joachimiak, A.; Coates, L.; Kovalevsky, A. Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. *Nat. Commun.* **2020**, *11*, 3202.

https://doi.org/10.1021/acs.jproteome.1c00156
J. Proteome Res. XXXX, XXX, XXX−XXX