


Sequence analysis

# Identification of evolutionarily stable functional and immunogenic sites across the SARS-CoV-2 proteome and greater coronavirus family

Chen Wang <sup>1,†</sup>, Daniel M. Konecki<sup>2,†</sup>, David C. Marciano<sup>1,\*</sup>,  
Harikumar Govindarajan<sup>1</sup>, Amanda M. Williams<sup>3</sup>, Brigitta Wastuwidyaningtyas<sup>1</sup>,  
Thomas Bourquard<sup>1</sup>, Panagiotis Katsonis<sup>1</sup> and Olivier Lichtarge<sup>1,2,3,4,\*</sup>

<sup>1</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA, <sup>2</sup>Quantitative and Computational Biosciences Graduate Program, Baylor College of Medicine, Houston, TX 77030, USA, <sup>3</sup>Cancer and Cell Biology Graduate Program, Baylor College of Medicine, Houston, TX 77030, USA, and <sup>4</sup>Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, TX 77030, USA

\*To whom correspondence should be addressed.

<sup>†</sup>The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.

Associate Editor: Alfonso Valencia

Received on February 26, 2021; revised on May 10, 2021; editorial decision on May 22, 2021; accepted on May 26, 2021

## Abstract

**Motivation:** Since the first recognized case of COVID-19, more than 100 million people have been infected worldwide. Global efforts in drug and vaccine development to fight the disease have yielded vaccines and drug candidates to cure COVID-19. However, the spread of SARS-CoV-2 variants threatens the continued efficacy of these treatments. In order to address this, we interrogate the evolutionary history of the entire SARS-CoV-2 proteome to identify evolutionarily conserved functional sites that can inform the search for treatments with broader coverage across the coronavirus family.

**Results:** Combining coronavirus family sequence information with the mutations observed in the current COVID-19 outbreak, we systematically and comprehensively define evolutionarily stable sites that may provide useful drug and vaccine targets and which are less likely to be compromised by the emergence of new virus strains. Several experimentally validated effective drugs interact with these proposed target sites. In addition, the same evolutionary information can prioritize cross reactive antigens that are useful in directing multi-epitope vaccine strategies to illicit broadly neutralizing immune responses to the betacoronavirus family. Although the results are focused on SARS-CoV-2, these approaches stem from evolutionary principles that are agnostic to the organism or infective agent.

**Availability and implementation:** The results of this work are made interactively available at <http://cov.lichtargelab.org>.

**Contact:** [lichtarge@bcm.edu](mailto:lichtarge@bcm.edu) or [david.marciano@bcm.edu](mailto:david.marciano@bcm.edu)

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

COVID-19 is a worldwide affliction. Since first being reported in December 2019 in Wuhan, Hubei province, China, the World Health Organization (WHO) has tallied more than 2 million COVID-19 related deaths and over 100 million infections worldwide (as of February 8, 2021) (Dong *et al.*, 2020). Although timely public health interventions can successfully curtail incidence, the threat of subsequent waves of infections and new strains that evade current treatments remains widespread (Ho *et al.*, 2021; Kraemer

*et al.*, 2020; Wang *et al.*, 2021; Wibmer *et al.*, 2021; Wu *et al.*, 2021). The novel betacoronavirus (SARS-CoV-2) that is causing the pandemic is closely related to other human coronavirus pathogens SARS-CoV, MERS-CoV (Chan *et al.*, 2020; Lu *et al.*, 2020), HCoV OC43, HKU1 and is more distantly related to the human infectious alphacoronaviruses HCoV 229E and HCoV NL63 (Lei *et al.*, 2018). Finding ways to control and prevent further infection are top priorities which include targeted drug discovery to impair viral mechanisms (Kim *et al.*, 2021; Li *et al.*, 2020; Rut *et al.*, 2020) and antigenic epitopes to develop robust vaccines (Poh *et al.*, 2020; van

Doremalen *et al.*, 2020). This study addresses both by utilizing evolutionary information from SARS-CoV-2 sequence and structural data to search for actionable functional sites for each protein in the SARS-CoV-2 genome.

First, we note that the approval of new drugs under normal circumstances often takes over 10 years (Dhama *et al.*, 2020; Pillaiyar *et al.*, 2020). In order to hasten the response, many current clinical trials for COVID-19 enlist antiviral agents that have targeted Zika, SARS-CoV, Ebola and MERS-CoV in the past (Dhama *et al.*, 2020; Jogalekar *et al.*, 2020). To test a broader variety of potential drugs to repurpose them for COVID-19 treatment, some studies screened thousands of clinical-stage or FDA-approved small molecules for antiviral activity (Riva *et al.*, 2020; White *et al.*, 2021). However, the antiviral activity in these large-scale screens may, in part, be cell-line specific (Hoffmann *et al.*, 2020) and therefore of unclear clinical relevance. Another approach to screen potential drugs for repurposing is to perform docking (Goodsell *et al.*, 2020) of clinical-stage or FDA-approved drugs to the SARS-CoV-2 proteome (Gupta *et al.*, 2020b; Ortega *et al.*, 2020). However, selection of the correct binding sites on the target proteins is crucial and difficult as protein surface cavities far exceed actual ligand binding sites that modulate function (Gupta *et al.*, 2018). Here we systematically suggest potential drug target sites for most SARS-CoV-2 proteins based on evolutionary information. As these sites are chosen for their conserved functional roles, broad pan-coronavirus/betacoronavirus relevance and minimal variability across all known SARS-CoV-2 variants, they should be prioritized in docking studies for drug repurposing.

Second, we note that understanding the immune response to SARS-CoV-2 infection is critical for vaccine development (Grifoni *et al.*, 2020b). Most early SARS-CoV-2 immune epitope discovery studies rely heavily on bioinformatic prediction tools as well as sequence and epitope work already done in SARS-CoV and MERS-CoV. B-cell linear and discontinuous epitope prediction tools have been used to identify possible SARS-CoV-2 epitopes (Ahmed *et al.*, 2020; Bhattacharya *et al.*, 2020; Grifoni *et al.*, 2020a), and several more recent studies experimentally determined SARS-CoV-2 immune epitopes (Le Bert *et al.*, 2020; Poh *et al.*, 2020). Interestingly, several groups have reported significant T-cell reactivity against SARS-CoV-2 epitopes in individuals without virus exposure (Grifoni *et al.*, 2020b; Le Bert *et al.*, 2020; Mateus *et al.*, 2020). Mateus *et al.* suggested that this could be due to cross reactivity between SARS-CoV-2 and other common human coronaviruses, such as OC43, HKU1, NL63 and 229E (Mateus *et al.*, 2020). Here we report an evolutionary metric, which can accurately separate cross-reactive epitopes from those that are not, and use this metric to suggest potential cross-reactive epitopes in SARS-CoV-2. Prioritizing these cross-reactive epitopes in vaccine development can potentially lead to broadly neutralizing immunity across the betacoronavirus family, provide starting points for mAb development and offer guidance in updating current vaccines in the face of variants which make them less effective. The AstraZeneca, Moderna and Pfizer vaccines currently being administered target the SARS-CoV-2 Spike glycoprotein (Baden *et al.*, 2021; Polack *et al.*, 2020; Voysey *et al.*, 2021), while the cross-reactive epitopes highlighted here occur in less variable regions of the Spike protein and in other SARS-CoV-2 proteins.

Here, we use the Evolutionary Trace (ET) method, which predicts the importance of protein sequence positions, from most important (0.0) to least important (100.0). This relative phylogenetic ranking reflects the variation entropy of each sequence position within and across the branches of an associated evolutionary tree, revealing evolutionary pressure points that correspond to functional and structural determinants, and the protein sites at which they often cluster (Mihalek *et al.*, 2004). Many studies validated this approach for predicting binding and catalytic functional sites (Lichtarge *et al.*, 1996), guiding protein engineering (Peterson *et al.*, 2015; Shenoy *et al.*, 2006; Sowa *et al.*, 2001) and predicting function (Amin *et al.*, 2013). When viewed as the gradient of the evolutionary landscape, ET rankings of residue importance can be combined with amino acid substitution log odds to estimate the likely impact, or Evolutionary Action (EA), of coding variations on protein function (Katsonis and Lichtarge, 2014). This first ET and EA analysis of a

full viral proteome identifies evolutionary important residues and functional sites in SARS-CoV-2.

## 2 Materials and methods

A brief description of the methods can be found here; for a more in-depth description of specific methods please see Supplementary Text.

### 2.1 Evolutionary trace

In order to map functional determinants in SARS-CoV-2 proteins, we applied the Evolutionary Trace (ET) approach (Lichtarge *et al.*, 1996; Mihalek *et al.*, 2004). This method ranks each amino acid position from most to least important during evolution by tracking how they vary along with divergences in the coronavirus phylogenetic tree. These rankings vary based on the precise choice of multiple sequence alignment (MSA). In order to produce robust ET rankings, three separate alignments were generated for each protein in the SARS-CoV-2 Wuhan-Hu-1 reference genome (NC\_045512.2) (Wu *et al.*, 2020), by querying three protein databases (UniRef90, UniRef100 and NCBI NR) for sequences with identity between 25% and 98%. This procedure filtered out sequences that were either overly distant or redundant. Only two proteins had too few matches for ET, NSP11 and ORF10, both of which have unknown function (Gadhav *et al.*, 2021; Pancer *et al.*, 2020) and have very short reference sequences (13 and 38 amino acids, respectively, Supplementary Fig. S1, Supplementary Dataset S1). The ET scores for all other proteins for each alignment and for the average scores across alignments were evaluated with the previously presented Selection Cluster Weighting (SCW)  $z$ -score (Mihalek *et al.*, 2004; Wilkins *et al.*, 2010). The  $z$ -scores for each structure were then ranked 1–4 which showed that ET scores from each of the three databases performed similarly well but the average ET of the three provided better  $z$ -scores in most cases (Supplementary Fig. S1C). ET rankings were further investigated by comparing the highest scoring regions with known functional sites.

### 2.2 Prediction of variant adjusted ET sites

Variant adjusted ET sites were predicted based on both the linear sequence as well as structural constraints. Residues were nominated as members of potential therapeutic sites based on their ET rankings, lack of variants as found in SARS-CoV-2 sequences retrieved from GISAID (Shu and McCauley, 2017), Genbank (Benson *et al.*, 2018) and the China National Center for Bioinformation (CNCB) (Zhao *et al.*, 2020), as well as surface accessibility, and structural proximity. Structurally identified therapeutic sites were compared to drug binding sites for agents known to bind to SARS-CoV-2 proteins. To generalize this approach to proteins without structure, linear sites were predicted based on ET rankings, current mutational profile and linear connectivity. Structural and linear predicted sites were compared to one another using Jaccard Similarity and Fisher's Exact test, to determine the usefulness of this method in the absence of a protein structure. Several ET metrics were also interrogated to determine their ability to highlight potential cross-reactive immunogenic epitopes. The best metric, sumEA/sum(100-ET ranking), was used to predict cross-reactive T-cell epitopes which are good potential therapeutic sites. The variant impact score used in this formula, EA, is described in greater detail in Supporting Information. Briefly, it combines the evolutionary gradient (approximated by ET) with the magnitude of a substitution (approximated by substitution log odds) to compute the evolutionary distance of a given variant travels in the fitness landscape. Summing the EA scores for variations observed between a SARS-CoV-2 epitope and the corresponding common cold coronavirus epitope is intended to estimate their functional distance from each other. This sumEA score is then adjusted by the overall phylogenetic conservation of the epitope as determined by the summation of ET scores of the amino acid positions covered by the epitope.

## 3 Results

### 3.1 Evolutionary trace of SARS-CoV-2

In order to map functional sites and determinants in SARS-CoV-2 proteins, we applied ET. With the multiple sequence alignments (Supplementary Fig. S1A, Supplementary Dataset S1) and the corresponding phylogenetic trees (Supplementary Figs S2–S4) for 24 of the 26 SARS-CoV-2 proteins (see Supplementary Methods and Materials), our protocol calculated the ET ranking of importance for 99.5% of SARS-CoV-2 amino acid residue positions (Supplementary Dataset S2) generated from each of three protein databases (UniRef90, UniRef100, NCBI NR) and combined them into a single average. In addition to considering the variety and breadth of sequences in the alignments, we assessed the quality of these ranks using a statistical measure that quantifies the distribution of ET rankings in the 3D structure; the Selection Cluster Weighting (SCW)  $z$ -score (Mihalek *et al.*, 2004). This metric measures how well top-ranked ET residues cluster structurally relative to a randomized distribution of scores on the structure (see Supplementary Materials and Methods). In previous studies, residues with better ET rankings (closer to 0) tended to cluster together at active sites, protein-protein interaction sites or other functional sites (Lichtarge *et al.*, 1996; Mihalek *et al.*, 2004; Wilkins *et al.*, 2010). Here, such clustering of top-ranked residues was particularly prominent in several SARS-CoV-2 proteins and complexes including the NSP5 main protease, the NSP7/NSP8/NSP12 RNA-dependent RNA polymerase complex and the NSP10/NSP16 RNA cap methyltransferase complex and can be visualized as groups of warm colored residues in the protein structure (Fig. 1). For 28 out of 38 protein structures analyzed, the SCW  $z$ -score is 2 standard deviations above the randomized background, suggesting the distribution of highly ranked ET residues are not random in these protein structures and confirming that the alignments are informative and that the resulting ET rankings are meaningful (Supplementary Fig. S1, Supplementary Dataset S3). For the proteins that do not reach significant  $z$ -scores there is a clear correlation to a lack of sequences in the alignments (e.g. NSP1, E, ORF3 and ORF7a), or, the structure belongs to a small domain within a larger protein (e.g. the macrodomain within NSP3 and the HR2 domain within the S protein).

To probe these smaller domains within large proteins we further investigated the ADP-ribose-phosphatase (ADPRP) and papain-like protease (PL<sup>pro</sup>) domains of NSP3. NSP3 was an intriguing case because top-ranked ET residues cluster well in its PL<sup>pro</sup> domain but not in its ADPRP domain (Supplementary Dataset S3). In order to better resolve ET rankings for NSP3, we generated new alignments, phylogenetic trees and ET residue rankings for the subsequences specific to each NSP3 domain structure (see Supplementary Materials and Methods). In this focused analysis, the PL<sup>pro</sup> domain now yielded ~50% more sequences leading to an increase in the clustering of top-ranked residues (Supplementary Fig. S5). For the ADPRP domain, thousands of additional sequences spanning the three domains of life and distantly related viruses were included which resulted in ET rankings that rivaled the significance of clustering in the PL<sup>pro</sup> domain. The differences we find in the phylogenetic trees of specific NSP3 domains confirm previous observations of alternate domain configurations in different coronavirus genera and even within clades of betacoronavirus (Lei *et al.*, 2018). The improvement in SCW  $z$ -score corresponds to a cluster of highly ranked ET residues within the ligand binding site of the ADPRP domain (Supplementary Fig. S5D) which was missing in the analysis of the full NSP3 reference sequence. Having better resolved ET rankings in the NSP3 domains, we returned to the main dataset to see how well ET rankings captured functional sites in other proteins.

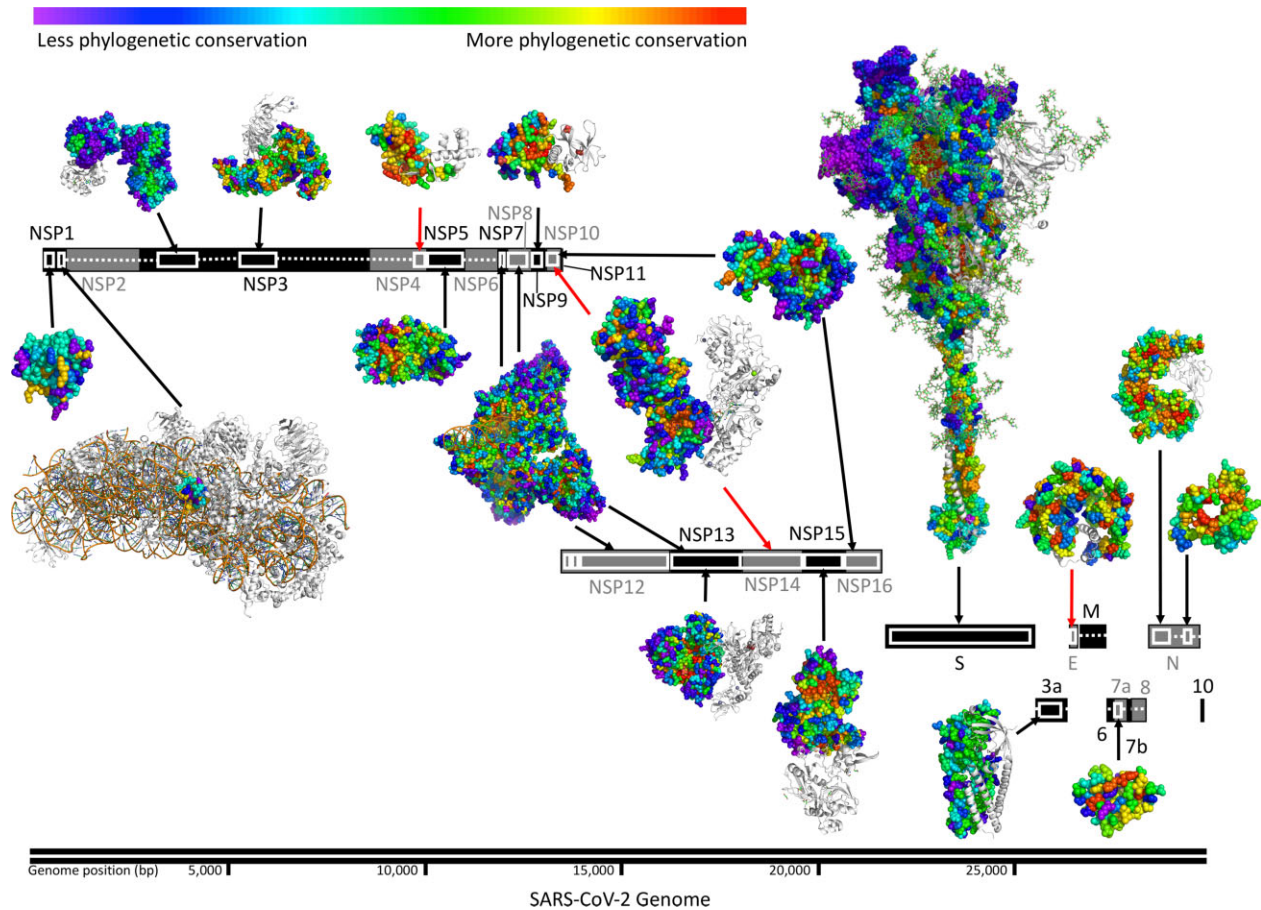
### 3.2 Phylogenetically conserved ligand binding sites

A catalog of SARS-CoV-2 ligand binding sites could serve as a timely resource for prioritizing therapeutic targets. Previous studies have shown that evolutionary sequence information correlates well enough with enzyme active sites so as to serve as 3D-templates for functional signatures (Amin *et al.*, 2013) and identify allosteric sites (Bhat *et al.*, 2020; Rodriguez *et al.*, 2010). Here we used NSP12,

NSP15 and NSP16 as examples to show how the evolutionary sequence information captured by ET can predict ligand binding sites for virus proteins. As shown in Figure 2A–C, top ranked ET residues cluster around the native ligands of NSP12 (RNA), NSP15 (GpU) and NSP16 (m7GpppA and SAM). In order to quantify this result, the enrichment of residues with ET ranking  $\leq 30$  within 5 Å of the ligands of NSP3, NSP12, NSP13, NSP14, NSP15 and NSP16 was ascertained by a Fisher's exact test. In each case, there is a statistically significant enrichment (Supplementary Fig. S6, Supplementary Dataset S4). We also find that in nine of the eleven ligand binding sites analyzed, our analysis better identifies known ligand binding sites than a previously reported analysis of the SARS-CoV-2 proteome using a percent identity metric (Supplementary Fig. S6, Supplementary Table S4) (Gupta *et al.*, 2020a). Moreover, several new functional sites are also predicted by ET (Fig. 2D and E). On the spike protein (S), one such ET cluster partially overlaps the S2' protease cleavage site and fusion peptide that are critical for membrane fusion and infectivity of the SARS virus (Madu *et al.*, 2009). On the N-terminal domain of the nucleoprotein (N), a cluster of highly ranked ET residues overlap (Lin *et al.*, 2014; Saikatendu *et al.*, 2007) the putative RNA binding site and may contribute to formation of the N protein-RNA helical filaments that package the RNA genome (Chen *et al.*, 2007). These results indicate ET can provide alternative drug target sites with no currently available ligand-bound structures.

In addition to being important to protein function, ideal drug target sites should be rarely mutated in the current outbreak so as to avoid the potential emergence of drug resistance. Thus, we focused on positions that do not have any mutations observed in the 139 607 high quality, full length SARS-CoV-2 sequences that were available as of December 8th, 2020. In order to translate proteome-wide ET ranks and mutational profiles into useful information for drug development, we defined clusters of mutation-free, surface-exposed residues that are highly ranked by ET and fall within 5 Å of each other (Supplementary Dataset S5) as variant adjusted 3D sites. The resulting catalog of potential drug targets includes 103 sites at ~4 sites per structure with the largest structure (full-length model of Spike, 6vsb\_1\_1\_1) having the highest number of sites. For NSP12, NSP15 and NSP16, the variant adjusted 3D sites overlap known ligand binding sites.

In order to evaluate whether these variant adjusted 3D sites may correspond to druggable target sites, we examined their overlap with sites observed in five SARS-CoV-2 protein–drug complex crystal structures. It is important to note that all five drugs showed an inhibitory effect in either cellular or biochemical assays (see Supplementary Data for details). The variant adjusted 3D sites were mapped onto the five SARS-CoV-2 protein–drug complexes and, as shown in Figure 3, all five drugs reside in protein surface pockets that are within or are very close to at least one residue of a predicted variant adjusted 3D site. The variant adjusted 3D site for NSP5 is not well recovered mostly due to a single SARS-CoV-2 sequencing entry (strain MT745875) wherein several residues in the protease active site are mutated (G143S, S144E and C145I), including the catalytic cystine residue. S144E and C145I are both caused by two nucleotide substitutions in the codon, and only observed in this strain (sampled on 06/24/20). It is unclear whether this is a sequencing artifact or represents a genuine active site plasticity that compromises NSP5's active site as a stable drug target. It does however illustrate the importance of accurately detecting emerging sequence variations when choosing drug targets. A clearer example of this is the tipiracil bound NSP15 active site which, although evolutionarily important to the overall coronavirus family, is not predicted to be a good drug target due to the presence of multiple variants observed in the current outbreak. Overall, these results show that predicted variant adjusted 3D sites can recover experimentally tested drug binding pockets and suggest new sites that can be targeted in computational docking approaches. In addition, because these sites are conserved across multiple coronavirus genera, these predicted variant adjusted 3D sites are anticipated to be relevant for identifying inhibitors of SARS-CoV-2 as well as more distantly related coronaviruses.



**Fig. 1.** Structural and sequence information permits identification of evolutionarily important sites in SARS-CoV-2. Structurally determined regions are highlighted (white boxes) and their residues colored by ET rank. Black arrows connect SARS-CoV-2 structures to their gene, red arrows indicate structures of homologous proteins. Ribosomal proteins in the NSP1 complex are shown (white). For multimeric structures, one monomer is also shown in white. List of structures used: 7k7p (NSP1), 6zlw (NSP1 C-term), 6woj (NSP3), 6w9c (NSP3), ExPasy NSP4 model 01 (NSP4), 6yb7 (NSP5), 6wxd (NSP9), 6xez (NSP7, 8, 12 and 13), 6zsl (NSP13), 5c8s (NSP10 and 14), 6wlc (NSP15), 6w4h (NSP10 and NSP16), 6vsb\_1\_1\_1 model (S), 6xdc (ORF3a), 5x29 (E), 6w37 (ORF7a), 6vyo (N) and 6zco (N)

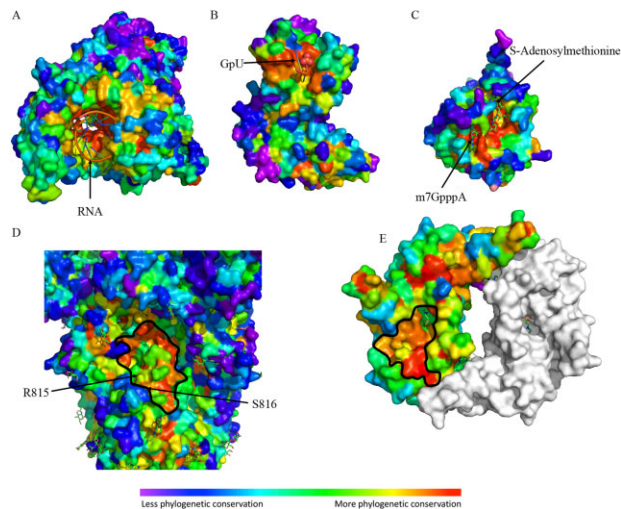
### 3.3 Conserved linear sites

Variant adjusted 3D sites may prove valuable in guiding drug design, but these approaches are dependent upon having high-resolution crystal structures and some structures are either not yet available (e.g. NSP2, NSP6, M and several accessory proteins), do not cover a majority of the protein (NSP3 and NSP4) or are too low in resolution for accurate docking studies (NSP12, NSP14, ectodomain of S, N, ORF3a and ORF7a). However, ET operates over protein sequences and can therefore identify phylogenetically important linear sequence fragments even in the absence of a 3D structure (Lichtarge et al., 2002). As in our approach to discover variant adjusted 3D sites, we combined ET residue ranking information with sequencing data from SARS-CoV-2 isolates to arrive at linear peptides along the proteome that are evolutionarily important and also show no variation in the current outbreak (Supplementary Fig. S7, Supplementary Dataset S6). In order to assess the value of these variant adjusted linear sites, we asked whether they could recapitulate variant adjusted 3D sites. Variant adjusted linear sites for NSP12 were mapped onto an available NSP12 structure and, as illustrated in Figure 4A, the majority of the 3D and linear sites overlap with each other. 84 of the 143 (59%) residues in the 3D sites were also identified as linear sites (total of 92 residues as linear sites) for NSP12. Variant adjusted linear sites and 3D sites also overlap well for other SARS-CoV-2 proteins, which was quantified by Jaccard Similarity and Fisher's exact test (Supplementary Dataset S7). These data suggest that variant adjusted linear sites contain functionally relevant information since they recapitulate variant adjusted 3D sites for proteins or domains without requiring 3D

structural data. In the absence of a protein structure, these linear sites could be useful in designing inhibitory peptides (Gu et al., 2005; Wu et al., 2020).

These linear sites are also connected to a second main approach toward resolving the pandemic, vaccine and mAb development. Although vaccines for COVID-19 are now available, several new variants that arose in the UK (B.1.1.7) (Rambaut et al., 2020) and South Africa (B.1.351, also known as 501Y.V2) (Tegally et al., 2020) have multiple substitutions in the Spike protein's receptor binding domain. Both are resistant to several classes of mAbs (Ho et al., 2021; Wibmer et al., 2021). While B.1.1.7 is ~2 fold more resistant to convalescent plasma, B.1.351 is more concerning as it can be ~11- to 33-fold more resistant to the convalescent plasma obtained from ~80% of patients (Ho et al., 2021; Wibmer et al., 2021). Ideally, effective protection against future outbreaks from related coronaviruses would include a broadly neutralizing effect wherein the immune system recognizes epitopes shared among coronavirus species. The prospect of raising a broad antibody response is bolstered by a study that naïve patients, never exposed to SARS-CoV-2, were found to possess a subset of T-cells that can cross-react to homologous epitopes shared by common cold coronaviruses and SARS-CoV-2 (Mateus et al., 2020). In this context, we note that ET rankings reflect the degree of homology over the phylogenetic tree, so we reasoned that summing ET scores over the length of an identified T-cell epitope may be able to estimate its potential for cross-reactivity.

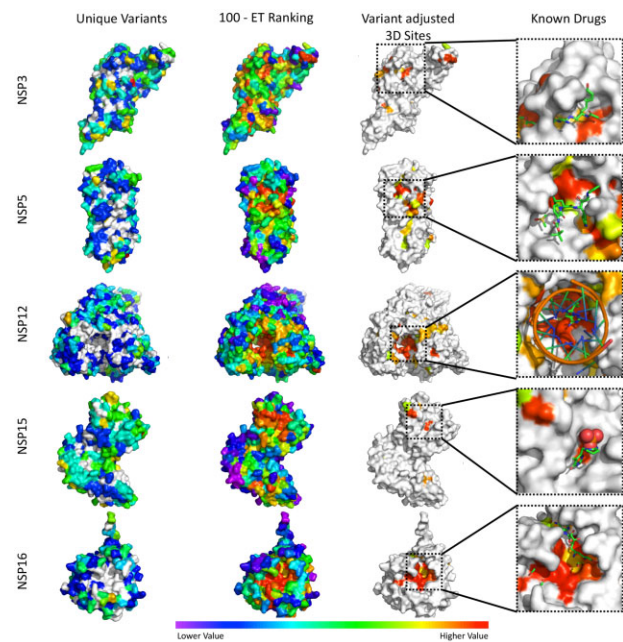
As a first step, we summed the ET ranks for each of the 40 SARS-CoV-2 epitopes that had been shown to react with patient-



**Fig. 2.** Top ET ranking residues overlap with known functional sites. ET recovers (A) the RNA binding site of NSP12 (RNA dependent RNA polymerase, pdb:6xez), (B) the active site of NSP15 (uridine-specific endoribonuclease, pdb:6x1b), (C) the substrates binding sites of NSP16 (RNA-cap methyltransferase, pdb:6wvn). ET also recovers a key functional (D) S2' protease cleavage site of S (key residues: R815 and S816, pdb:6v5b), and predicts (E) a site associated with the putative RNA binding site of N (pdb:6vyo). The S cleavage site between R815 and S816 is labeled and the putative sites of panels D and E are highlighted with a black outline. The site in panel D is a high priority target as it was found that residues 815-825, which overlap this site, comprise the most frequently recognized epitope among naïve and COVID-19 patients (Shrock *et al.*, 2020)

derived T-cells so that they could be ranked by predicted cross-reactivity to 161 common cold coronavirus epitopes assayed by Mateus *et al.* Although summing ET ranks could identify SARS-CoV-2 epitopes that are more likely to be cross-reactive (Supplementary Fig. S8), it did not account for the specific amino acid differences in the potentially cross-reactive homolog. In other words, ET ranks can predict whether or not a SARS-CoV-2 epitope will be cross-reactive in general, but not which epitope homologs will cross react.

In order to improve resolution of our predictions to specific epitope homologs, we next combined EA, a predictor of mutational impact, with the summed ET rankings. EA calculates the predicted impact of amino acid variations on protein function aiding in the interpretation of coding variants (Katsonis and Lichtarge, 2014). Summing the predicted impact of amino acid changes between a SARS-CoV-2 epitope and a homologous epitope in another virus (sumEA) while adjusting for the SARS-CoV-2 epitope's overall evolutionary importance [sum(100-ET ranking)] produced a metric that was able to separate cross-reactive epitopes from those that did not cross react (Fig. 4B, Supplementary Fig. S8, Supplementary Dataset S8). This metric, sumEA/sum(100-ET ranking), was then applied to 21 untested SARS-CoV-2 T-cell epitopes and their common cold homologs (Mateus *et al.*, 2020). From a total of 92 homologs we identified 23 with potential to cross react to one of five SARS-CoV-2 epitopes (Fig. 4C, Supplementary Dataset S9). These 5 SARS-CoV-2 epitopes along with the 9 others experimentally shown to possess cross-reactivity could be used in a multi-epitope vaccination strategy that provides a broad neutralizing response to currently circulating coronaviruses, SARS-CoV-2 and, possibly, future outbreaks. Subsequent to this analysis, further confirmation came when it was found that residues 815–825 of the Spike protein compose the most frequently recognized epitope among naïve and COVID-19 patients (Shrock *et al.*, 2020). These 11 residues are specifically highlighted by ET as being particularly evolutionarily conserved amino acids and are thereby responsible for our metric's prediction of cross reactivity in the 15 amino acid long epitopes used in the study by Mateus *et al.* This result and the generality of our approach suggest highly cross-reactive epitopes could be quickly identified in other families of pathogens.



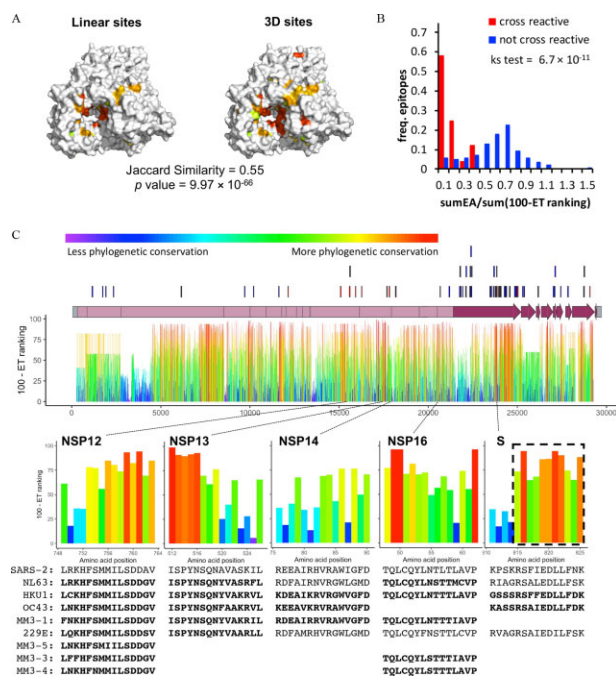
**Fig. 3.** Identification of variant adjusted 3D sites (5 Å) and their colocalization with known drug binding sites. Variant adjusted 3D sites for NSP3 (6w9c), NSP5 (6yb7), NSP12 (7bv1), NSP15 (6wlc) and NSP16 (6w4h) were identified as clusters of surface residues with low ET ranks and a lack of mutations in the current outbreak. In the known drugs panels, variant adjusted 3D sites were identified using apo form structures, then mapped to the co-structures of NSP3 with peptide inhibitor vir251 (PDB:6wx), NSP5 with potential drug 13 b (PDB:6y2f), NSP12 with drug remdesivir (7bv2), NSP15 in complex with potential drug tipiracil (PDB:6wxc) and NSP16 with sinefungin (PDB:6wkq). For structures in the 'Unique Variants' column 'Lower Values' in the color scale correspond to fewer variants, while 'Higher Values' correspond to more variants and white residues have no reported variants in the analyzed SARS-CoV-2 strains. For the '100—ET Ranking', 'Variant Adjusted 3D Sites' and 'Known Drugs' columns, 'Lower Values' correspond to less phylogenetic conservation while 'Higher Values' correspond to more phylogenetic conservation

## 4 Discussion

Rapid progress has been made in response to the SARS-CoV-2 pandemic; from sequencing, to structural determination, and drug and vaccine development (Jeong *et al.*, 2020; Li *et al.*, 2020; Peng *et al.*, 2020; Sanders *et al.*, 2020; Wu *et al.*, 2020). Here, we make use of coronavirus phylogenetics, and sequence and structure information to provide a functional map of sites that are not only stable across coronavirus families but are also stable to mutations in the current pandemic. These sites are favored strategic targets for pan coronavirus/betacoronavirus therapeutics that are less likely to be subjected to the rapid emergence of resistance from SARS-CoV-2 variants. In addition to focusing therapeutic studies, the data presented here can guide directed mutagenesis studies aimed at identifying the mechanism of action for successful therapies, not only in the context of the current outbreak but across future coronavirus outbreaks.

There are limitations to this study. The quality of our results depends on the number and range of homologous sequences available and a few of the SARS-CoV-2 proteins such as NSP1 and the accessory proteins do not reach significant z-scores or have many diverse sequences in their final alignments. The inability to recover more sequence information could be due to a higher evolutionary rate in these proteins that limits our ability to recognize distantly related homologs with very little sequence identity. More likely, these peripheral genes have been more recently recruited through the frequent recombination events that occur in the coronavirus family (Su *et al.*, 2016).

The equivalent of gene recruitment has occurred at the domain level in the NSP3 protein with its variable number of domains (10–16), some of which are unique to the betacoronavirus clade b containing SARS-CoV and SARS-CoV-2. Therefore, it is unsurprising



**Fig. 4.** Identification of variant adjusted linear sites with Evolutionary Trace. (A) Mapping of variant adjusted linear and structural sites on the surface of NSP12 (7bv1) with Jaccard Similarity value and Fisher's Exact test *P* value indicated. (B) Relative frequency distributions of sumEA/sum(100-ET ranking) for T-cell epitopes shown to either be cross reactive (red) or not (blue). SumEA/sum(100-ET ranking) metric predicts the functional impact of variants (EA) relative to the overall Evolutionary Trace rankings in the epitope. A Kolmogorov-Smirnov test (*ks* test) shows a significant difference in the distributions. (C) T-cell epitopes reported in Mateus et al. (2020) are shown above the SARS-CoV-2 genome (lines) and ET rankings within each protein are shown below. Shown are five SARS-CoV-2 epitopes (NSP12, NSP13, NSP14, NSP16 or S) that are predicted to cross react with the indicated common human coronavirus epitopes (bold text). Closely related coronavirus epitopes that did not meet our stringent threshold are also shown (normal text). The dashed box highlights the 11 amino acid stretch subsequently shown to be the most cross-reactive Spike protein epitope among naive and COVID-19 patients (Shrock et al., 2020).

that the initial alignments and corresponding ET rankings for full-length NSP3 are heavily influenced by the less divergent PL<sup>pro</sup> domain that is present across coronavirus clades and families. Domain-specific analysis of NSP3 greatly improved both the number of sequences returned, phylogenetic coverage and the resolution of ET results. This suggests that future work should include domain specific analyses for multidomain proteins. Such analyses are likely to provide ET rankings that identify important functional sites for individual domains, while full-length analyses can provide insight into how particular domains were recruited in specific branches of the phylogenetic tree.

Several other groups have focused on experimentally screening clinical-stage or FDA-approved small molecules with the hope of identifying and repurposing drugs for SARS-CoV-2 treatment. However, drug efficacy of top hits might be cell line specific (Hoffmann et al., 2020) and the mechanisms of drug action may be unclear or acting through modulation of the tissue culture cell. In silico docking studies (Deshpande et al., 2020; Gupta et al., 2020b) take a more targeted approach toward specific SARS-CoV-2 sites and benefit from knowledge of ligand binding sites. Although structural characterization of SARS-CoV-2 proteins is unprecedented, the structural information available is far from comprehensive. In order to bridge these knowledge gaps, we identified 3D clusters of surface residues that have low ET rankings and a lack of mutations in the current outbreak as potential drug target sites. Many of these variant adjusted 3D sites correspond to ligand bound active sites, but others map to evolutionarily important sites that have yet to be fully characterized. These variant adjusted sites are putative drug targets

which can guide docking studies to sites not immediately apparent from currently available structural information.

The depth of the phylogenetic tree for the sequences analyzed with ET can set expectations for how broadly a drug may inhibit homologs in different species. For instance, the active site of NSP12 is conserved throughout a deep phylogenetic tree of RNA viruses and an inhibitor targeting it, remdesivir, is effective against SARS-CoV-2, SARS-CoV, MERS and the distantly related Ebola RNA virus (de Wit et al., 2020; Eastman et al., 2020). Likewise, most of the other NSPs and the structural M (membrane) and N (nucleocapsid) proteins have deep RNA virus phylogenies and targeting them may provide broadly effective inhibitors. The most obvious candidates are the variant adjusted 3D sites that overlap with the ligand binding sites of NSP12, NSP13, NSP14 and 16. A less apparent drug target revealed by our analysis includes a putative RNA binding site on the N terminal domain of the N protein (Kang et al., 2020). The equivalent site in HCoV-OC43's N protein is targeted by compound PJ34 where it inhibits RNA binding activity and viral replication (Lin et al., 2014; Peng et al., 2020). In contrast, another inhibitor (5-benzyloxygramine), that induces aggregation of the MERS-CoV N protein and shows potent antiviral activity against that virus (Lin et al., 2020), binds two hydrophobic pockets. However, these pockets have undergone variation in the current SARS-CoV-2 outbreak, suggesting compounds targeting these areas are more likely to become susceptible to resistant SARS-CoV-2 variants.

In contrast to the aforementioned deep phylogenies predominantly composed of RNA virus sequences, the ADP ribose phosphatase sub-domain of NSP3 has a phylogenetic tree with few coronavirus sequences among a multitude of sequences that span three domains of life. Drugs targeting this domain may inhibit coronavirus infectivity but could also inhibit host ADP ribose phosphatases. ADP ribose phosphatase inhibitors have already been developed for cancer treatment and their application toward SARS-CoV-2 treatment is warranted (Kassab et al., 2020) but care should be taken to ensure unwanted side effects do not overshadow any benefits as a viral inhibitor.

Along with small molecule viral inhibitors, the development of immunological therapeutics to address COVID-19 can also be guided by the use of evolutionary information. We performed evolutionary analysis on SARS-CoV-2 T-cell epitopes capable of cross reacting with homologous peptides in other human coronaviruses (Le Bert et al., 2020; Mateus et al., 2020). This led to a new metric, sumEA/sum(100-ET ranking), that can better predict which epitopes will cross-react. In general, knowledge of cross-reactive epitopes could inform multi-epitope vaccine development efforts to direct the immune system toward a broadly neutralizing response.

The S protein evolves to bind different receptors (Fehr and Perlman, 2015), suggesting the high variation rate of the binding site is due to both this adaptive function and the avoidance of the host's adaptive immune system. Though several vaccines and mAbs are now in use, new strains are arising rapidly (Faria et al., 2021; Fiorentini et al., 2021; Rambaut et al., 2020; Tegally et al., 2020) and show signs of evading existing treatments (Wibmer et al., 2021; Wu et al., 2021). The two most worrisome mutations in those strains occur at N501 and E484 in the receptor binding domain of the S protein. These residues have ET ranks of 96 and 95.8 respectively, indicative of particularly rapid phylogenetic change across the coronavirus family and represent poor targets for therapeutics meant to remain effective against emerging SARS-CoV-2 variants. Despite the variability of the S protein, we identified a relatively conserved site (Fig. 2D) that corresponds to a fusion peptide adjacent to the S2' cleavage site (Madu et al., 2009) that is also the most cross-reactive epitope among naive and COVID-19 patients (Shrock et al., 2020) (Fig. 4C). This site is particularly appealing when considering that the 5H10 human mAb targeting the equivalent region in SARS-CoV was very effective in preventing disease a *Rhesus macaque* infection model (Miyoshi-Akiyama et al., 2011). The appearance of new strains makes it very likely that additional vaccine and mAb development will be necessary. We believe that targeting the SARS-CoV-2 variant-adjusted linear ET site corresponding to the S2' cleavage site and fusion peptide region and other sites highlighted by our study,

may provide protection that is less susceptible to the emergence of resistant variants.

This study was motivated by the current pandemic and uses evolutionary sequence information to guide the development of therapeutics for COVID-19. Although we are presently in the grip of COVID-19, this pandemic was preceded by the SARS and MERS outbreaks and it should be anticipated that related coronaviruses will cause future outbreaks. And while this study is focused upon SARS-CoV-2, it draws upon pieces of sequence information taken from the whole of the coronavirus family and thereby the findings are extendable to other coronavirus species, including those that have not yet been encountered. Indeed, the tools we present could be applied to any family of pathogen. Putting a pandemic virus into the evolutionary context of related viruses can expose a path to managing recovery and may offer therapeutics that cover future outbreaks.

## Data availability

The data and analyses presented here can be viewed and accessed from the interactive web page (<http://cov.lichtargelab.org>) and in [Supplementary Material](#).

## Funding

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under BAA-17-01 [contract #2019-19071900001 to O.L.]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. This work was also supported by the National Science Foundation [DBI-2032904 to O.L.], the Oskar Fischer Foundation and the National Institutes of Health [GM079656, GM066099 and AG061105 to O.L.].

*Conflict of Interest:* none declared.

## References

- Ahmed,S.F. *et al.* (2020) Preliminary identification of potential vaccine targets for the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. *Viruses*, **12**, 254.
- Amin,S.R. *et al.* (2013) Prediction and experimental validation of enzyme substrate specificity in protein structures. *Proc. Natl. Acad. Sci. USA*, **110**, E4195–202.
- Baden,L.R. *et al.* (2021) Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N. Engl. J. Med.*, **384**, 403–416.
- Benson,D.A. *et al.* (2018) GenBank. *Nucleic Acids Res.*, **46**, D41–D47.
- Bhat,A.S. *et al.* (2020) Recent advances suggest increased influence of selective pressure in allostery. *Curr. Opin. Struct. Biol.*, **62**, 183–188.
- Bhattacharya,M. *et al.* (2020) Development of epitope-based peptide vaccine against novel coronavirus 2019 (SARS-COV-2): immunoinformatics approach. *J. Med. Virol.*, **92**, 618–631.
- Chan,J.F.-W. *et al.* (2020) Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.*, **9**, 221–236.
- Chen,C.Y. *et al.* (2007) Structure of the SARS coronavirus nucleocapsid protein RNA-binding dimerization domain suggests a mechanism for helical packaging of viral RNA. *J. Mol. Biol.*, **368**, 1075–1086.
- Deshpande,R.R. *et al.* (2020) In silico molecular docking analysis for repurposing therapeutics against multiple proteins from SARS-CoV-2. *Eur. J. Pharmacol.*, **886**, 173430.
- Dhama,K. *et al.* (2020) COVID-19, an emerging coronavirus infection: advances and prospects in designing and developing vaccines, immunotherapeutics, and therapeutics. *Hum. Vaccin. Immunother.*, **16**, 1232–1238.
- Dong,E. *et al.* (2020) An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.*, **20**, 533–534.
- van Doremalen,N. *et al.* (2020) ChAdOx1 nCoV-19 vaccine prevents SARS-CoV-2 pneumonia in rhesus macaques. *Nature*, **586**, 578–582.
- Eastman,R.T. *et al.* (2020) Remdesivir: a review of its discovery and development leading to emergency use authorization for treatment of COVID-19. *ACS Cent. Sci.*, **6**, 672–683.
- Faria,N.R. *et al.* (2021) Genomic characterisation of an emergent SARS-CoV-2 lineage in Manaus: preliminary findings. <https://virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586>.
- Fehr,A.R. and Perlman,S. (2015) Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.*, **1282**, 1–23.
- Fiorentini,S. *et al.* (2021) First detection of SARS-CoV-2 spike protein N501 mutation in Italy in August, 2020. *The Lancet Infectious Diseases*, **21**, e147 10.1016/S1473-3099(21)00007-4
- Gadhave,K. *et al.* (2021) Conformational dynamics of NSP11 peptide of SARS-CoV-2 under membrane mimetics and different solvent conditions. *bioRxiv*, (preprint) 10.1101/2020.10.07.330068.
- Goodsell,D.S. *et al.* (2021) The AutoDock suite at 30. *Protein Science : a Publication of the Protein Society*, **30**, 31–43. 10.1002/pro.3934
- Grifoni,A. *et al.* (2020a) A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. *Cell Host Microbe*, **27**, 671–680.e2.
- Grifoni,A. *et al.* (2020b) Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. *Cell*, **181**, 1489–1501.e15.
- Gu,P. *et al.* (2005) Evolutionary trace-based peptides identify a novel asymmetric interaction that mediates oligomerization in nuclear receptors. *J. Biol. Chem.*, **280**, 31818–31829.
- Gupta,M. *et al.* (2018) Docking techniques in pharmacology: how much promising? *Comput. Biol. Chem.*, **76**, 210–217.
- Gupta,R. *et al.* (2020a) SARS-CoV-2 (COVID-19) structural and evolutionary dynamicome: insights into functional evolution and human genomics. *J. Biol. Chem.*, **295**, 11742–11753.
- Gupta,S. *et al.* (2020b) Identification of potential natural inhibitors of SARS-CoV2 main protease by molecular docking and simulation studies. *J. Biomol. Struct. Dyn.*, **11**, 1–12.
- Ho,D. *et al.* (2021) Increased resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7 to antibody neutralization. *Res. Sq*, (preprint) 10.21203/rs.3.rs-155394/v1
- Hoffmann,M. *et al.* (2020) Chloroquine does not inhibit infection of human lung cells with SARS-CoV-2. *Nature*, **585**, 588–590.
- Jeong,G. *et al.* (2020) Therapeutic strategies against COVID-19 and structural characterization of SARS-CoV-2: a review. *Front. Microbiol.*, **11**, 1723.
- Jogalekar,M.P. *et al.* (2020) Novel 2019 coronavirus: genome structure, clinical trials, and outstanding questions. *Exp. Biol. Med. (Maywood)*, **245**, 964–969.
- Kang,S. *et al.* (2020) Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm. Sin. B*, **10**, 1228–1238.
- Kassab,M.A. *et al.* (2020) Targeting dePARylation for cancer therapy. *Cell Biosci.*, **10**, 7.
- Katsonis,P. and Lichtarge,O. (2014) A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Res.*, **24**, 2050–2058.
- Kim,Y. *et al.* (2021) Tipiracil binds to uridine site and inhibits Nsp15 endoribonuclease NendoU from SARS-CoV-2. *Commun Biol*, **4**, 19
- Kraemer,M.U.G. *et al.*; Open COVID-19 Data Working Group. (2020) The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, **368**, 493–497.
- Le Bert,N. *et al.* (2020) SARS-CoV-2-specific T cell immunity in cases of COVID-19 and SARS, and uninfected controls. *Nature*, **584**, 457–462.
- Lei,J. *et al.* (2018) Nsp3 of coronaviruses: structures and functions of a large multi-domain protein. *Antiviral Res.*, **149**, 58–74.
- Li,H. *et al.* (2020) Coronavirus disease 2019 (COVID-19): current status and future perspectives. *Int. J. Antimicrob. Agents*, **55**, 105951.
- Lichtarge,O. *et al.* (1996) An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.*, **257**, 342–358.
- Lichtarge,O. *et al.* (2002) Evolutionary traces of functional surfaces along G protein signaling pathway. In: yengar R. and Hildebrandt J.D. (ed.) *Methods in Enzymology*. Vol. 344, Academic Press Inc., pp. 536–556.
- Lin,S.M. *et al.* (2020) Structure-based stabilization of non-native protein–protein interactions of coronavirus nucleocapsid proteins in antiviral drug design. *J. Med. Chem.*, **63**, 3131–3141.
- Lin,S.Y. *et al.* (2014) Structural basis for the identification of the N-terminal domain of coronavirus nucleocapsid protein as an antiviral target. *J. Med. Chem.*, **57**, 2247–2257.

- Lu, R. *et al.* (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet*, **395**, 565–574.
- Madu, I.G. *et al.* (2009) 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.*, **83**, 7411–7421.
- Mateus, J. *et al.* (2020) Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. *Science*, **370**, 89–94.
- Mihalek, I. *et al.* (2004) A family of evolution-entropy hybrid methods for ranking protein residues by importance. *J. Mol. Biol.*, **336**, 1265–1282.
- Miyoshi-Akiyama, T. *et al.* (2011) Fully human monoclonal antibody directed to proteolytic cleavage site in Severe Acute Respiratory Syndrome (SARS) coronavirus S protein neutralizes the virus in a Rhesus Macaque SARS model. *J. Infect. Dis.*, **203**, 1574–1581.
- Ortega, J.T. *et al.* (2020) Class A G protein-coupled receptor antagonist famotidine as a therapeutic alternative against SARS-CoV2: an in silico analysis. *Biomolecules*, **10**, 954.
- Pancer, K. *et al.* (2020) The SARS-CoV-2 ORF10 is not essential in vitro or in vivo in humans. *PLoS Pathog*, **16**, e1008959.
- Peng, Y. *et al.* (2020) Structures of the SARS CoV2 nucleocapsid and their perspectives for drug design. *EMBO J.*, **39**.
- Peterson, S.M. *et al.* (2015) Elucidation of G-protein and  $\beta$ -arrestin functional selectivity at the dopamine D2 receptor. *Proc. Natl. Acad. Sci. USA*, **112**, 7097–7102.
- Pillaiyar, T. *et al.* (2020) Recent discovery and development of inhibitors targeting coronaviruses. *Drug Discov. Today*, **25**, 668–688.
- Poh, C.M. *et al.* (2020) Two linear epitopes on the SARS-CoV-2 spike protein that elicit neutralising antibodies in COVID-19 patients. *Nat. Commun.*, **11**, 2806.
- Polack, F.P. *et al.*; C4591001 Clinical Trial Group. (2020) Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. *N. Engl. J. Med.*, **383**, 2603–2615.
- Rambaut, A. *et al.* (2020) Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. <https://virological.org/t/preliminary-genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-the-uk-defined-by-a-novel-set-of-spike-mutations/563>.
- Riva, L. *et al.* (2020) Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature*, **586**, 113–119. [10.1038/s41586-020-2577-1](https://doi.org/10.1038/s41586-020-2577-1)
- Rodriguez, G.J. *et al.* (2010) Evolution-guided discovery and recoding of allosteric pathway specificity determinants in psychoactive bioamine receptors. *Proc. Natl. Acad. Sci. USA*, **107**, 7787–7792.
- Rut, W. *et al.* (2020) Activity profiling and crystal structures of inhibitor-bound SARS-CoV-2 papain-like protease: A framework for anti-COVID-19 drug design. *Sci. Adv.*, **6**.
- Saikatendu, K.S. *et al.* (2007) Ribonucleocapsid formation of severe acute respiratory syndrome coronavirus through molecular action of the N-terminal domain of N protein. *J. Virol.*, **81**, 3913–3921.
- Sanders, J.M. *et al.* (2020) Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review. *JAMA*, **323**, 1824–1836.
- Shenoy, S.K. *et al.* (2006)  $\beta$ -arrestin-dependent, G protein-independent ERK1/2 activation by the  $\beta$ 2 adrenergic receptor. *J. Biol. Chem.*, **281**, 1261–1273.
- Shrock, E. *et al.*; MGH COVID-19 Collection & Processing Team16. (2020) Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science*, **370**, eabd4250.
- Shu, Y. and McCauley, J. (2017) GISAID: global initiative on sharing all influenza data – from vision to reality. *Euro Surveill.*, **22**, 30494.
- Sowa, M.E. *et al.* (2001) Prediction and confirmation of a site critical for effect or regulation of RGS domain activity. *Nat. Struct. Biol.*, **8**, 234–237.
- Su, S. *et al.* (2016) Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol.*, **24**, 490–502.
- Tegally, H. *et al.* (2020) Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *medRxiv*, (preprint) [10.1101/2020.12.21.20248640](https://doi.org/10.1101/2020.12.21.20248640).
- Voysey, M. *et al.* (2021) Safety and efficacy of the ChAdOx1 nCoV-19 vaccine (AZD1222) against SARS-CoV-2: an interim analysis of four randomised controlled trials in Brazil, South Africa, and the UK. *Lancet*, **397**, 99–111.
- Wang, P. *et al.* (2021) Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature*, **593**, 130–135.
- White, K.M. *et al.* (2021) Plitidepsin has potent preclinical efficacy against SARS-CoV-2 by targeting the host protein eEF1A. *Science*, **371**, 926–931.
- Wibmer, C.K. *et al.* (2021) SARS-CoV-2 501Y.V2 escapes neutralization by South African COVID-19 donor plasma. *Nature Medicine*, **27**, 622–625. [10.1038/s41591-021-01285-x](https://doi.org/10.1038/s41591-021-01285-x).
- Wilkins, A.D. *et al.* (2010) Sequence and structure continuity of evolutionary importance improves protein functional site discovery and annotation. *Protein Sci.*, **19**, 1296–1311.
- de Wit, E. *et al.* (2020) Prophylactic and therapeutic remdesivir (GS-5734) treatment in the rhesus macaque model of MERS-CoV infection. *Proc. Natl. Acad. Sci. USA*, **117**, 6771–6776.
- Wu, F. *et al.* (2020) A new coronavirus associated with human respiratory disease in China. *Nature*, **579**, 265–269.
- Wu, K. *et al.* (2021) mRNA-1273 vaccine induces neutralizing antibodies against spike mutants from global SARS-CoV-2 variants. *bioRxiv Prepr. Serv. Biol.* (preprint) [10.1101/2021.01.25.427948](https://doi.org/10.1101/2021.01.25.427948).
- Zhao, W.-M. *et al.* (2020) The 2019 novel coronavirus resource. *Yi Chuan Hered.*, **42**, 212–221.