HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2022 May 29.

Published in final edited form as:

Nat Methods. 2021 December ; 18(12): 1477–1488. doi:10.1038/s41592-021-01318-w.

A 3D Structural SARS-CoV-2-Human Interactome to Explore Genetic and Drug Perturbations

Shayne D. Wierbowski[1,2], **Siqi Liang**[1,2], **Yuan Liu**[2], **You Chen**[2,3], **Shagun Gupta**[1,2], **Nicole M. Andre**[4], **Steven M. Lipkin**[5], **Gary R. Whittaker**[4,6], **Haiyuan Yu**[1,2,†]

[1]Department of Computational Biology, Cornell University, Ithaca, NY 14853, USA

[2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY 14853, USA

[3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA

[4]Department of Microbiology and Immunology, Cornell University, Ithaca NY 14853, USA

[5]Department of Medicine, Weill-Cornell Medicine, New York, NY 10021

[6]Master of Public Health Program, Cornell University, Ithaca NY 14853, USA

Abstract

Emergence of new viral agents is driven by evolution of interactions between viral proteins and host targets. For instance, increased infectivity of SARS-CoV-2 compared to SARS-CoV-1 arose in part through rapid evolution along the interface between the Spike protein and its human receptor ACE2, leading to increased binding affinity. To facilitate broader exploration of how pathogen-host interactions might impact transmission and virulence in the ongoing COVID-19 pandemic, we performed state-of-the-art interface prediction followed by molecular docking to construct a 3D structural interactome between SARS-CoV-2 and human. We additionally carried out downstream meta-analyses to investigate enrichment of sequence divergence between SARS-CoV-1 and SARS-CoV-2 or human population variants along viral-human protein interaction interfaces, predict changes in binding affinity by these mutations/variants, and further prioritize drug repurposing candidates predicted to competitively bind human targets. We believe this resource (http://3D-SARS2.yulab.org) will aid in development and testing of informed hypotheses for SARS-CoV-2 etiology and treatments.

†Correspondence should be addressed to H.Y. (haiyuan.yu@cornell.edu).

Author Contributions Statement

H.Y. conceived and oversaw all aspects of the study, supervised research, and provided constructive feedback. Computational predictions, simulations, and analyses were run by S.D.W. Analysis of the NHGRI-EBI GWAS Catalog was performed by Y.C. The development and maintenance of the 3D-SARS2 website was done by S.L., Y.C., and S.G. Key contributions and useful feedback particularly for the introduction and expertise in virology were provided by N.M.A., S.M.L., and G.R.W. The manuscript was written by S.D.W. and H.Y. with additional contributions and edits from N.M.A., S.M.L., and G.R.W.

Competing Interests Statement

The authors declare no competing interests.

Code Availability

Scripts used for guided docking and relevant analyses are available on GitHub (https://github.com/hyulab/3D_SARS2) and DOI minted on Zenodo[123].

Introduction

The ongoing global COVID-19 pandemic has resulted in over 210 million SARS-CoV-2 infections and over 4.4 million deaths worldwide[1]. The coronavirus family of enveloped viruses causes respiratory and enteric tract infections in avian and mammalian hosts[2]. Seven well characterized human coronaviruses[3–5] exhibit symptoms ranging from mild respiratory illness to severe pneumonia and acute respiratory distress syndrome (ARDS). These coronaviruses are either highly transmissible yet generally not highly pathogenic (e.g. HCoV-229E, HCoV-OC43) or highly pathogenic but poorly transmissible (SARS-CoV-1 and MERS-CoV). Unique from these, SARS-CoV-2 is both highly transmissible and capable of causing severe disease with infectivity and pathogenesis differing between individuals[6,7]. While ~25–35% of infected individuals experience only mild or minimal symptoms, ~1–2% of infected patients die primarily from severe respiratory failure and ARDS[8,9]. Differences in morbidity, hospitalization, and mortality among different ethnic groups[10–15] are not fully explained by cardiometabolic, socioeconomic, or behavioral factors, suggesting a role for human genetic variation in SARS-CoV-2 pathogenicity. Insights into the evolution of SARS-CoV-2, its elevated transmission relative to SARS-CoV-1, and dynamic range of symptoms have been key areas of interest. These traits are likely driven by molecular mechanisms of pathology including interactions between the virus and its host, but specific causes are yet to be fully characterized.

Networks of protein-protein interactions between pathogens and their hosts provide one avenue to understand mechanisms of infection and pathology. Viral-human interactome maps have been compiled for SARS-CoV-1[16], HIV[17], Ebola virus[18], and Dengue and Zika viruses[19] among others. Recent, affinity-purification mass-spectrometry experiments on 29 SARS-CoV-2 proteins identified 332 viral-human interactions[20]. Inter-species interactions contribute to disease progression by facilitating pathogen entry into host cells[21–26], inhibiting host response proteins and pathways[27–29], and hijacking cell signaling or metabolism to accelerate cellular—and consequentially viral—replication[30–32]. Structures and dynamics of these interactions can provide insights into their roles. For instance, the viral-human binding interface between poxvirus chemokine inhibitor vCCI and human MIP-1β is shown to occlude domains vital to chemokine homodimerization, receptor binding, and interactions with GAG, thus explaining poxvirus' inhibitory effect on chemokine signaling[29]. Additionally, the dynamics of a herpesvirus cyclin and human cdk2 interaction induce a conformational change on cdk2 that matches its interaction with human cyclin A leading to dysregulated cell cycle progression[31].

Because protein-protein interactions mediate the majority of protein function[33–35], targeted disruption by small molecule inhibitors that compete for the same binding site provide a precise toolkit to modulate cellular function[33,35–38]. For instance, BCL-2 inhibitors that displace bound anti-apoptotic BCL-X interactors can treat chronic lymphocytic leukemia pathogenesis[39,40]. This approach can be particularly effective in viral networks and several potent inhibitors of key interactions have been developed. Disruption of viral complexes involved in viral replication has been successful in vaccinia virus[41] and human papilloma virus therapies[42,43]. Specifically, disruption of viral-host protein-protein interactions involved in early viral infection is an important therapeutic strategy. Discovery

that a population variant in the membrane protein CCR5 conferred resistance to HIV-1 by disrupting its interaction with the viral envelope glycoprotein led to the development of Maraviroc as an FDA approved treatment for HIV-1 that functions by blocking the interface for this interaction[23,44].

Here we apply a full-interactome modeling framework to construct a 3D structural interactome between SARS-CoV-2 and human proteins. Our framework first applies our previous ECLAIR framework[45] to identify interface residues for the whole SARS-CoV-2-human interactome and leverages these predictions to guide atomic-resolution interface modeling and docking in HADDOCK[46,47]. We additionally carried out in-silico scanning mutagenesis in PyRosetta[48] to predict the impact of mutations on interaction binding affinity and explored the overlap between protein-protein and protein-drug binding sites. All results from our 3D structural interactome are provided as a user-friendly web server allowing exploration of individual interactions or bulk download and analysis of the whole dataset. We further explore the utility of our 3D interactome modeling approach in identifying key interactions undergoing evolution along viral protein interfaces, highlighting population variants on human interfaces that could modulate the strength of viral-host interactions to confer protection from or susceptibility to COVID-19, and prioritizing drug candidates predicted to bind competitively at viral-human interaction interfaces, some of which could potentially be used for therapeutic purposes. Cumulatively these predictions and analyses are intended as a resource to facilitate investigation and further characterization of SARS-CoV-2-human interactions.

Results

Enrichment of Variation on the spike-ACE2 Binding Interface

We highlight the utility of computational and structural approaches to model the SARS-CoV-2-human interactome, from the interaction between the SARS-CoV-2 spike protein (S) and human angiotensin-converting enzyme 2 (ACE2) (Fig 1.a). This interaction mediates viral entry into human cells[3] and is among the only viral-human interactions solved in both SARS-CoV-1[49] and SARS-CoV-2[50–52]. Recent sequence divergences of the S protein are highly enriched at the S-ACE2 interaction interface (Fig 1.a; Log2OddsRatio=2.82, p=1.97e-5), indicating functional evolution around this interaction. We predicted the impact of these mutations on the binding affinity (ΔΔG) between the SARS-CoV-1 and SARS-CoV-2 versions of the S-ACE2 interaction using the Rosetta energy function[53] (Fig 1.b and 1.c). The negative ΔΔG value of −14.66 Rosetta Energy Units (REU) indicates an increased binding affinity using the SARS-CoV-2 S protein driven by better optimized solvation and hydrogen bonding potential fulfillment. Our result is consistent with the hypothesis that increased stability of the S-ACE2 interaction contributes to the elevated transmission of SARS-CoV-2[54]. Experimental kinetics assays have confirmed that compared to SARS-CoV-1, SARS-CoV-2 S protein binds ACE2 with 10–20-fold higher affinity[55] supporting the conclusions from our computational modeling.

A wide range in severity of and susceptibility to SARS-CoV-2 exists between individuals[6,7,56]. Genetic predisposition hypotheses explaining this range include both expression regulating and protein-coding variants[57,58]. For instance, an RNA-sequencing

analysis suggested higher expression of ACE2 in Asian males could facilitate viral entry and explain increased susceptibility among this population[59]. Alternatively, missense population variants in ACE2 could strengthen or weaken the S-ACE2 interaction, thereby modulating susceptibility to infection. We used a mutation scanning pipeline in PyRosetta[60,61] to predict the impact of six missense variants reported in gnomAD[62] that occur on the S-ACE2 interface (Fig 1.d). The three variants with the largest predicted impact on S-ACE2 binding affinity—ACE2_E37K (ΔΔG=1.50), ACE2_M82I (ΔΔG=2.95), and ACE2_G326E (ΔΔG=5.74)—were consistent with previous experimental screens identifying them as putative protective variants exhibiting decreased binding of ACE2 to S[63,64]. Our results highlight utility for a 3D structural interactome modeling approach in identifying interactions and mutations important for viral infection, pathogenesis, and transmission.

Constructing the 3D Structural SARS-CoV-2-Human Interactome

To facilitate similar investigation and hypothesis development at the full interactome scale, we next compiled a comprehensive 3D structural interactome between SARS-CoV-2 and human proteins based on 332 viral-human interactions uncovered in an early interactome screen by Gordon *et al.*[20]. First, we modeled SARS-CoV-2 proteins supplementing solved structures from the Protein Data Bank (PDB)[65] (16 of 29 proteins) with homology derived from SARS-CoV-1 templates (12 of 29 proteins). Homology models added one new structure for nsp14 (Extended Data Figure 1.a) while comparison against the available SARS-CoV-2 PDB structures from the remaining 11 validated the quality of our modeling approach (Extended Data Figure 1.b–c). For human interactors all models were obtained from the PDB or Modbase[66] (Extended Data Figure 2.a). We then predicted the interface residues for each interaction using our ECLAIR framework[45]. In total, our pipeline identified 679 interface residues across 21 SARS-CoV-2 proteins with an average 18.23 residues per interface and 5,790 across 189 human proteins with an average 17.4 residues per interface.

To provide structural interaction models for visualization and downstream analysis we performed guided docking in HADDOCK[46,47] using our high-confidence ECLAIR-predicted interface residues as restraints to refine the search space. To avoid potential biases in interface identification from docking low coverage models (Extended Data Figure 2.b) we only performed docking for 138 out of 332 interactions for which either 1) at least 33% of the full-length proteins were covered by available structures, or 2) available structures included at least one high-confidence ECLAIR prediction to use as docking restraint. In total we report 1,248 docked interface residues across 15 SARS-CoV-2 proteins with an average 33.4 residues per interface and 4,604 across 138 human proteins with an average 32.4 residues per interface. For all analyses, docked interface annotations were prioritized over initial ECLAIR predictions. The full interface annotations from our ECLAIR and docking predictions are available in Supplemental Table 1 and Supplemental Table 2, respectively.

Benchmarking ECLAIR and Guided Docking Predictions

Our specific applications of ECLAIR—for interspecies interactions—and HADDOCK—performing data-driven docking with computational rather than experimental priors—are unique from those these tools were previously validated for. To ensure the robustness and

quality of these methods for our interface prediction task, we constructed a comprehensive human-pathogen PDB benchmark set consisting of 509 interactions between a human protein and a viral or bacterial interactor (Fig 2.a). The full list of interactions in this benchmark set alongside the PDB sources plus true and predicted interfaces are provided in Supplemental Table 3.

To validate ECLAIR's applicability to inter-species interactions, we compared its published performance the test set of 200 human-human interactions to its performance on our human-pathogen PDB benchmark set. Both tasks achieved comparable performance (ROC AUC=0.69 vs. 0.74), although the intra-species task slightly outperformed inter-species (Fig 2.b). We note feature availability between sets—for instance, co-evolution features can only be calculated for intra-species interactions—may confound direct comparisons between different interaction sets. Overall, the evaluation of our benchmark conclusively shows that ECLAIR retains predictive power for inter-species interactions.

To evaluate the benefit of using ECLAIR predicted interfaces as restraints in HADDOCK docking, we compared our ECLAIR data driven protocol against a raw protocol with no restraints. From the original 509 inter-species interactions, 153 fit our criteria for docking. We compared interface annotations from each protocol based on precision and recall (Fig 2.c). Overall interface quality was comparable between both raw and guided protocols (precision=0.21 vs 0.19, p=0.15), however, the guided docking better recovered the total interface (recall=0.21 vs 0.29, p=5.88e-6). Previous evaluation on the HADDOCK framework confirms accurate interface predictions can be achieved even if the precise binding orientation is not recovered. While our main evaluation of interest is correct identification of interface residues, by evaluating the RMSD between docked and reference structures, we further demonstrate that the guided docking better recapitulated the true co-crystal structures (Fig 2.d; average RMSD=9.45 vs. 11.79, p=0.04).

Our aim in performing guided docking based on ECLAIR predicted interfaces was to produce atomic-resolution structures that reflected our residue level predictions for use in downstream analyses. However, we also hypothesized that docking would be effective in expanding accurate interface annotations to nearby residues if ECLAIR only identified a few high-confidence interface residues (Fig 2.e). Comparison of the precision and recall between ECLAIR and our guided docking (Fig 2.c) is consistent with this hypothesis and clearly demonstrates improvement in our guided docking approach over both raw docking and ECLAIR predictions.

Depletion of Human Disease Mutation at SARS-CoV-2 Interfaces

We explored evidence of interface-specific variation by mapping gnomAD[62] reported human population variants (Supplemental Table 4) and sequence divergences between SARS-CoV-1 and SARS-CoV-2 (Supplemental Table 5) onto predicted interfaces. Conserved residues generally cluster along protein-protein interfaces[67], and an analysis of SARS-CoV-2 structure and evolution similarly concluded highly conserved surface residues likely drove protein-protein interactions[68]. Consistent with these prior studies, we observed significant interactome-wide depletion for both viral and human variation along predicted interfaces comparable to that observed along solved human-human interfaces (Fig 3.a).

Nonetheless, considering each interaction individually, we identify 11 interaction interfaces enriched for human population variants (Fig 3.b), and 4 enriched for recent viral sequence divergences (Fig 3.c). Supplemental Table 6 provides the log odds enrichments for each interface. Similar to the S-ACE2 interface, a high degree of variation on these viral interfaces may indicate recent functional evolution around specific viral-human interactions. Because human evolution is slower, enrichment of population variants along the human interfaces is unlikely to be a selective response to the virus. Rather, interfaces with high population variation may represent edges in the interactome most prone to modulation by existing variation between individuals or populations. Alternatively, enrichment and depletion of variation along the human-viral interfaces could help distinguish viral proteins that bind along existing—likely conserved—human-human interfaces from those that bind using novel interfaces—unlikely to be under selective pressure.

To further explore the functional importance of variations within human interactors of SARS-CoV-2, we considered phenotypic associations reported in HGMD[69], ClinVar[70] or the NHGRI-EBI GWAS Catalog[71]. Interactors of SARS-CoV-2 were enriched for phenotypic variants from each database (Fig 3.d). Notably, several of the individual disease categories enriched among interactors, were consistent with SARS-CoV-2 comorbidities including heart disease, respiratory tract disease, and metabolic disease[72,73] (Fig 3.e; Supplemental Table 7). Disruption of native protein-protein interactions is one mechanism of disease pathology, and disease mutations are known to be enriched along protein interfaces[74,75]. Variants on predicted human-viral interfaces matched allele frequency distributions of variants off the interfaces, but were considered overall to be more deleterious by SIFT[76] and PolyPhen[77] (Extended Data Figure 3). However, while we show annotated disease mutations were significantly enriched along known human-human interfaces, enrichment was drastically reduced (HGMD) or insignificant (ClinVar) on human-viral interfaces (Fig 3.f). This is likely because mutations that disrupt human-viral interfaces would not disrupt natural cell function, and hence would be unlikely to manifest as disease phenotypes. Our finding that disease mutations and viral proteins affect human proteins at distinct sites is consistent with a two-hit hypothesis of comorbidities whereby proteins whose function is already affected by genetic background may be further compromised by viral infection.

Changes in Binding Affinity Between SARS-CoV-1 and SARS-CoV-2

Using a PyRosetta pipeline[48,60,61] we predicted the impact of sequence divergences between SARS-CoV-2 and SARS-CoV-1 on the binding energy (ΔΔG) of 138 viral-human interactions amenable to docking. Although the binding energy for most interactions was unchanged, we note that the divergence from SARS-CoV-1 to SARS-CoV-2 was biased towards a decreased binding energy (i.e. more stable interaction) (Fig 4.a; Supplemental Table 8). The significant outliers in these ΔΔG predictions may help pinpoint key differences between the viral-human interactomes of SARS-CoV-1 and SARS-CoV-2.

To further explore and validate the biological relevance of these predicted changes, we performed yeast two-hybrid (Y2H) screens to test 30 human interactors against both SARS-CoV-1 and SARS-CoV-2 baits. Our Y2H experiments reconstituted 6 of these interactions (20%) using the SARS-CoV-2 bait. Extensive prior studies across many species and

hundreds of well-validated interactions show inherent limits in assay sensitivity for all high-throughput interaction assays (detection rates span 15–25%)[78–81]. This is due in part to inability to match native expression, proper folding, or post-translational modifications under assay conditions. Our 20% reproducibility rate—in line with expected sensitivity of the Y2H system—indicates good quality of the published interactome. In each of the 6 reproduced interactions we predicted no changes in binding affinity between SARS-CoV-2 and SARS-CoV-1. Consistent with this prediction, each interaction was also detected using the SARS-CoV-1 bait (Fig 4.c). Docked models for these interactions suggest sequence divergences between SARS-CoV-1 and SARS-CoV-2 occurred away from the interface and would be unlikely to affect binding (Fig 4.c).

We additionally performed co-immunoprecipitation (co-IP) assays for the interaction between human DNA Primase Subunit 2 (PRIM2) and SARS-CoV-2 nsp1 (Fig 4.b; Source Data Figure 4, predicted $\Delta\Delta G=-17.3$ REU). Several deviations in nsp1 were predicted to cumulatively stabilize this interaction near the edges of its interface. Results from the co-IP validated our prediction showing that SARS-CoV-2 nsp1 was more effective at pulling down human PRIM2 than was SARS-CoV-1 nsp1. Moreover, a follow-up quantitative mass spectrometry comparison of SARS-CoV-2, SARS-CoV-1, and MERS-CoV by Gordon et al.[82] included 5 interactions we predicted to be more stable in SARS-CoV-2. Consistent with our predictions 3 of these (RNF41-nsp15, PRIM2-nsp1, and SNIP1-N) showed interaction preferences for the SARS-CoV-2 protein. Specifically, the interaction between RNF41 and nsp15 was exclusively detected in SARS-CoV-2. Overall, these independent experimental results together with our co-IP result thoroughly validate the accuracy of our 3D interactome modelling approach and demonstrate its utility in identifying functional differences between SARS-CoV-1 and SARS-CoV-2.

Impact of Population Variants on Binding Affinity

We hypothesized the dynamic range of patient responses and symptoms reported for SARS-CoV-2 infection can be explained in part by missense variations and their impact on viral-human interactions. This is consistent with previous reports that up to 10.5% of missense population variants can disrupt native protein-protein interactions[83] and that underlying genetic variation can explain up to 15% of variation in patient response and viral load in other viruses including HIV[84]. To explore this hypothesis we employed a previously benchmarked scanning mutagenesis protocol provided through PyRosetta[48,60] to identify candidate binding energy hotspot mutations for all docked interfaces. Out of 2,023 population variants on eligible interfaces, we identify 90 (4.4%) as predicted disruptive hotspots, and 51 (2.5%) as predicted stabilizing hotspots (Fig 5.d).

To validate our predictions for the impact of population variants, we generated a Ras GTPase-activating protein-binding protein 2 (G3BP2) variant, G3BP2_P121T (RS ID=rs1185000405) using site-directed mutagenesis as described previously[85]. We annotated this variant as strongly disruptive (predicted $\Delta\Delta G=10.3$ REU) and had confirmed earlier the interaction between N and wildtype G3BP2 could be recapitulated using Y2H. Comparing the Y2H results between wildtype and mutant G3BP2 confirmed complete disruption of the G3BP2-N protein interaction by G3BP2_P121T (Fig 4.e). Analysis of the docked models,

suggests this disruption is driven by steric clashes between the mutated residue in G3BP2 and Glu-323 and Thr-325 of the N protein. The unfavorable polar interaction and steric bulk from the hydroxyl side chain of the threonine variant was also predicted to induce a rotation in the Trp-330 of N disrupting hydrophobic interaction with Trp-282.

G3BP2 is implicated in cardiovascular diseases[86], potentially linking this interaction to known comorbidities. Moreover, G3BP2 alongside G3BP1 is an important target in viral etiology; sequestration of both proteins by SARS-CoV-2 N protein results in an inhibition of stress granule formation and suppression of host innate immune responses[87,88]. Therefore, the existence of naturally occurring variation disrupting this interaction is of particular interest. Although the G3BP2_P121T variant is rare (AF=0.00043%), it may affect SARS-CoV-2 progression in roughly 30,000 individuals who carry it worldwide. Overall, our computational and experimental work concretely shows that human population variants can modulate the SARS-CoV-2-human interactome network and that our interface and energy modelling predictions can help identify such variants. The full predicted impact of all 2,023 population variants along SARS-CoV-2 interaction interfaces is provided in Supplemental Table 9 and may inform future studies investigating genetic contribution to COVID-19.

Comparing Binding Sites of Drugs and SARS-CoV-2 Proteins

Drugs that directly interfere with viral-host interactions—for instance by competing for the same binding site—could provide promising clinical leads to target viral infection or replication. On this basis we consider potential for our 3D interactome modelling approach to inform drug repurposing strategies. We aimed to further prioritize a current candidate set including 76 expert-reviewed drugs targeting one or more of the 332 identified human interactors of SARS-CoV-2[20] based on potential for competitive binding. We performed protein-ligand docking using smina[89] to identify drug binding sites for 30 out of 76 candidate drug-target pairs that have available human receptor structures Supplemental Table 10. Smina a fork of the widely used AutoDock Vina, competes competitively in pose prediction challenges[89], as is validated by us to robustly identify the true binding site from the full protein surface on a published benchmark set of 4,399 experimentally solved protein-ligand complexes (Fig 5.a)[90].

We compared the overlap of predicted drug binding sites with the corresponding docked viral-human interaction interface for 16 cases with both predictions available. Overall drug binding sites were significantly enriched at the interaction interface compared to the rest of the protein surface (Fig 5.b; Log2OddsRatio=1.38, p=2.1e-7). Individually, we further prioritize 8 drugs that exhibited significant overlap between the drug- and viral-protein-binding sites (Fig 5.c), several of which have been explored by recent independent studies. A retroactive association study identified prior treatment with metformin as an independent factor associated with reduced mortality in diabetic patients[91], although a precise mechanism was not explored at the time. Ongoing phase 2 and phase 4 clinical trials are being conducted or planned for silmitasertib and valproic acid respectively[92,93].

As an example, we highlight orf9b-MARK3 interaction whose interface we predicted could be blocked by ZINC95559591 (MRT-68601 hydrochloride) (Fig 5.d and 5.e). MARK3 is a serine / threonine protein kinase involved in microtubule organization with implicated roles

in modulating gene expression by activating histone deacetylation proteins. Our models suggest both ZINC95559591 and orf9b bind and make several polar contacts with MARK3 (e.g. one with Tyr-134) near its active ATP site. Consistent with its known role as an inhibitor of MARK3[94] our model shows ZINC95559591 binds deep within the ATP active site of MARK3. By contrast the N-terminal tail of orf9b forms looser contact only entering the periphery of the active pocket. Therefore, we suspect ZINC95559591 may outcompete orf9b for this pocket; thus making it a prime candidate to explore targeted disruption of SARS-CoV-2-human protein-protein interactions through drug repurposing.

While this example fits our criteria for prioritized drug repurposing and competitive binding, it does raise further questions to consider. Namely, the functional role of a SARS-CoV-2-human interaction—whether the viral protein co-opts vs. disrupts native human protein function or if interaction is part of an immune response against the virus—is needed to inform potential clinical utility of drug repurposing. Since both orf9b and ZINC95559591 bind within the same MARK3 active site, both may induce an inhibitory effect and ZINC95559591 could be counterproductive; even if it outcompetes orf9b, it may replace a harmful viral inhibitor with a more potent chemical one. In this scenario, exploration of the predicted binding sites of SARS-CoV-2 proteins could still help uncover an inhibitory role in viral etiology. Moreover, it may be possible to design analogs of inhibitor drugs that retain high binding affinity to their receptor but lose their inhibitor activity. Therefore, while these factors may complicate the prospects of drug repurposing, we are optimistic that our 3D interactome modelling approach can facilitate understanding of viral mechanisms and may aid development of new treatments.

The SARS-CoV-2-Human 3D Structural Interactome Web Server

We constructed the SARS-CoV-2-human 3D interactome web server (http://3D-SARS2.yulab.org) to provide our computational predictions and modeling as a comprehensive resource to the public. All results and analyses described herein are directly available for bulk download or users can quickly navigation through the reported interactome to see a summary of our analyses for specific interactions of interest (Fig 6).

The interface comparison panel (Fig 6 top left) visualizes the interface annotation along a linear sequence and provides comparison against all other known or predicted interfaces from the same protein. This comparison may reveal biologically meaningful insights about the interface overlap and possible competition between viral and human interactors.

The mutations panel (Fig 6 top right) presents information on variation within each interaction partner; divergences from the SARS-CoV-1 or gnomAD population variants. We provide a log odds enrichment or depletion of variation along the interface which can help highlight interactions undergoing functional evolution for further characterization.

For interactions amenable to docking, the ΔΔG Information panel (Fig 6 lower left) compiles the predicted impact of all possible mutations across the docked interface on binding affinity. Individual mutations are colored by their z-score normalized ΔΔG prediction and can be toggled to only show the impacts of known variants. On the viral side, a cumulative ΔΔG

value compares binding affinity between the SARS-CoV-1 and SARS-CoV-2 versions of the protein.

Finally, the drug panel (Fig 6 lower right) describes any drugs known to target human proteins and provides information for each drug alongside display options for visualizing predicted binding conformations. The overlap between the drug binding site and interface with the viral protein is reported.

The SARS-CoV-2 human 3D structural interactome web server currently includes 332 viral-human interactions reported by Gordon *et al.*[20]. We will continue support for the web server with periodic updates as additional interactome screens between SARS-CoV-2 and human are published. As we update, a navigation option to select between the current or previous stable releases of the web server will be provided.

Discussion

Our 3D SARS-CoV-2-human interactome provides a comprehensive resource to supplement ongoing and future investigations into COVID-19. The analyses provided and discussed throughout highlight potential applications of these predictions to inform structure-based hypotheses regarding the roles of individual interactions and prioritize further functional characterization of: evolutionarily relevant interactions, causal links connecting population variation with differences in response to infection, and drug candidates that may interfere with interaction-mediated disease pathology. Our observation that perturbation from underlying disease mutations and viral protein binding occur at distinct sites on human proteins may warrant further investigation into whether the combined role of these two sources of perturbation is clinically relevant to mechanisms of comorbidities.

Although we have experimentally validated several of our predictions, we emphasize that further experimental characterization should be conducted to corroborate any hypotheses derived from individual predictions. Moreover, these predictions are not without limitation. Interface predictions may not be applicable to some published human targets identified by mass spectrometry[20] if they represent indirect complex associations rather than direct binary interactions[78]. Further, while structural coverage from SARS-CoV-2 proteins was robust, per-residue coverage of the human proteome is less complete (Extended Data Figure 2). Though we only performed molecular docking for low coverage structures when strong prior ECLAIR interface restraints were available, coverage restrictions can nonetheless introduce bias and may prohibit identification of true interface residues. Recent advances in protein-folding predictions[95–97] may ameliorate this restriction in the future. In the meantime, initial ECLAIR interface annotations—not susceptible to structural coverage limitations—may provide orthogonal value to docked models.

Additionally we caution that direct quantitative interpretation of Rosetta-predicted ΔΔG values is often difficult. In particular, relative importance of scoring function terms may differ between proteins and interactions of varying sizes and compositions. For these reasons, we only evaluate normalized predictions to compare the relative qualitative differences from our scanning mutagenesis results. Moreover, because mutated structure

optimization focuses only on side-chain repacking, our analysis is limited to mutations at or near the interface where side-chain repacking can have a direct effect. We expect mutations that significantly impact binding affinity through refolding or other allosteric effects exist but cannot be captured by our method.

Importantly, users can tailor the use of our raw predictions to their own interests; thus expanding upon the concepts and applications our analyses explore. For instance, we limited investigation of druggable interactions to repurposing known drugs that overlap and might disrupt viral-host interactions which we hypothesized would elicit the most promising clinical responses. However, this approach reduces the scope of the SARS-CoV-2-human interactome to only a few interactions that already have known drug candidates. An alternative application could prioritize candidate druggable interfaces throughout the whole SARS-CoV-2-human interactome by overlapping our interface annotations with predictions of druggable protein surfaces using recent deep-learning approaches[98] with the aim of designing novel protein-protein interaction inhibitors.

Overall, we believe our 3D structural SARS-CoV-2-human interactome web server (http://3D-SARS2.yulab.org) will prove to be a key resource in informing hypothesis-driven exploration of the mechanisms of SARS-CoV-2 pathology and host response. The scope, and potential impacts of our webserver will continue to grow as we incorporate the results of ongoing and future interactome screens between SARS-CoV-2 and human. Finally, we note our 3D structural interactome framework can be rapidly deployed to analyze future viruses.

Methods

Generation and Validation of SARS-CoV-2 Homology Models

Homology-based modeling of all 29 SARS-CoV-2 proteins was performed in Modeller[99] using a multiple template modeling procedure consistent with previous high-profile homology modelling resources[100]. In brief, candidate template structures for each query protein were selected by running BLAST[101] against all sequences in the Protein Data Bank (PDB)[65] retaining only templates with at least 30% identify. Remaining templates were ranked using a weighted combination of percent identity and coverage described previously[100]. The final set of overlapping templates to use was first seeded with the top ranked template with additional templates being added iteratively if: 1) overall coverage increase from the template was at least 10%, and 2) percent identify of the new template was no less than 25% the identity of the initial seed template. Query-template Pairwise alignments were generated in Modeller using default settings and were manually trimmed to remove large gaps (≥5 gaps in a 10 residue window). Finally, modelling was carried out using the Modeller automodel function.

This approach generated homology models for 18 out of 29 proteins. Based on manual inspection of the template quality and sources, homology models were further filtered to 12 models for which a high-quality template from a SARS-CoV-1 homolog was available. Moreover, during revision of this manuscript, newly deposited PDB structures for many SARS-CoV-2 proteins (https://rcsb.org/covid19) allowed independent validation of homology model quality based on the root-mean-square deviation (RMSD) following

alignment and refinement in PyMol[102]. Visual representations these alignments between modelled and solved structures are provided in Extended Data Figure 1. For all analyses SARS-CoV-2 PDB structures were prioritized where available, and only the homology model for nsp14 was retained.

Interface Prediction Using ECLAIR

Interface predictions for all 332 interactions reported by Gordon *et al.*[20] were made in two phases. In phase one, we leveraged our previously validated ECLAIR frameworkd[45] to perform initial residue-level predictions across all interactions. ECLAIR compiles five sets of features; biophysical, conservation, coevolution, structural, and docking. In brief, biophysical features are compiled using a windowed average of several ExPASy ProtScales[103], conservation features are derived from the Jensen-Shannon divergence[104,105] from known homologs for each protein, coevolution features between interacting proteins are derived from direct coupling analysis (DCA)[106] and statistical coupling analysis (SCA)[107] among paired homologs, structural features are obtained by calculating the solvent accessible surface area of available PDB[65] or ModBase[66] models using NACCESS[108], and docking features are the average inter-chain distance and surface occlusion per residue from a consensus of independent Zdock[109] trials.

Slight alterations were made to accommodate SARS-CoV-2-human predictions. First, construction of multiple sequence alignment (MSA) for SCA and DCA calculations require at least 50 species containing homologs of both interacting proteins. Therefore, co-evolution features could not be calculated for inter-species interactions. Second, MSAs for conservation features typically only allow one homolog per species. Because viral species classifications are less precise and are often subdivided into unique strains (and because all higher-order ECLAIR classifiers require protein conservation features) we modified the MSAs for viral proteins to include homologs from various strains in a single species. The initial prediction results from ECLAIR are provided in Supplemental Table 1.

Interface Prediction Using Guided HADDOCK Docking

Interface predictions for all 332 interactions reported by Gordon *et al.*[20] were made in two phases. In phase two, we leveraged high-confidence interface predictions from ECLAIR to perform guided docking in HADDOCK[46,47]. For a thorough introduction to protein-protein docking in HADDOCK, see https://www.bonvinlab.org/education/HADDOCK-protein-protein-basic/.

In brief, HADDOCK is designed to perform data-driven docking using (traditionally experimentally derived) priors about the interface. These data (e.g. scanning mutagenesis) often indicate sets of residues involved in the interface but no pairwise information linking interface residues between each protein. These residues (termed active residues) are used in conjunction with any neighboring surface residues (termed passive residues) to drive rigid body docking, by introducing a scoring penalty for any active residue on one protein not in proximity of an active or passive residue on the other. This approach is formalized as a set of ambiguous interaction restraints (AIR) that evaluate the distances of each active residue to the active or passive residues on the other protein. The approach ensures experimental

priors about interface composition are enforced, but leaves the exact orientation and pairing of residues flexible to HADDOCK's energy based scoring function.

To incorporate computational interface predictions from ECLAIR we use the standard HADDOCK protein-protein docking framework. Active residues are encoded as all high-confidence ECLAIR predictions at the surface (≥15% SASA). Passive residues are identified as all surface residues (≥40% SASA) within 6 Å of an active residue. For definition of surface residues, the 15% SASA cutoff is for consistency with our definition of interface residues, while the 40% SASA cutoff is for consistency with the typical recommendation in HADDOCK. All SASA calculations were carried out using NACCESS[108] and neighboring residues were selected using PyMol[102]. Following HADDOCK recommendations to reduce computational burden from using many restraints, we defined our AIR using only the alpha carbons and increased the upper distance limit for from 2 Å to 3 Å. All other HADDOCK run parameters were left at the default. In total 1000 rigid body docking trials were performed, and the top 200 scored orientations were retained for subsequent iterations refinement and analysis.

For each interaction we identified available PDB or homology model structures to determine whether the interaction should be eligible for docking. Previous benchmark evaluations show the HADDOCK performs using homology models, but that performance drops off for models produced from low sequence identity templates[110]. In all cases PDB models were prioritized over homology models. We next evaluated risks of using low coverage structures for protein-protein docking; using structure fragments that completely exclude the true interface residues will produce false interface predictions. We aimed to minimize this risk while maximizing the dockable interactome by setting two conditions for determining structure eligibility. First, protein structures covering at least 33% of the total protein length were considered sufficiently large for docking. Second, protein structures at least 50 residues in length and containing at least one high-confidence ECLAIR predicted interface residue to use as an active residue were made eligible. Inclusion of an ECLAIR-defined active residue gives us reasonable confidence that part of the interface is covered, and therefore, true docked interface predictions should be possible. When multiple structures were available for one protein, ranking was based on the sum of ECLAIR scores for all residues covered by each structure; we always selected the available structure most likely to include the true interface.

In total we performed guided HADDOCK docking on 138 out of 332 interactions. The remaining 194 interactions did not have reliable 3D models for both interactors. The top scored docked conformation from each HADDOCK run was retained. The final docked interface annotations are provided in Supplemental Table 2.

Definition of Interface Residues

We annotate interface residues from atomic resolution docked models, using an established definition for interface residues[45]. The solvent accessible surface area (SASA) for both bound and unbound docked structures was calculated using NACCESS[108]. We define as interface residue, any residue that is both 1) at the surface of a protein (defined as ≥ 15%

relative accessibility) and 2) in contact with the interacting chain (defined by a ≥ 1.0 $Å^2$ decrease in absolute accessibility).

Human-Pathogen Co-Crystal Structure Benchmark Set

We constructed a benchmark set of experimentally determined co-crystal structures to evaluate the performance of both our ECLAIR and guided HADDOCK docking interface predictions on inter-species interactions (Fig 2.a). First, we parsed 165,567 PDB structures, identified all interacting chains by interface residue calculation, and mapped PDB chains to UniProt protein IDs using SIFT[76] to identify a total of 33,242 unique protein-protein interactions. Using taxonomic lineages from UniProt we filtered this set to 7,738 interactions involving human proteins, of which 6,256 represented human-human intra-species interactions, and 1,482 represented inter-species interactions between human and some other species. Finally, to provide the most relevant set of interactions that would biologically similar to SARS-CoV-2-human interactions, we only considered interactions between human and viral proteins (346) or between human and bacterial proteins (163). We refer to this collective set of 509 co-crystal structures as our human-pathogen PDB benchmark set. The full list of structures and interface annotations for this benchmark set is provided in Supplemental Table 3.

To validate performance of ECLAIR predictions on the human-pathogen PDB benchmark, ECLAIR predictions were run as described above for SARS-CoV-2-human interactions. Evaluation of raw prediction probabilities was done by area under the receiver operating characteristic curve (AUROC) in python using scikit-learn and was compared against ECLAIR's original test set containing 200 intra-species interactions[45]. Precision and recall metrics were calculated based on ECLAIR's binary definition for high-confidence vs. non-interface predictions.

To validate HADDOCK guided docking performance using our human-pathogen PDB benchmark, we compared performance with a raw HADDOCK docking protocol. Guided docking was performed as described for SARS-CoV-2-human interactions. No PDB protein chains from the human-pathogen benchmark were used during docking. For raw HADDOCK docking no experimental constraints (AIR) were provided and the ranair and surfrest parameters in the run.cns were set to true. Using these parameters, each rigid dock generates one random AIR between one surface residue from each protein A and B which is used to ensure the two protein chains slide together during docking. Overall performance of protocols was evaluated based on precision and recall of the true interface (Fig 2.c). Secondary evaluation of was done based on root-mean-squared deviation (RMSD) in PyMol before refinement between the docked and co-crystal structures (Fig 2.d). When multiple co-crystal structures were used to define the interfaces, the RMSD was reported as the average RMSD against all co-crystal structures.

Compilation of Sequence Variation Sets

For analysis of genetic variation that may impact the viral-human interactome, two sets of mutations were compiled; 1) viral mutations, and 2) human population variants.

For viral mutations, we identified sequence divergences between SARS-CoV-1 and SARS-CoV-2 versions of each protein based on alignment. Representative sequences for 16 SARS-CoV-1 proteins were obtained from UniProt (Proteome ID UP000000354)[111,112]. Sequences for 29 SARS-CoV-2 proteins were reported by Gordon *et al.*[20] and based on genbank accession MN985325[113,114]. Notably, UniProt accessions for the SARS-CoV-1 proteome report two sequences for the uncleaved ORF1a and ORF1a-b which correspond to NSP1 through NSP16 in SARS-CoV-2. Sequence divergences were reported after pairwise Needleman Wench alignment[115,116] (using Blosum62 scoring matrix, gap open penalty of 10 and gap extension penalty of 0.5) between the corresponding protein sequences from each species. A total of 1,003 missense variants were detected among 23 SARS-CoV-2 proteins. No suitable alignment form a SARS-CoV-1 sequence was available for orf3b, orf8, or orf10. Additionally, orf7b, nsp3, and nsp16 were excluded because they were not involved in any viral-human interactions. The full list of SARS-CoV-2 mutations is reported in Supplemental Table 5.

We obtained human population variants for all 332 human proteins interacting with SARS-CoV-2 proteins from gnomAD[62]. We used gnomAD's graphQL API to run programmatic queries to fetch all missense variants per gene. Details on performing gnomAD queries in this manner are available in the gnomad-api github page (https://github.com/broadinstitute/gnomad-browser/tree/master/projects/gnomad-api). We used the Ensembl Variant Effect Predictor (VEP)[117] to map gnomAD DNA-level SNPs to equivalent protein-level UniProt annotations. After VEP mapping, variants were parsed to ensure the reported reference amino acid and position agree with the UniProt sequence and roughly 4.4.6% of variants that did not match were dropped from our dataset because they could not reliably be mapped to UniProt coordinates. In total 127,528 human population variants were curated. The full list of human population variants from GnomAD is reported in Supplemental Table 4.

Log Odds Enrichment Calculations

To determine enrichment or depletion, odds ratios were calculated as described previously[118]

$$OR = \frac{a \ / \ c}{b \ / \ d}$$

Where, a, b, c, and d describe values in a contingency table between case and exposure criteria. For a particular application, where we are interested in the enrichment of viral mutations or human populations variants (case: Variant vs. NonVariant) along predicted interaction interfaces (exposure: Interface vs. NonInterface), we would have…

$$a = Number\,of\,Variant\,Interface\,Residues$$
$$b = Number\,of\,NonVariant\,Interface\,Residues$$
$$c = Number\,of\,Variant\,NonInterface\,Residues$$
$$d = Number\,of\,NonVariant\,NonInterface\,Residues$$

Statistical tests for enrichment or depletion were performed by calculating the z-statistic and corresponding two-sided p-value for the odds ratio (unadjusted for multiple hypothesis testing)…

$$z = \frac{\ln OR}{\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}}$$

All reported odds ratios were $\log_2$ transformed to maintain interpretable symmetry between enriched and depleted values. To avoid arbitrary odds ratio inflation or depletion from missing data, in all cases where the interface residues were predicted by molecular docking, the odds ratio was altered to only account for positions that were included in the structural models used for docking.

Curation of Disease Associated Variants

To explore whether human proteins interacting with SARS-CoV-2 proteins were enriched for disease or trait associated variants, three datasets were curated; the Human Gene Mutation Database (HGMD)[69], ClinVar[70], and the NHGRI-EBI GWAS Catalog[71]. Disease annotations for HGMD and ClinVar were downloaded directly from these resources and mapped to UniProt. To calculate enrichment of individual disease terms, we reconstructed the disease ontology from NCBI MedGen term relationships (https://ftp.ncbi.nlm.nih.gov/pub/medgen/MGREL.RRF.gz) and propagated counts up through all parent nodes up to a singular root node. Significant terms were reported as the most general term with no more significant ancestor term (Supplemental Table 7, sheet 1). Raw enrichment values for all terms are also provided (Supplemental Table 7, sheet 2).

For curation of disease and trait associations from NHGRI-EBI GWAS Catalog (http://www.ebi.ac.uk/gwas/)[71], lead SNPs (p-value<5e-8) for all diseases/traits were retrieved on June 16, 2020. Proxy SNPs in high linkage disequilibrium (LD) (Parameters: $R^2 > 0.8$; pop: “ALL”) for individual lead SNPs were obtained through programmatic queries to the LDproxy API[119], which used phase 3 haplotype data from the 1000 Genomes Project as reference for calculating pairwise metrics of LD. Both lead SNPs and proxy SNPs were filtered to only retain missense variants.

In-silico Scanning Mutagenesis and ΔΔG Estimation

To explore importance of each SARS-CoV-2-human interface residue and the impact of all possible mutations along the interface, we performed in-silico scanning mutagenesis. We use a setup provided by the PyRosetta documentation (https://graylab.jhu.edu/pyrosetta/downloads/scripts/demo/D090_Ala_scan.py) designed around an approach previously benchmarked to correctly identify nearly 80% of interface hotspot mutations[60]. For consistency, we replaced the PyRosetta implementation’s definition of interface residues (≤ 8.0 Å away from partner chain), with our definition described above.

We encourage reference to the original well-documented demo for details, but in brief, we consider all interface residue positions and, begin by estimating the wildtype binding energy for the interaction. The complex state energy is calculated following a PackRotamersMover operation to optimize the side-chains of residues within 8.0 Å of the interface residue to be mutated. The chains are separated 500.0 Å to eliminate any interchain energy contributions

and energy for the unbound state is calculated the same way. The difference between these two values provides the binding energy for the wildtype structure.

$$\Delta G_{WT} = E_{complex} - E_{unbound}$$

To estimate the binding energy for all 19 amino acid mutations possible at the given position, each mutation is made iteratively, and the ΔG_{Mut} is as above using the mutated structures. Finally, the change in binding energy from each mutation is the difference between these two binding energies.

$$\Delta\Delta G = \Delta G_{Mut} - \Delta G_{WT}$$

The scoring function used for these calculations is as described previously[60] using the following weights; fa_atr=0.44, fa_rep=0.07, fa_sol=1.0, hbond_bb_sc=0.5, hbond_sc=1.0. To account stochasticity of the PackRotamersMover optimization between trials, all ΔΔG values are reported from an average of 10 independent trials. To test whether an a mutation had a significantly non-zero impact on binding energy, a two-sided z-test between the 10 independent trials was performed. To account for average impact of other same amino acid mutations at other positions along the interface, each average ΔΔG was z-normalized relative to the rest of the interface and outliers were called at ≥ 1 standard deviation away from the mean. Mutations that passed both criteria were identified as significant interface binding affinity hotspots. No adjustments were made for multiple hypothesis corrections.

Predicting ΔΔG from SARS-CoV-1 and SARS-CoV-2 Divergences

Estimates of the overall impact of the cumulative set of mutations between SARS-CoV-1 and SARS-CoV-2 were made based on the in-silico mutagenesis framework modified to introduce multiple mutations at a time. We generated interaction models using the SARS-CoV-1 protein by applying all amino acid substitutions between the two viruses to initial docked models containing the SARS-CoV-2 protein. A minority of mutations that comprised insertions or deletions could not be modelled under this framework. The ΔΔG calculation here was identical to the single mutation ΔΔG described above, except that side-chain rotamer optimization involved all residues within 8.0 Å of any of the mutated residues. The ΔΔG were calculated considering the SARS-CoV-1 as the wildtype such that a negative ΔΔG indicates the interaction is more stable (lower binding energy) in the SARS-CoV-2 version of the interaction compared to the SARS-CoV-1 version of the interaction…

$$\Delta\Delta G = \Delta G_{SARSCoV2} - \Delta G_{SARSCoV1}$$

To account for stochasticity between trials for these predictions (which notably had a larger impact likely due to the decreased constraints on rotamer optimization in these cases), this set of ΔΔG values was reported as an average of 50 trials. Significant outliers for overall binding affinity change from SARS-CoV-1 to SARS-CoV-2 were called based on similar criteria to the individual mutations, except the z-score normalization was performed relative to all other interactions.

Protein-ligand Docking Using Smina

To further prioritize 76 previously reported candidate drugs targeting human proteins in the SARS-CoV-2-human interactome[20], we performed protein-ligand docking for, 30 interaction-drug pairs (involving 25 unique drugs) that were amenable to docking. For docking, we excluded any human protein targets whose structures were below 33% coverage. To prep for docking, 3D structures for all ligands were first generated using Open Babel[120] and the command:

```
obabel -:"[SMILES_STRING]" --gen3d -opdb -O [OUT_FILE] -d
```

Protein-ligand docking was executed using smina[89] with the following parameters. The autobox_ligand option was turned on and centered around the receptor PDB file with an autobox_add border size of 10 Å. To increase the number of independent stochastic sampling trajectories and increase the likelihood of identifying a global minimum, the exhaustiveness was set to 40and the num_modes was set to retain the top 1000 ranked models. To reduce real wall time each docking process was run using 5 CPU cores (no impact on net CPU time). The final smina command used was as follows:

```
smina -r [RECEPTOR] -l [LIGAND] --autobox_ligand [RECEPTOR] --autobox_add 10
-o [OUT_FILE] --exhaustiveness 40 --num_modes 1000 --cpu 5 --seed [SEED]
```

Each protein-ligand docking command was repeated 10 times (essentially the same as one trial with exhaustiveness set to 400) with a unique seed in order to saturate the ligand binding search space as thoroughly as possible. We note that a single run with exhaustiveness ranging from 30–50 is considered sufficient for most applicaitons[89]. To retain candidate poses covering different low-energy binding sites, a final set of up to 10 of the best scoring poses with centers at least 1 Å away from one another was selected. Results described in this manuscript are reported based the top ranked pose. Protein residues involved in drug binding sites were annotated using the same criteria used to define interface residues. The Record Type for all ligand atoms was first manually changed from HETATM to ATOM because NACCESS otherwise excluded ligand atoms from the solvent accessible surface area calculations.

Validation of Smina Docking to Identify Drug Binding Sites

Past evaluation of smina show competitive performance across numerous Community Structure-Activity Resources (CSAR)[89,90]. However, tradition docking evaluation tasks, focus on sampling and correctly scoring docked conformations within a single known binding site and may frequently restrict the docking space to a few angstroms bounding box around the known ligand conformation. The focus is on recovering precisely how a ligand orients within a binding site rather than identifying the binding site from the whole protein surface.

Because this performance metric may not provide sufficient confidence in smina's ability to identify a binding site from scratch (our application in this manunscript) we re-benchmarked smina's performance using an established drug docking benchmark set containing 4,399 protein-ligand complexes representing 95 protein targets[90]. We defined true ligand binding site residues from the available crystal structure and evaluated the fraction correctly recovered by smina's top-ranked dock across the full protein surface.

Docking was performed as above and evaluated based on both redocking—ligand docked back into the exact receptor structure it came from—and crossdocking—ligand docked into an alternate conformation of the receptor it came from—conditions. Because the conformation of the binding pocket from an alternate receptor may not perfectly accommodate the ligand, crossdocking is considered more difficult, but also more representative of real conditions when making new predictions.

To provide a reference for whether smina selectively recovered the true binding site we calculated a baseline random expectation. Artificial binding sites were defined by selecting a single surface residue and its N nearest neighbors where N is the number of binding site residues in the true binding site. The average recovery of the true binding site from all such artificial binding sites was used as the null expectation for each drug-target pair.

Construction of plasmids for Y2H and co-IP

Clones of all human proteins tested were picked from the hORFeome 8.1 library[121]. Clones for all SARS-CoV-1 and SARS-CoV-2 proteins tested were designed to match GenBank entries AY357076 and MN908947 respectively. To construct plasmids for testing by Y2H viral genes were PCR amplified and cloned into PDEST-AD and PDEST-DB vectors (for Y2H). For co-immunoprecipitation (co-IP) Gateway LR reactions were used to transfer bait SARS-CoV-2 nsp1 protein into a pQXIP (ClonTech, 631516) vector modified to include a Gateway cassette featuring a carboxy-terminal 3×FLAG.

Yeast two-hybrid (Y2H) screens

Y2H experiments were carried out as previously described[78,83,122] in order to 1) confirm that SARS-CoV-2-human interactions previously detected by immunoprecipitation mass-spectrometry (IP-MS) could be recapitulated in Y2H, 2) compare the occurrence of interactions using SARS-CoV-1 vs. SARS-CoV-2 viral baits, and 3) profile the disruption of SARS-CoV-2-human interactions by human population variants. In brief human and viral clones were transferred into Y2H vectors pDEST-AD and pDEST-DB by Gateway LR reactions then transformed into *MAT*a Y8800 and *MAT*α Y8930, respectively. For comparisons of interest, the viral-human interactions were screened in both orientations; namely viral DB-ORF MATα transformants were mated against corresponding human AD-ORF *MAT*a transformants and vice versa. All DB-ORF yeast cultures were also mated against *MAT*a yeast transformed with empty pDEST-AD vector to screen for autoactivators. Mated transformants were incubated overnight at 30 °C, before being plated onto selective Synthetic Complete agar media lacking leucine and tryptophan (SC-Leu-Trp) to select for mated diploid yeast. After another overnight incubation at at 30 °C, diploid yeast were plated onto two sets of SC-Leu-Trp agar selection plates; one lacking histidine

and supplemented with 1 mM of 3-amino-1,2,4-triazole (SC-Leu-Trp-His+3AT), the other lacking adenine (SC-Leu-Trp-Ade). After overnight incubation at 30 °C, plates were replica-cleaned and incubated again for three days at 30 °C for final interaction calling.

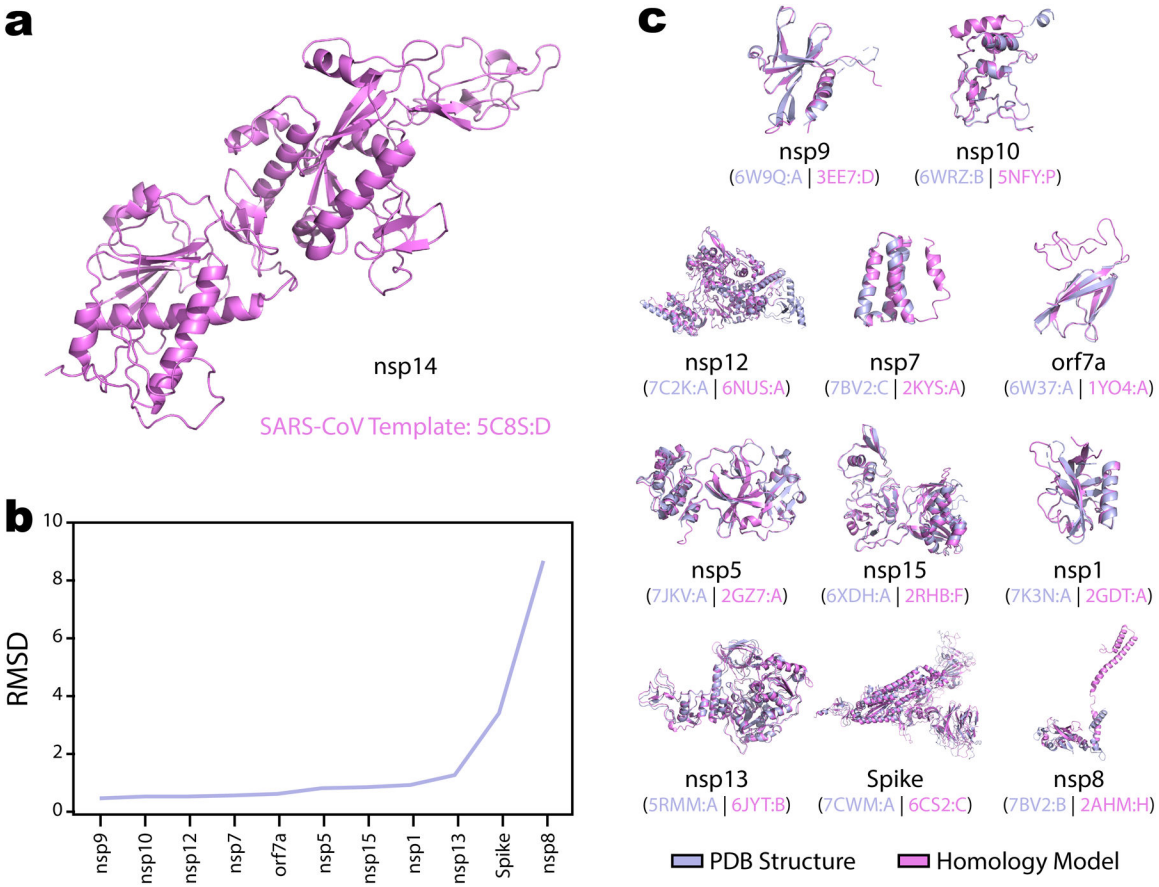
Cell culture, co-immunoprecipitation and western blotting

HEK 293T cells (ATCC, CRL-3216) were maintained in complete DMEM medium supplemented with 10% FBS. Cells were seeded onto 6-well dishes and incubated until 70–80% confluency. Cells were then transfected with 1 μg of either empty vector, SARS-CoV-1 nsp1 or SARS-CoV-2 nsp1, respectively, and combined with 10 μl of 1 mg ml−1 PEI (Polysciences, 23966) and 150 μl OptiMEM (Gibco, 31985–062). After 24 h incubation, cells were gently washed three times in 1×PBS and then resuspended in 200 μl cell lysis buffer (10 mM Tris-Cl pH 8.0, 137 mM NaCl, 1% Triton X-100, 10% glycerol, 2 mM EDTA and 1×EDTA-free Complete Protease Inhibitor tablet (Roche)) and incubated on ice for 30 min. Extracts were cleared by centrifugation for 10 min at 16,000g at 4 °C. For co-immunoprecipitation, 100 μl cell lysate per sample was incubated with 5 μl EZ view Red Anti-FLAG M2 Affinity Gel (Sigma, F2426) for 2 h at 4 °C under gentle rotation. After incubation, bound proteins were washed three times in cell lysis buffer and then eluted in 50 μl elution buffer (10 mM Tris-Cl pH 8.0, 1% SDS) at 65 °C for 10 min. Cell lysates and co-immunoprecipitated samples were then treated in 6×SDS protein loading buffer (10% SDS, 1 M TrisCl pH 6.8, 50% glycerol, 10% β-mercaptoethanol, 0.03% bromophenol blue) and subjected to SDS–PAGE. Proteins were then transferred from gels onto PVDF (Amersham) membranes. Anti-FLAG (Sigma, F1804) and anti-PRIM2 (abcam, ab241990) at 1:3,000 dilutions were used for immunoblotting analysis.

Cloning human population variants through site-directed mutagenesis

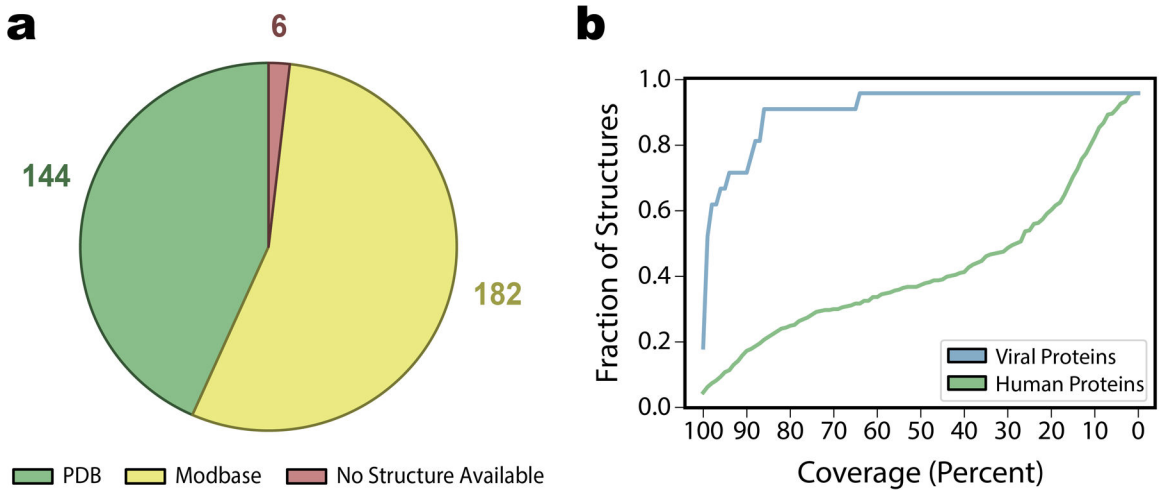
Generation of mutant clones containing human population variants was done using site-directed mutagenesis as described previously[85]. In brief, WT G3BP2 was picked from the hORFeome 8.1 library[121] and used as a template for site-directed mutagenesis. Site-specific mutagenesis primers (Eurofins) for mutagenesis were designed using the webtool primer.yulab.org. To minimize sequencing artifacts, PCR was limited to 18 cycles using Phusion polymerase (NEB, M0530). PCR products were digested overnight with DpnI (NEB, R0176) then transformed into competent bacteria cells to isolate single colonies. To confirm successful mutagenesis single colonies were then hand-picked, incubated for 21 h at 37 °C under constant vibration, and submitted for Sanger sequencing to ensure the desired single base-pair mutation—an no other mutations—had been introduced.

Extended Data



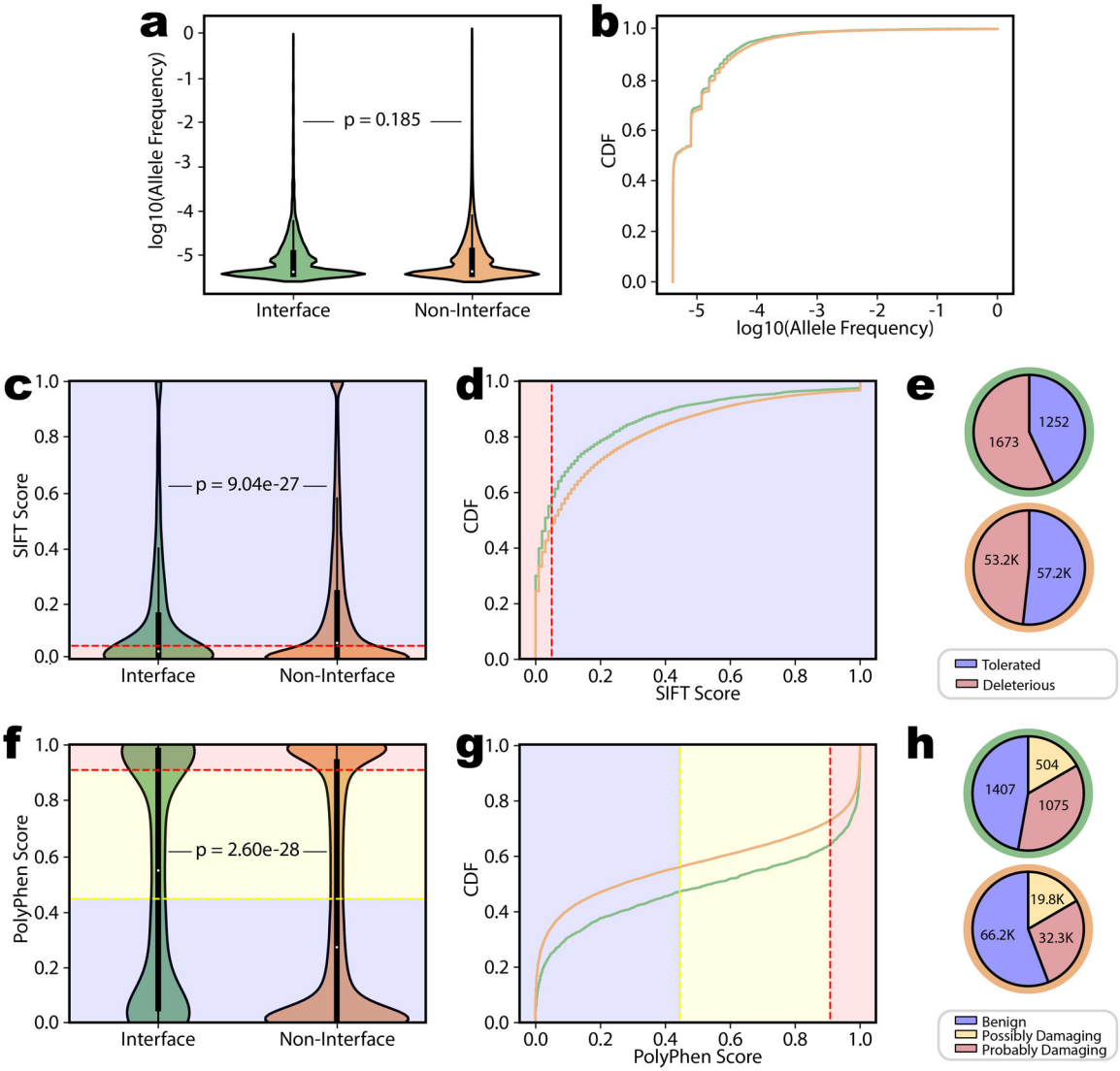
Extended Data Fig. 1. Homology modeling for SARS-CoV-2 proteins

a, Homology models for SARS-CoV-2 nsp14 modeled from a high-quality template for from SARS-CoV-1 nsp14 (PDB 5C8S:D). The nsp14 homology model was retained and used in downstream computational predictions. b, Quality assessment on 11 SARS-CoV-2 models generated using the same method as the nsp14 model. For these 11 proteins solved crystal-structures for the SARS-CoV-2 protein were deposited into the PDB during submission and revision of this manuscript and validated the quality of the homology modelling. Assessment is based on the on root-mean-square deviation (RMSD) following alignment of the homology model and PDB structure using PyMol. c, Visual representation of the alignment between all homology models (magenta) against their available PDB structure (light blue). PDB IDs and chains used for both the homology template and the reference PDB structure are indicated.



Extended Data Fig. 2. Source and coverage of available protein structures

a, Breakdown of the source of all structures available for the 332 human interactors of SARS-CoV-2 as being either a experimentally solved structure from the Protein Data Bank (n=144) a homology model from Modbase (n=182), or no available structure (n=6). b, Analysis of the coverage of all available structures for both human (green) and viral (blue) proteins. The fraction of structures retained with coverage greater than or equal to a range of coverage thresholds is shown. For our purposes, all available structures were used for solvent accessibility feature calculations for ECLAIR predictions, but structures were only retained for docking if either 1) total coverage was at least 33% of 2) the structure covered at least one high-confidence interface prediction from ECLAIR.



Extended Data Fig. 3. Summary of human population variant frequency and deleteriousness

a, b, Summary of allele frequency for human population variants either on (n=2,925) or off (n=118,042) the predicted human-viral interface presented as either a raw distribution or a cumulative density respectively. Variants in either category had roughly identical allele frequency distributions. Interior boxplots represent the distribution quartiles with whiskers representing the most extreme non-outlier values. c, d, Equivalent plots to a and b for the distribution of the SIFT deleteriousness scores for the same human population variant sets. Plots are colored based on the split between SIFT tolerated and deleterious categories. Population variants on the interface were significantly more likely to be classified deleterious by two-sample Kolmogorov-Smirnov test. e, Pie chart breakdown of SIFT categories. Pie chart outlines distinguish interface (green) from non-interface (orange). f, g, Equivalent plots to a and b for the distribution of the PolyPhen deleteriousness scores for the same human population variant sets. Plots are colored based on the split between PolyPhen benign, possibly damaging, and probably damaging categories. Population variants on the interface were significantly more likely to be classified deleterious by two-sample

Kolmogorov-Smirnov test. h, Pie chart breakdown of PolyPhen categories as in e. All p-values based on two-sided two-sample Kolmogorov-Smirnov test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by grants from NIGMS (R01 GM124559 and R01 GM125639), NIDDK (R01 DK115398), NICHD (R01 HD082568), NHGRI (UM1 HG009393), NSF (DBI-1661380), SFARI (575547) to H.Y.; NIH (R01AI35270) to G.R.W.; and William Strong COVID19 Biobank to S.M.L. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. Viral clones for SARS-CoV-1 and SARS-CoV-2 were synthesized and provided as generous gifts from Professor Pei Hui Wang from the Shandong University.

Data Availability

Protein-protein interaction sets and drug-target candidate pairs were obtained from the data reported in Gordon et al. 2020. Protein sequences were obtained from UniProt and Genebank. Population variants were mined from gnomAD using their batch query API (https://gnomad.broadinstitute.org/api). Disease and phenotypic variations were downloaded directly from HGMD, ClinVar, and the NHGRI-EBI GWAS Catalog. The data from these resources were modified and reformatted by some post-processing using custom python scripts. Wherever restrictions on relevant data did not apply (e.g. HGMD is restricted access) the formatted data are provided in the supplemental tables accompanying this manuscript or through the downloads page for our SARS-CoV-2-Human Interactome Browser (http://3D-SARS2.yulab.org/downloads).

Homology-modeling for SARS-CoV-2 proteins was performed through a custom script using modeller (based on their multiple templates modeling example https://salilab.org/modeller/manual/node21.html). Protein structures either presented as raw structures or used as templates in homology modelling were obtained from the Protein Data Bank (PDB). Homology models for human proteins were obtained from ModBase. Guided protein-protein docking and in-silico mutagenesis were performed in HADDOCK and PyRosetta respectively using these structures. Relevant analysis summaries for these experiments are provided in the supplemental tables that accompany this manuscript. Raw .pdb files for the original undocked structures and for all docking trials are provided through our downloads page (http://3D-SARS2.yulab.org/downloads).

We believe that all data has been accounted for, however, should any additional piece of data supporting the findings of this study later become of interest, the authors will strive to make it available upon request. Please address any additional requests or clarifications to Shayne Wierbowski (sdw95@cornell.edu) and Haiyuan Yu (haiyuan.yu@cornell.edu).

References

1. COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University. 2020 [cited 2020 August 26, 2020]; Available from: https://coronavirus.jhu.edu/map.html.

2. Fehr AR and Perlman S, Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol, 2015. 1282: p. 1–23. [PubMed: 25720466]
3. Zhou P, et al. , A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 2020. 579(7798): p. 270–273. [PubMed: 32015507]
4. McIntosh K and Perlman S, Coronaviruses, Including Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS). Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases, 2015: p. 1928–1936.e2.
5. Zhou H, et al. , A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. Curr Biol, 2020. 30(11): p. 2196–2203 e3. [PubMed: 32416074]
6. Gupta A, et al. , Extrapulmonary manifestations of COVID-19. Nat Med, 2020. 26(7): p. 1017–1032. [PubMed: 32651579]
7. Wang D, et al. , Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. JAMA, 2020. 323(11): p. 1061–1069. [PubMed: 32031570]
8. Yang X, et al. , Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. Lancet Respir Med, 2020. 8(5): p. 475–481. [PubMed: 32105632]
9. Zhou F, et al. , Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. Lancet, 2020. 395(10229): p. 1054–1062. [PubMed: 32171076]
10. Palaiodimos L, et al. , Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. Metabolism, 2020. 108: p. 154262. [PubMed: 32422233]
11. Ferdinand KC and Nasser SA, African-American COVID-19 Mortality: A Sentinel Event. J Am Coll Cardiol, 2020. 75(21): p. 2746–2748. [PubMed: 32330545]
12. Killerby ME, et al. , Characteristics Associated with Hospitalization Among Patients with COVID-19 - Metropolitan Atlanta, Georgia, March-April 2020. MMWR Morb Mortal Wkly Rep, 2020. 69(25): p. 790–794. [PubMed: 32584797]
13. Raisi-Estabragh Z, et al. , Greater risk of severe COVID-19 in Black, Asian and Minority Ethnic populations is not explained by cardiometabolic, socioeconomic or behavioural factors, or by 25(OH)-vitamin D status: study of 1326 cases from the UK Biobank. J Public Health (Oxf), 2020.
14. Moore JT, et al. , Disparities in Incidence of COVID-19 Among Underrepresented Racial/Ethnic Groups in Counties Identified as Hotspots During June 5–18, 2020 – 22 States, February-June 2020. MMWR Morb Mortal Wkly Rep, 2020. 69(33): p. 1122–1126. [PubMed: 32817602]
15. Mahajan UV and Larkins-Pettigrew M, Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties. J Public Health (Oxf), 2020. 42(3): p. 445–447. [PubMed: 32435809]
16. Pfefferle S, et al. , The SARS-coronavirus-host interactome: identification of cyclophilins as target for pan-coronavirus inhibitors. PLoS Pathog, 2011. 7(10): p. e1002331. [PubMed: 22046132]
17. Jager S, et al. , Global landscape of HIV-human protein complexes. Nature, 2011. 481(7381): p. 365–70. [PubMed: 22190034]
18. Batra J, et al. , Protein Interaction Mapping Identifies RBBP6 as a Negative Regulator of Ebola Virus Replication. Cell, 2018. 175(7): p. 1917–1930 e13. [PubMed: 30550789]
19. Shah PS, et al. , Comparative Flavivirus-Host Protein Interaction Mapping Reveals Mechanisms of Dengue and Zika Virus Pathogenesis. Cell, 2018. 175(7): p. 1931–1945 e18. [PubMed: 30550790]
20. Gordon DE, et al. , A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature, 2020. 583(7816): p. 459–468. [PubMed: 32353859]
21. Niemann HH, et al. , Structure of the human receptor tyrosine kinase met in complex with the Listeria invasion protein InlB. Cell, 2007. 130(2): p. 235–46. [PubMed: 17662939]
22. Hoffmann M, et al. , SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. Cell, 2020. 181(2): p. 271–280 e8. [PubMed: 32142651]
23. Xu GG, Guo J, and Wu Y, Chemokine receptor CCR5 antagonist maraviroc: medicinal chemistry and clinical applications. Curr Top Med Chem, 2014. 14(13): p. 1504–14. [PubMed: 25159165]

24. Hayouka Z, et al. , Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. Proc Natl Acad Sci U S A, 2007. 104(20): p. 8316–21. [PubMed: 17488811]
25. Peat TS, et al. , Small molecule inhibitors of the LEDGF site of human immunodeficiency virus integrase identified by fragment screening and structure based design. PLoS One, 2012. 7(7): p. e40147. [PubMed: 22808106]
26. Maginnis MS, Virus-Receptor Interactions: The Key to Cellular Invasion. J Mol Biol, 2018. 430(17): p. 2590–2611. [PubMed: 29924965]
27. Daczkowski CM, et al. , Structural Insights into the Interaction of Coronavirus Papain-Like Proteases and Interferon-Stimulated Gene Product 15 from Different Species. J Mol Biol, 2017. 429(11): p. 1661–1683. [PubMed: 28438633]
28. Yao J, et al. , Mechanism of inhibition of retromer transport by the bacterial effector RidL. Proc Natl Acad Sci U S A, 2018. 115(7): p. E1446–E1454. [PubMed: 29386389]
29. Zhang L, et al. , Solution structure of the complex between poxvirus-encoded CC chemokine inhibitor vCCI and human MIP-1beta. Proc Natl Acad Sci U S A, 2006. 103(38): p. 13985–90. [PubMed: 16963564]
30. Jonker HR, et al. , Structural properties of the promiscuous VP16 activation domain. Biochemistry, 2005. 44(3): p. 827–39. [PubMed: 15654739]
31. Card GL, et al. , Crystal structure of a gamma-herpesvirus cyclin-cdk complex. EMBO J, 2000. 19(12): p. 2877–88. [PubMed: 10856233]
32. Smith M, Honce R, and Schultz-Cherry S, Metabolic Syndrome and Viral Pathogenesis: Lessons from Influenza and Coronaviruses. J Virol, 2020. 94(18).
33. Vidal M, A unifying view of 21st century systems biology. FEBS Lett, 2009. 583(24): p. 3891–4. [PubMed: 19913537]
34. Robinson CV, Sali A, and Baumeister W, The molecular sociology of the cell. Nature, 2007. 450(7172): p. 973–82. [PubMed: 18075576]
35. Barabasi AL, Gulbahce N, and Loscalzo J, Network medicine: a network-based approach to human disease. Nat Rev Genet, 2011. 12(1): p. 56–68. [PubMed: 21164525]
36. Scott DE, et al. , Small molecules, big targets: drug discovery faces the protein-protein interaction challenge. Nat Rev Drug Discov, 2016. 15(8): p. 533–50. [PubMed: 27050677]
37. Arkin MR, Tang Y, and Wells JA, Small-molecule inhibitors of protein-protein interactions: progressing toward the reality. Chem Biol, 2014. 21(9): p. 1102–14. [PubMed: 25237857]
38. Rooklin D, et al. , AlphaSpace: Fragment-Centric Topographical Mapping To Target Protein-Protein Interaction Interfaces. J Chem Inf Model, 2015. 55(8): p. 1585–99. [PubMed: 26225450]
39. Lampson BL and Davids MS, The Development and Current Use of BCL-2 Inhibitors for the Treatment of Chronic Lymphocytic Leukemia. Curr Hematol Malig Rep, 2017. 12(1): p. 11–19. [PubMed: 28116634]
40. VENCLEXTA combination regimens for CLL work through 2 distinct cytotoxic mechanisms of action. 2019; Available from: https://www.venclextahcp.com/cll/venclexta-efficacy/mechanism-of-action.html.
41. Schormann N, et al. , Identification of protein-protein interaction inhibitors targeting vaccinia virus processivity factor for development of antiviral agents. Antimicrob Agents Chemother, 2011. 55(11): p. 5054–62. [PubMed: 21844323]
42. White PW, et al. , Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1-E2 protein interaction. J Biol Chem, 2003. 278(29): p. 26765–72. [PubMed: 12730224]
43. Goudreau N, et al. , Optimization and determination of the absolute configuration of a series of potent inhibitors of human papillomavirus type-11 E1-E2 protein-protein interaction: a combined medicinal chemistry, NMR and computational chemistry approach. Bioorg Med Chem, 2007. 15(7): p. 2690–700. [PubMed: 17306550]
44. Brito AF and Pinney JW, Protein-Protein Interactions in Virus-Host Systems. Front Microbiol, 2017. 8: p. 1557. [PubMed: 28861068]
45. Meyer MJ, et al. , Interactome INSIDER: a structural interactome browser for genomic studies. Nat Methods, 2018. 15(2): p. 107–114. [PubMed: 29355848]

46. Dominguez C, Boelens R, and Bonvin AM, HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc, 2003. 125(7): p. 1731–7. [PubMed: 12580598]
47. van Zundert GCP, et al. , The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol, 2016. 428(4): p. 720–725. [PubMed: 26410586]
48. Chaudhury S, Lyskov S, and Gray JJ, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics, 2010. 26(5): p. 689–91. [PubMed: 20061306]
49. Kirchdoerfer RN, et al. , Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. Sci Rep, 2018. 8(1): p. 15701. [PubMed: 30356097]
50. Wang Q, et al. , Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell, 2020. 181(4): p. 894–904 e9. [PubMed: 32275855]
51. Wrobel AG, et al. , SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. Nat Struct Mol Biol, 2020. 27(8): p. 763–767. [PubMed: 32647346]
52. Walls AC, et al. , Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell, 2020. 181(2): p. 281–292 e6. [PubMed: 32155444]
53. Alford RF, et al. , The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J Chem Theory Comput, 2017. 13(6): p. 3031–3048. [PubMed: 28430426]
54. Shang J, et al. , Structural basis of receptor recognition by SARS-CoV-2. Nature, 2020. 581(7807): p. 221–224. [PubMed: 32225175]
55. Wrapp D, et al. , Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science, 2020. 367(6483): p. 1260–1263. [PubMed: 32075877]
56. Jordan RE and Adab P, Who is most likely to be infected with SARS-CoV-2? The Lancet Infectious Diseases, 2020. 20(9): p. 995–996. [PubMed: 32422197]
57. Cao Y, et al. , Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. Cell Discov, 2020. 6: p. 11. [PubMed: 32133153]
58. Darbeheshti F and Rezaei N, Genetic predisposition models to COVID-19 infection. Med Hypotheses, 2020. 142: p. 109818. [PubMed: 32416414]
59. Zhao Y, et al., 2020.
60. Kortemme T and Baker D, A simple physical model for binding energy hot spots in protein-protein complexes. Proc Natl Acad Sci U S A, 2002. 99(22): p. 14116–21. [PubMed: 12381794]
61. Shulman-Peleg A, et al. , Spatial chemical conservation of hot spot interactions in protein-protein complexes. BMC Biol, 2007. 5: p. 43. [PubMed: 17925020]
62. Karczewski KJ, et al. , The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 2020. 581(7809): p. 434–443. [PubMed: 32461654]
63. Stawiski EW, et al. , Human ACE2 receptor polymorphisms predict SARS-CoV-2 susceptibility. bioRxiv, 2020.
64. Procko E, The sequence of human ACE2 is suboptimal for binding the S spike protein of SARS coronavirus 2. bioRxiv, 2020.
65. Berman HM, et al. , The Protein Data Bank. Nucleic Acids Res, 2000. 28(1): p. 235–42. [PubMed: 10592235]
66. Pieper U, et al. , ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res, 2014. 42(Database issue): p. D336–46. [PubMed: 24271400]
67. Guharoy M and Chakrabarti P, Conserved residue clusters at protein-protein interfaces and their use in binding site identification. BMC Bioinformatics, 2010. 11: p. 286. [PubMed: 20507585]
68. Gupta R, et al. , SARS-CoV2 (COVID-19) Structural/Evolution Dynamicome: Insights into functional evolution and human genomics. bioRxiv, 2020.
69. Stenson PD, et al. , Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat, 2003. 21(6): p. 577–81. [PubMed: 12754702]

70. Landrum MJ, et al. , ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res, 2018. 46(D1): p. D1062–D1067. [PubMed: 29165669]
71. Buniello A, et al. , The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res, 2019. 47(D1): p. D1005–D1012. [PubMed: 30445434]
72. Killerby ME, et al. , Characteristics Associated with Hospitalization Among Patients with COVID-19 - Metropolitan Atlanta, Georgia, March-April 2020. MMWR Morb Mortal Wkly Rep, 2020. 69(25): p. 790–794. [PubMed: 32584797]
73. Yang J, et al. , Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. Int J Infect Dis, 2020. 94: p. 91–95. [PubMed: 32173574]
74. Sahni N, et al. , Widespread macromolecular interaction perturbations in human genetic disorders. Cell, 2015. 161(3): p. 647–660. [PubMed: 25910212]
75. Wang X, et al. , Three-dimensional reconstruction of protein networks provides insight into human genetic disease. Nat Biotechnol, 2012. 30(2): p. 159–64. [PubMed: 22252508]
76. Sim NL, et al. , SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res, 2012. 40(Web Server issue): p. W452–7. [PubMed: 22689647]
77. Adzhubei IA, et al. , A method and server for predicting damaging missense mutations. Nat Methods, 2010. 7(4): p. 248–9. [PubMed: 20354512]
78. Yu H, et al. , High-quality binary protein interaction map of the yeast interactome network. Science, 2008. 322(5898): p. 104–10. [PubMed: 18719252]
79. Braun P, et al. , An experimentally derived confidence score for binary protein-protein interactions. Nat Methods, 2009. 6(1): p. 91–7. [PubMed: 19060903]
80. Vo TV, et al. , A Proteome-wide Fission Yeast Interactome Reveals Network Evolution Principles from Yeasts to Human. Cell, 2016. 164(1–2): p. 310–323. [PubMed: 26771498]
81. Luck K, et al. , A reference map of the human binary protein interactome. Nature, 2020. 580(7803): p. 402–408. [PubMed: 32296183]
82. Gordon DE, et al. , Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. Science, 2020. 370(6521).
83. Fragoza R, et al. , Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. Nat Commun, 2019. 10(1): p. 4141. [PubMed: 31515488]
84. Fellay J, et al. , A whole-genome association study of major determinants for host control of HIV-1. Science, 2007. 317(5840): p. 944–7. [PubMed: 17641165]
85. Wei X, et al. , A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. PLoS Genet, 2014. 10(12): p. e1004819. [PubMed: 25502805]
86. Hong HQ, et al. , G3BP2 is involved in isoproterenol-induced cardiac hypertrophy through activating the NF-kappaB signaling pathway. Acta Pharmacol Sin, 2018. 39(2): p. 184–194. [PubMed: 28816235]
87. Lu S, et al. , The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. Nat Commun, 2021. 12(1): p. 502. [PubMed: 33479198]
88. Nabeel-Shah S, et al. , Nucleus-specific linker histones Hho1 and Mlh1 form distinct protein interactions during growth, starvation and development in Tetrahymena thermophila. Sci Rep, 2020. 10(1): p. 168. [PubMed: 31932604]
89. Koes DR, Baumgartner MP, and Camacho CJ, Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. J Chem Inf Model, 2013. 53(8): p. 1893–904. [PubMed: 23379370]
90. Wierbowski SD, et al. , Cross-docking benchmark for automated pose and ranking prediction of ligand binding. Protein Sci, 2020. 29(1): p. 298–305. [PubMed: 31721338]
91. Crouse AB, et al. , Metformin Use Is Associated With Reduced Mortality in a Diverse Population With COVID-19 and Diabetes. Front Endocrinol (Lausanne), 2020. 11: p. 600439. [PubMed: 33519709]

92. Silmitasertib (CX-4945) in Patients With Severe Coronavirus Disease 2019 (COVID-19) (CX4945). 2020; Available from: https://clinicaltrials.gov/ct2/show/NCT04668209.

93. Valproate Alone or in Combination With Quetiapine for Severe COVID-19 Pneumonia With Agitated Delirium. 2020; Available from: https://clinicaltrials.gov/ct2/show/NCT04513314.

94. McIver EG, et al. , Synthesis and structure-activity relationships of a novel series of pyrimidines as potent inhibitors of TBK1/IKKepsilon kinases. Bioorg Med Chem Lett, 2012. 22(23): p. 7169–73. [PubMed: 23099093]

95. Jumper J, et al. , Highly accurate protein structure prediction with AlphaFold. Nature, 2021.

96. Tunyasuvunakool K, et al. , Highly accurate protein structure prediction for the human proteome. Nature, 2021.

97. Baek M, et al. , Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021. 373(6557): p. 871–876. [PubMed: 34282049]

98. Gainza P, et al. , Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nat Methods, 2020. 17(2): p. 184–192. [PubMed: 31819266]

99. Eswar N, et al. , Comparative protein structure modeling using Modeller. Curr Protoc Bioinformatics, 2006. Chapter 5: p. Unit-5 6.

100. Mosca R, Ceol A, and Aloy P, Interactome3D: adding structural details to protein networks. Nat Methods, 2013. 10(1): p. 47–53. [PubMed: 23399932]

101. Altschul SF, et al. , Basic local alignment search tool. J Mol Biol, 1990. 215(3): p. 403–10. [PubMed: 2231712]

102. The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.

103. Gasteiger E, et al., Protein Identification and Analysis Tools on the ExPASy Server. (In) Walker John M. (ed): The Proteomics Protocols Handbook, Humana Press (2005), 2005: p. 571–607.

104. Lin JH, Divergence Measures Based on the Shannon Entropy. Ieee Transactions on Information Theory, 1991. 37(1): p. 145–151.

105. Capra JA and Singh M, Predicting functionally important residues from sequence conservation. Bioinformatics, 2007. 23(15): p. 1875–82. [PubMed: 17519246]

106. Morcos F, et al. , Direct coupling analysis for protein contact prediction. Methods Mol Biol, 2014. 1137: p. 55–70. [PubMed: 24573474]

107. Lockless SW and Ranganathan R, Evolutionarily conserved pathways of energetic connectivity in protein families. Science, 1999. 286(5438): p. 295–9. [PubMed: 10514373]

108. Lee B and Richards FM, The interpretation of protein structures: estimation of static accessibility. J Mol Biol, 1971. 55(3): p. 379–400. [PubMed: 5551392]

109. Pierce BG, Hourai Y, and Weng Z, Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS One, 2011. 6(9): p. e24657. [PubMed: 21949741]

110. Rodrigues JP, et al. , Defining the limits of homology modeling in information-driven protein docking. Proteins, 2013. 81(12): p. 2119–28. [PubMed: 23913867]

111. UniProt C, UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res, 2019. 47(D1): p. D506–D515. [PubMed: 30395287]

112. He R, et al. , Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochem Biophys Res Commun, 2004. 316(2): p. 476–83. [PubMed: 15020242]

113. Wu F, et al. , A new coronavirus associated with human respiratory disease in China. Nature, 2020. 579(7798): p. 265–269. [PubMed: 32015508]

114. Chan JF, et al. , Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect, 2020. 9(1): p. 221–236. [PubMed: 31987001]

115. Needleman SB and Wunsch CD, A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, 1970. 48(3): p. 443–53. [PubMed: 5420325]

116. Henikoff S and Henikoff JG, Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A, 1992. 89(22): p. 10915–9. [PubMed: 1438297]

117. McLaren W, et al. , The Ensembl Variant Effect Predictor. Genome Biol, 2016. 17(1): p. 122. [PubMed: 27268795]

118. Szumilas M, Explaining odds ratios. J Can Acad Child Adolesc Psychiatry, 2010. 19(3): p. 227–9. [PubMed: 20842279]

119. Machiela MJ and Chanock SJ, LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics, 2015. 31(21): p. 3555–7. [PubMed: 26139635]

120. O'Boyle NM, et al. , Open Babel: An open chemical toolbox. J Cheminform, 2011. 3: p. 33. [PubMed: 21982300]

121. Yang X, et al. , A public genome-scale lentiviral expression library of human ORFs. Nat Methods, 2011. 8(8): p. 659–61. [PubMed: 21706014]

122. Das J, et al. , Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. Sci Signal, 2013. 6(276): p. ra38. [PubMed: 23695164]

123. 3D_SARS2: 3D_SARS2 Analysis Code (10.5281/zenodo.4987957) 2021.

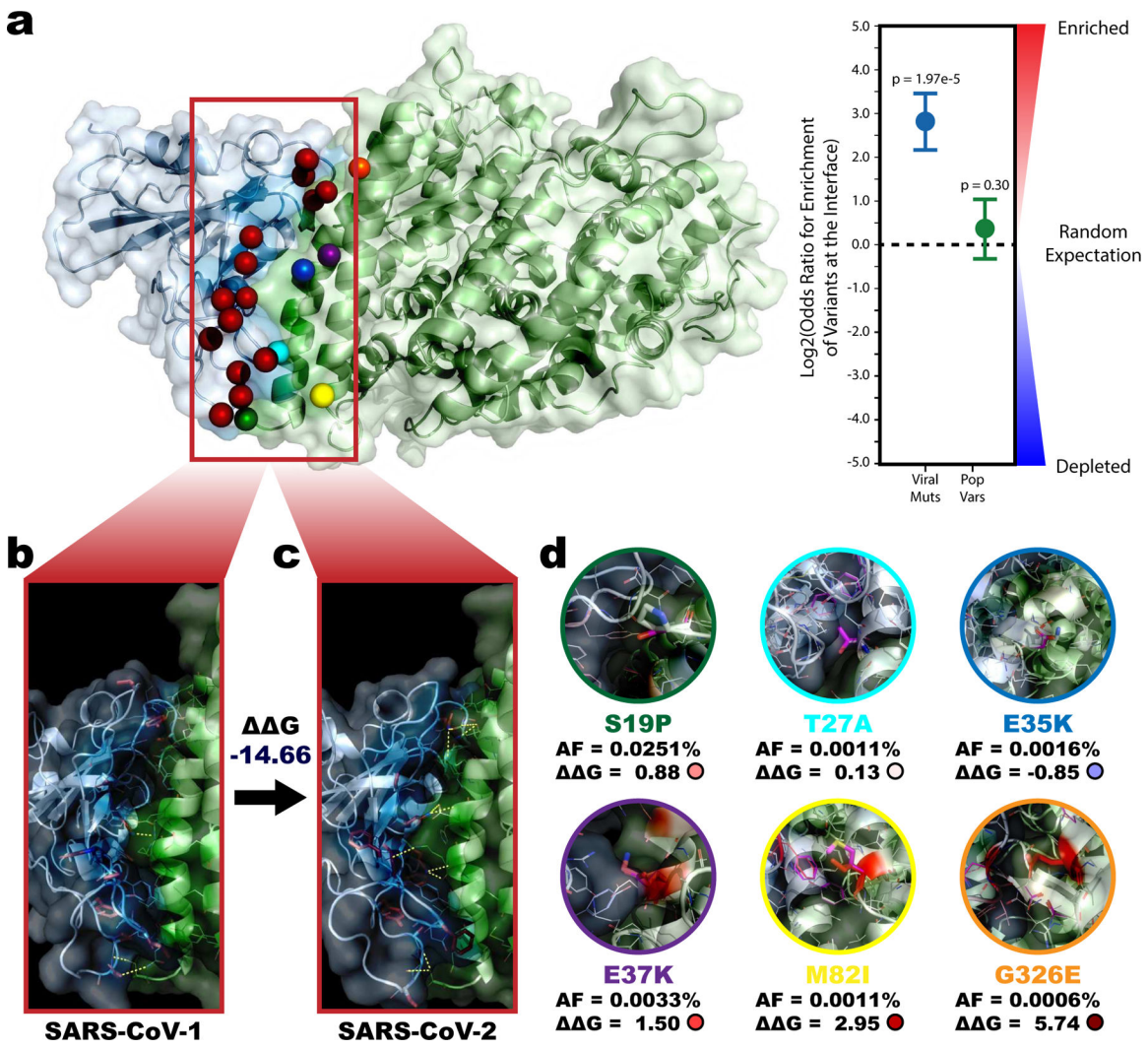


Figure 1. Enrichment and predicted impact of divergences between SARS-CoV-1 and SARS-CoV-2 along the S-ACE2 interface.

a, Co-crystal structure of the interaction between SARS-CoV-2 Spike protein (S) with human ACE2 (PDB 6LZG). All 15 sequence divergences between SARS-CoV-1 and SARS-CoV-2 Spike protein interfaces are highlighted as red spheres while all 6 population variants on the ACE2 protein interface are highlighted as green (ACE2_S19P), cyan (ACE2_T27A), blue (ACE2_E35K), purple (ACE2_E37K), yellow (ACE2_M82I), and orange (ACE2_G326E) spheres. Enrichment of these variants on the interface are reported for SARS-CoV-2 (Log2OR=2.82, p=1.97e-5 by two-sided z-test) and human (Log2OR=0.38, p=0.30 by two-sided z-test) shown to the right. Data presented as Log2OR ± SE. **b, c**, Expanded interface views for the SARS-CoV-1 S-ACE2 structure (PDB 6CS2) and SARS-CoV-2 S-ACE2 structure (PDB 6LZG). Sequence divergences are highlighted as red sticks. Inter-protein polar contacts that contribute to stabilizing the interaction are shown as yellow dashed lines. The negative predicted change in binding affinity (ΔΔG=−14.66 Rosetta Energy Units (REU)) indicates the interaction is more stable (lower energy) in the SARS-CoV-2 version of the interaction. **d**, Predicted impact each ACE2 population variant. Mutated structures superimposed over the wildtype structure (magenta). The mutated

residue is shown as sticks. Residues contributing to the overall change in binding energy are colored from blue (decreased ΔΔG) to white (no change) to red (increased ΔΔG). The gnomAD reported allele frequency and predicted ΔΔG for each mutation are reported.

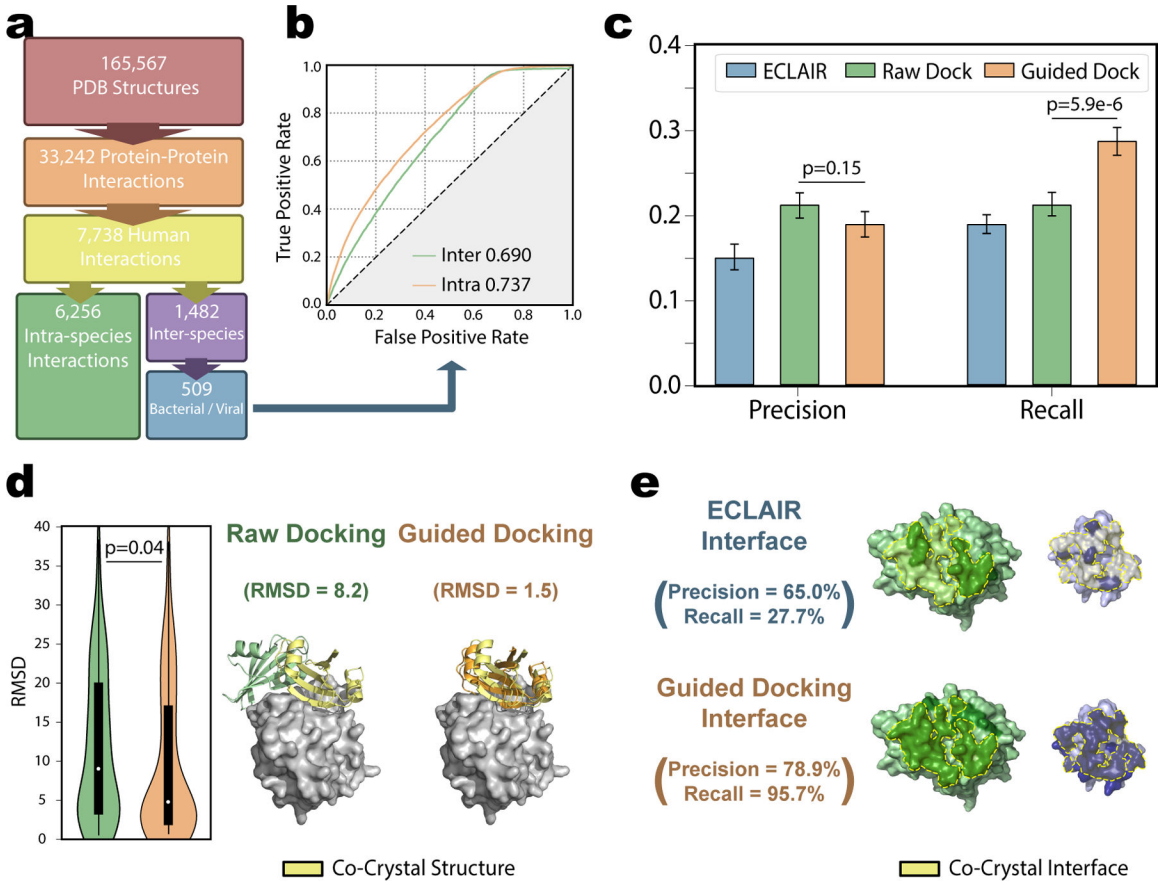


Figure 2. Validation of ECLAIR and Guided Docking Performance.

a, Steps taken to parse the Protein Data Bank and construct our human-pathogen PDB benchmark set. **b**, Comparison of ECLAIR performance on intra-species interactions (n=200 human-human interactions) against inter-species interactions (n=509 human-pathogen interactions). Area under the receiver operating characteristic (AUROC) evaluation indicates considerable predictive power is achieved in both tasks, (intra-species AUROC=0.737, inter-species AUROC=0.690). **c**, Comparison of final interface predictions across all residues in 153 dockable human-pathogen interactions using either ECLAIR (precision=0.15, recall=0.19), a raw docking HADDOCK protocol (precision=0.21, recall=0.21), or our guided docking HADDOCK protocol implementing ECLAIR predictions as restraints (precision=0.19, recall=0.29). Recall from guided docking significantly outperformed raw docking method (p=5.88e-6 by two-sided two proportion z-test) without sacrificing precision (p=0.15 by two-sided two proportion z-test). Data presented as precision or recall ± SD as estimated by 1000-fold bootstrapping sampling 153 interactions and interface predicitons with replacement each iteration. **d**, Distributions of root-mean-square deviation (RMSD) between the top-scored raw or guided docking output and the co-crystal structure (n=153 dockable human-pathogen interactions). Interior boxplots represent the distribution quartiles with whiskers representing the most extreme non-outlier values. Average RMSD from the guided docking (average RMSD=9.45) was significantly lower than the raw docking (average RMSD=11.79) based on two-sided t-test (p=0.04). To the right, an example where the guided docking accurately identifies the correct interaction orientation

missed by the raw docking (Human protein shown as gray surface, raw docking, guided docking, and co-crystal structure viral protein shown as green, orange, and yellow cartoon respectively). **e**, Example showing a best case scenario where a few true interface residues predicted by ECLAIR (top, recall=27.7%) are successfully expanded to identify the rest of the interface by the guided docking (bottom, recall=95.7%). Human and viral proteins shown to the left in green and to the right in blue respectively. Residues identified as interface in each approach are darkened. True interface from the co-crystal structure outlined and shaded in yellow.

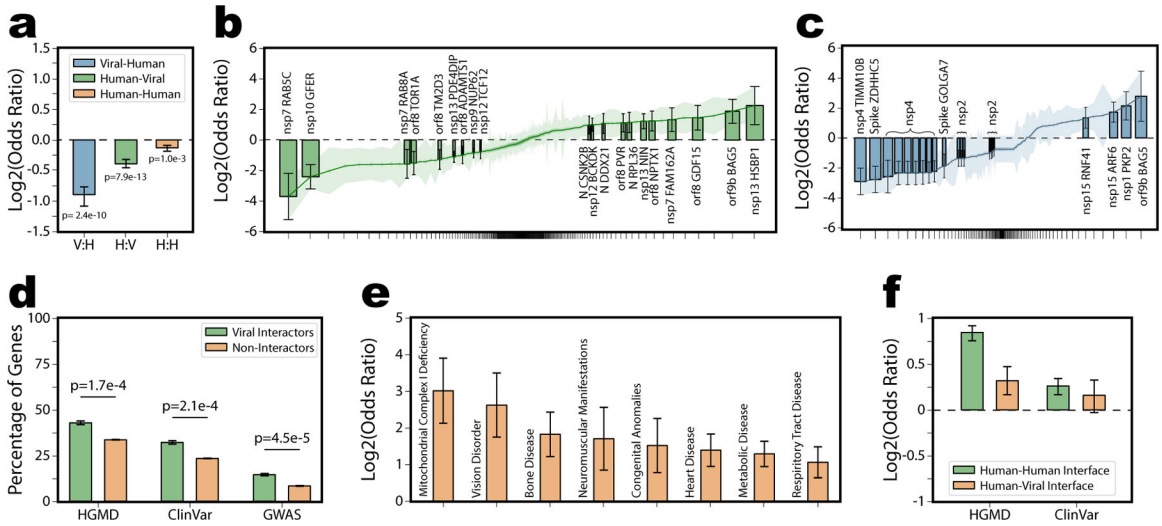


Figure 3. Enrichment of sequence divergences and disease mutations across all SARS-CoV-2-Human interaction interfaces.

a, Enrichment across 332 human genes interacting with SARS-CoV-2 for viral sequence divergence or human population variants along viral-human (V:H, Log2OR=−0.91, p=2.41e-10 by two-sided z-test) human-viral (H:V, Log2OR=−0.38, p=7.92e-13 by two-sided z-test) or human-human (H:H, Log2OR=−0.14, p=9.98e-4 by two-sided z-test) interfaces. Data presented as Log2OR ± SE. **b, c**, Individual enrichments (sorted from most depleted to most enriched) for human population variants and viral sequence divergences respectively on all 332 SARS-CoV-2-human interaction interfaces. Interfaces with statistically significant Log2OR (by two-sided z-test) are labeled and shown as bars, the remainder plotted as a line. Data presented as Log2OR ± SE. Clusters of SARS-CoV-2 enrichments involving the nsp4 interactions with (IDE, NUP210, DNAJC11, TIMM29, TIMM9, and TIMM10) and nsp2 interactions with (GIGYF2, FKBP15, WASHC4, EIF4E2, POR, and SLC27A2) were labeled as a group for legibility. **d**, Percentage of human genes that interact with (green, n=332) or do not interact with (orange, n=20,018) SARS-CoV-2 that contain disease annotations in HGDM (Log2OR=0.57, p=1.70e-4 by two-sided z-test), ClinVar (Log2OR=0.64, p=1.05e-4 by two-sided z-test), and GWAS (Log2OR=0.89, p=4.54e-5 by two-sided z-test) respectively. Genes targeted by SARS-CoV-2 proteins were significantly more likely to harbor disease mutations than non-interactors by log odds enrichment test. Data presented as percentage ± SE. **e**, Sample of individual disease terms enriched in human genes targeted by SARS-CoV-2. Full results reported in Supplemental Table 6. Data presented as Log2OR ± SE. **f**, Comparison of the enrichment of HGDM or ClinVar annotated mutations on human-vial interfaces or human-human interfaces for 332 genes interacting with SARS-CoV-2. Disease mutations enriched on human-human interfaces (HGMD, Log2OR=0.84, p<1e-20 by two-sided z-test; ClinVar, Log2OR=0.25, p=2.9e-3 by two-sided z-test), while human-viral interface show no or marginal enrichment (HGMD, Log2OR=0.31, p=0.048 by two-sided z-test; ClinVar, Log2OR=0.15, p=0.39 by two-sided z-test). GWAS category excluded from this analysis because most lead GWAS SNPs occur in non-coding regions. Data presented as Log2OR ± SE.

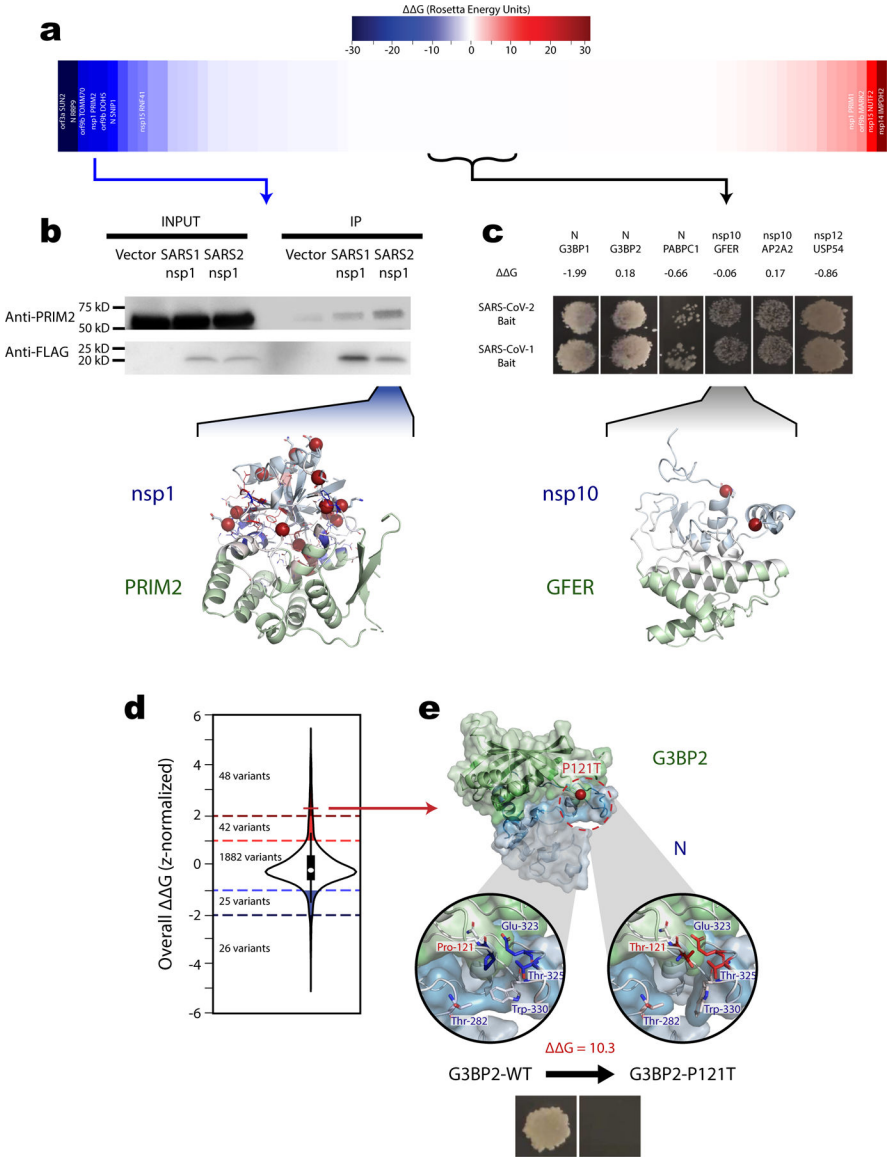


Figure 4. Predicted impact of sequence divergences on the binding affinity of SARS-CoV-2-Human interactions.

a, Predicted impact of SARS-CoV-1 to SARS-CoV-2 sequence divergences on binding affinity from docked structure for 83 applicable SARS-CoV-2-human interactions sorted from largest decrease (most stabilized relative to SARS-CoV-2) to largest increase (most destabilized relative to SARS-CoV-1) (mean=−0.57 REU, std=5.78 REU). Interaction labels shown wherever predicted ΔΔG exceeds mean ± 1 std. **b**, Representative cropped western blots (among 3 replicates) from co-immunoprecipitation (co-IP) comparing the interaction between human PRIM2 with SARS-CoV-1 or SARS-CoV-2 nsp1. More efficient PRIM2 pull down with SARS-CoV-2 bait validates the PRIM2-nsp1 ΔΔG prediction (ΔΔG=−17.3 REU, z-score=−2.9). Shown below, docked structure for PRIM2 with SARS-CoV-2 nps1 (green and blue cartoon respectively). SARS-CoV-1 to SARS-CoV-2 sequence divergences represented as spheres. Interface residues colored relative to overall ΔΔG contribution ranging from blue (more stabilizing in SARS-CoV-2) to white (little impact on ΔΔG), to

red (more stabilizing in SARS-CoV-1). Residue side chains shown as sticks in regions with high local ΔΔG. **c**, Representative Y2H results (among 3 replicates) confirming that 6 interactions with no predicted ΔΔG values can be detected using either SARS-CoV-2 or SASR-CoV viral protein as bait. The docked structure (visualized as in **b**) for human GFER and SARS-CoV-2 nsp10 (ΔΔG=−0.06) shown to highlight that sequence divergences in these 6 interactions did not localize near the interface. **d**, Distribution of the predicted changes in binding affinity from scanning mutagenesis for all 2,023 human population variants on SARS-CoV-2-human interfaces. Values were z-score normalized across for each residue type and on each interface. Shaded regions indicate putative interface binding energy hotspots annotated as strongly disruptive (z-score ≥ 2, 48 total variants), disruptive (1 ≤ z-score < 2, 42 total variants), stabilizing (−2 < z-score ≤ −1, 25 total variants), or strongly stabilizing (z-score ≤ −2, 26 total variants). Interior boxplot represents the distribution quartiles with whiskers representing the most extreme non-outlier values. **e**, Docked structure between SARS-CoV-2 N protein and human G3BP2, alongside expanded interface views comparing the wildtype interface (left) with a predicted strongly disruptive (ΔΔG=10.3 REU, z-score=2.3) population variant, G3BP2_P121T (right). Shown below, yeast two-hybrid results confirmed that the G3BP2_P121T variant completely disrupts the G3BP2-N interaction.

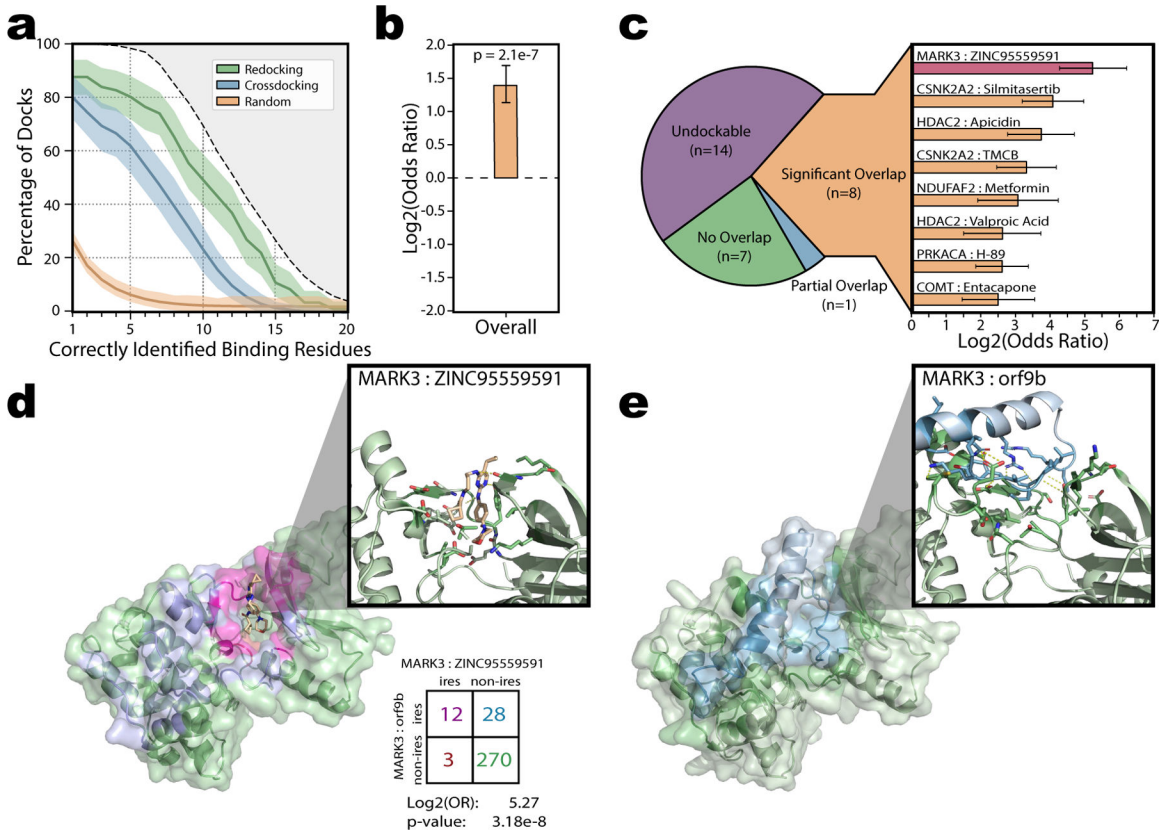


Figure 5. Drug Docking and Prioritization of SARS-CoV-2-Human Interaction Inhibitors.

a, Validation of smina's ability to identify the correct binding site from the full protein surface based on 4,399 drug-ligand pairs across 95 protein targets. Docking was carried out either by redocking each ligand back into its native protein structure, or cross-docking each ligand into a representative receptor structure. Baseline performance expectation derived from random selection of surface patches matching the size of the correct binding site is shown for comparison. Each line and shaded area indicates the percentage of docks that correctly identify X binding site residues ± SD as estimated by 1000-fold bootstrapping sampling 95 drug-target pairs with replacement each iteration. The gray shaded area at the top indicates the maximum fraction of docks whose true binding sites contain at least X residues. **b**, Protein-protein and protein-drug binding sites pooled across 16 applicable drug-target pairs were significantly enriched (Log2OR=1.38, p=2.1e-7 by two-sided z-test). Data presented as Log2OR ± SE. **c**, Individual breakdown of the overlap between the each of the protein-protein and protein-drug binding sites as either undockable (i.e. no protein-protein docked structure available for comparison; 14 total), no overlap (7 total), partial overlap (1 total) or significant overlap (8 total). The individual log2(Odds Ratios) for each of the significant drug-target pairs are shown. Data presented as Log2OR ± SE. The MARK3-ZINC95559591 pair (shown in **d**) is highlighted red. **d**, Docked structure for ZINC95559591 bound to human MARK3. MARK3 surface is colored either green (non-interface, n=270), blue (orf9b interface, n=28), red (ZINC95559591 interface, n=3), or magenta (shared interface, n=12). Cut-out display highlights the MARK3- ZINC95559591

binding site. Polar contacts between MARK3 and ZINC95559591 shown as dashed lines. **e**, Corresponding docked structure for SARS-CoV-2 orf9b bound to human MARK3.

Figure 6. 3D-SARS2 Structural Interactome Browser Overview.

Overview of the main results page for exploring a given interaction in our 3D-SARS2 structural interactome browser. The main display contains information for both the SARS-CoV-2 and human proteins including structural displays for either the docked or single crystal structures as well as a table summarizing the interface residues for both proteins. Interface residues are colored dark blue and dark green for the viral and human proteins respectively. By default the page will display the docked structure if available. The display can be toggled between docked structures and single structures using the button in the bottom middle. When single structures display is selected residues will instead be colored based on the initial ECLAIR interface definition. Four categories of expandable panels containing additional analyses are provided. **upper left**, The interface view shows a linear representation of the protein sequence with interface residues annotated in dark blue or dark green. Interfaces for other interactors of the protein are shown underneath for easy comparison. **upper right**, The mutations panel summarizes either human population variants or viral sequence divergences on the protein. Mutations on the interface are labeled. **lower left**, The ΔΔG information panel summarizes the results from in-silico mutagenesis scanning along the interface. Results for each mutation are z-score normalized relative to the rest of the interface and colored on a blue (negative ΔΔG, stabilizing) to yellow (minimal impact) to red (positive ΔΔG, destabilizing). The heatmap can be filtered to only show values corresponding to known mutations on the interface. **lower right**, The candidate drugs panel shows docking information for any known drug targets of the human protein.