nature methods

RESOURCE

https://doi.org/10.1038/s41592-021-01318-w

Check for updates

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] and Haiyuan Yu[1,2] ✉

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 three-dimensional 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.

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. These coronaviruses are either highly transmissible yet generally not highly pathogenic (for example HCoV-229E and 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 acute respiratory distress syndrome[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 (ref. [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]. Interspecies 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 the inhibitory effect of poxvirus 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]. 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[40] and human papilloma virus therapies[41,42]. 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 a US Food and Drug Administration-approved treatment for HIV-1 that functions by blocking the interface for this interaction[23,43].

Here we apply a full-interactome modeling framework to construct a three-dimensional (3D) structural interactome between SARS-CoV-2 and human proteins. Our framework first applies our previous ECLAIR framework[44] to identify interface residues

[1]Department of Computational Biology, Cornell University, Ithaca, NY, USA. [2]Weill Institute for Cell and Molecular Biology, Cornell University, Ithaca, NY, USA. [3]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, USA. [4]Department of Microbiology and Immunology, Cornell University, Ithaca, NY, USA. [5]Department of Medicine, Weill-Cornell Medicine, New York, NY, USA. [6]Master of Public Health Program, Cornell University, Ithaca, NY, USA. ✉e-mail: haiyuan.yu@cornell.edu

for the whole SARS-CoV-2–human interactome and leverages these predictions to guide atomic-resolution interface modeling and docking in HADDOCK[45,46]. We additionally carried out in silico scanning mutagenesis in PyRosetta[47] 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. 1a). This interaction mediates viral entry into human cells[3] and is among the only viral–human interactions solved in both SARS-CoV-1 (ref. [48]) and SARS-CoV-2 (refs. [49–51]). Recent sequence divergences of the S protein are highly enriched at the S–ACE2 interaction interface (Fig. 1a; $\log_2$odds ratio (OR) = 2.82, $P = 1.97 \times 10^{-5}$), 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[52] (Fig. 1b,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 (ref. [53]). Experimental kinetics assays have confirmed that compared to SARS-CoV-1, SARS-CoV-2 S protein binds ACE2 with 10–20-fold higher affinity[54], supporting the conclusions from our computational modeling.

A wide range in severity of and susceptibility to SARS-CoV-2 exists between individuals[6,7,55]. Genetic predisposition hypotheses explaining this range include both expression-regulating and protein-coding variants[56,57]. 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[58]. 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[59,60] to predict the impact of six missense variants reported in gnomAD[61] that occur on the S–ACE2 interface (Fig. 1d). 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[62,63]. 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[20]. First, we modeled SARS-CoV-2 proteins supplementing solved structures from the Protein Data Bank (PDB)[64] (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 Fig. 1a), while comparison against the available SARS-CoV-2 PDB structures from the remaining 11 validated the quality of our modeling approach (Extended Data Fig. 1b,c). For human interactors all models were obtained from the PDB or Modbase[65] (Extended Data Fig. 2a). We then predicted the interface residues for each interaction using our ECLAIR framework[44]. 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[45,46] 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 Fig. 2b) 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 Supplementary Tables 1 and 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. 2a). The full list of interactions in this benchmark set alongside the PDB sources plus true and predicted interfaces are provided in Supplementary Table 3.

To validate ECLAIR's applicability to interspecies interactions, we compared its published performance test set of 200 human–human interactions to its performance on our human–pathogen PDB benchmark set. Both tasks achieved comparable performance (receiver operating characteristic area under the curve = 0.69 versus 0.74), although the intraspecies task slightly outperformed interspecies (Fig. 2b). We note that feature availability between sets (for instance, coevolution features can only be calculated for intraspecies interactions) may confound direct comparisons between different interaction sets. Overall, the evaluation of our benchmark conclusively shows that ECLAIR retains predictive power for interspecies 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 interspecies interactions, 153 fit our criteria for docking. We compared interface annotations from each protocol based on precision and recall (Fig. 2c). Overall interface quality was comparable between both raw and guided protocols (precision = 0.21 versus 0.19, $P = 0.15$), however, the guided

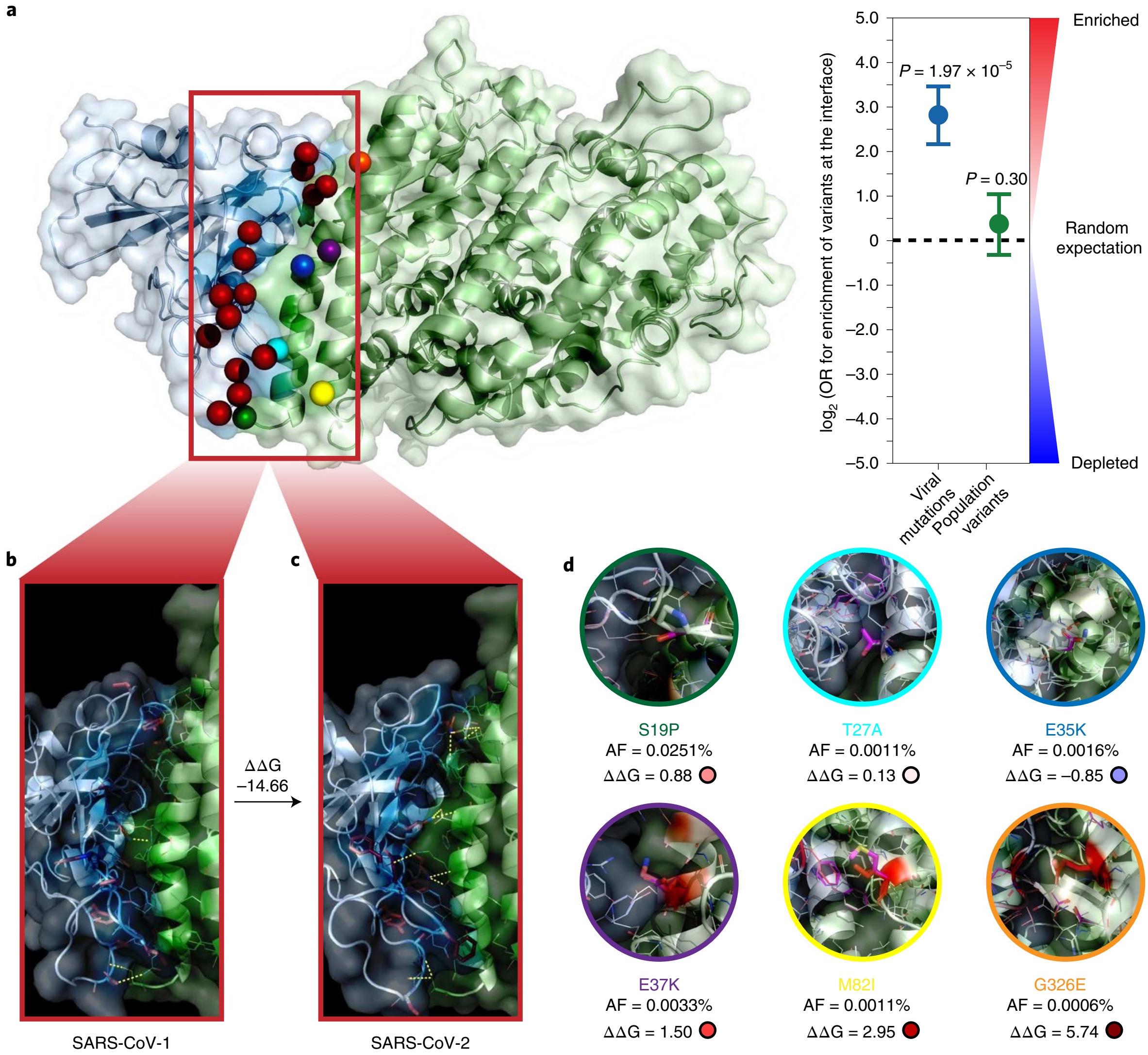


Fig. 1 | Enrichment and predicted impact of divergences between SARS-CoV-1 and SARS-CoV-2 along the S–ACE2 interface. a, Co-crystal structure of the interaction between SARS-CoV-2 S with human ACE2 (PDB 6LZG). All 15 sequence divergences between SARS-CoV-1 and SARS-CoV-2 S interfaces are highlighted as red spheres and 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 ($\log_2$OR = 2.82, $P = 1.97 \times 10^{-5}$ by two-sided z-test) and human ($\log_2$OR = 0.38, $P = 0.30$ by two-sided z-test) shown to the right. Data are presented as $\log_2$OR ± s.e.m. **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 ($\Delta\Delta G = -14.66$ REU) indicates the interaction is more stable (lower energy) in the SARS-CoV-2 version of the interaction. **d**, Predicted impact of each *ACE2* population variant. Mutated structures superimposed over the wild-type 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.

docking better recovered the total interface (recall = 0.21 versus 0.29, $P = 5.88 \times 10^{-6}$). 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 root-mean-square deviation (RMSD) between docked and reference structures, we further demonstrate that the guided docking better recapitulated the true co-crystal structures (Fig. 2d; average RMSD = 9.45 versus 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. 2e). Comparison of the precision and recall between ECLAIR and our guided docking (Fig. 2c) 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-reported[61] human population variants (Supplementary Table 4) and sequence divergences between SARS-CoV-1 and SARS-CoV-2 (Supplementary Table 5) onto predicted interfaces.

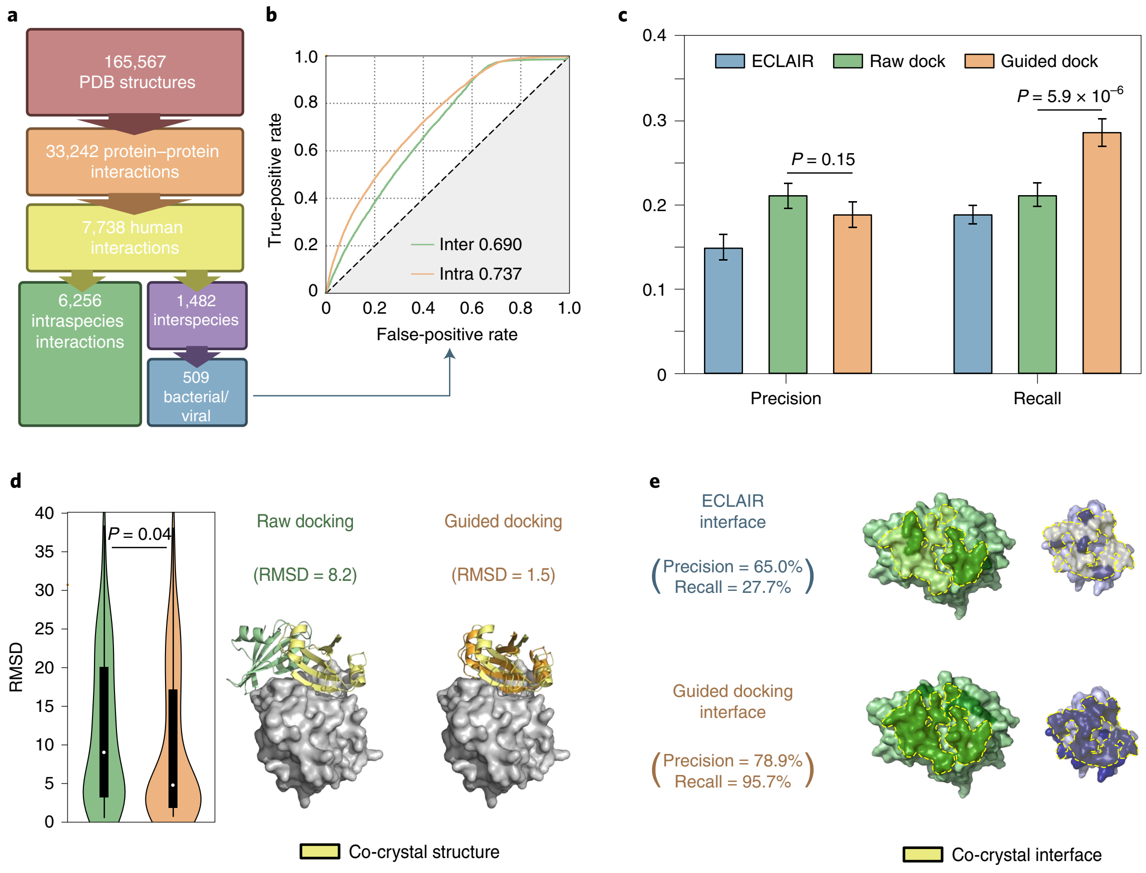


Fig. 2 | Validation of ECLAIR and guided docking performance. a, Steps taken to parse the PDB and construct our human–pathogen PDB benchmark set. **b**, Comparison of ECLAIR performance on intraspecies interactions ($n=200$ human–human interactions) against interspecies interactions ($n=509$ human–pathogen interactions). Area under the receiver operating characteristic (AUROC) evaluation indicates considerable predictive power is achieved in both tasks (intraspecies AUROC = 0.737 and interspecies AUROC = 0.690). **c**, Comparison of final interface predictions across all residues in 153 dockable human–pathogen interactions using either ECLAIR (precision = 0.15 and recall = 0.19), a raw docking HADDOCK protocol (precision = 0.21 and recall = 0.21) or our guided docking HADDOCK protocol implementing ECLAIR predictions as restraints (precision = 0.19 and recall = 0.29). Recall from guided docking significantly outperformed the raw docking method ($P=5.88\times10^{-6}$ by two-sided two proportion z-test) without losing precision ($P=0.15$ by two-sided two proportion z-test). Data are presented as precision or recall ± s.d. as estimated by 1,000-fold bootstrapping sampling 153 interactions and interface predictions with replacement each iteration. **d**, Distributions of RMSD between the top-scored raw or guided docking output and the co-crystal structure ($n=153$ dockable human–pathogen interactions). Interior box plots represent the distribution quartiles with whiskers representing the most extreme non-outlier values. Average RMSD from guided docking (average RMSD = 9.45) was significantly lower than raw docking (average RMSD = 11.79) based on a two-sided t-test ($P=0.04$). An example where the guided docking accurately identifies the correct interaction orientation missed by the raw docking (human protein shown by gray surface and raw docking, guided docking and co-crystal structure viral protein shown as green, orange and yellow cartoons, respectively) (right). **e**, Example showing a best-case scenario where a few true interface residues predicted by ECLAIR (recall = 27.7%) (top) are successfully expanded to identify the rest of the interface by the guided docking (recall = 95.7%) (bottom). Human and viral proteins shown in green (left) and in blue (right), respectively. Residues identified as an interface in each approach are darkened. True interfaces from the co-crystal structure are outlined and shaded in yellow.

Conserved residues generally cluster along protein–protein interfaces[66] and an analysis of SARS-CoV-2 structure and evolution similarly concluded highly conserved surface residues likely drove protein–protein interactions[67]. Consistent with these previous 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. 3a).

Nonetheless, considering each interaction individually, we identified 11 interaction interfaces enriched for human population variants (Fig. 3b) and 4 enriched for recent viral sequence divergences (Fig. 3c). Supplementary Table 6 provides the log odds enrichment values 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 new 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[68], ClinVar[69] or the NHGRI-EBI GWAS catalog[70]. Interactors of SARS-CoV-2 were enriched for phenotypic variants from each database (Fig. 3d). Notably, several of the individual disease categories enriched among interactors, were consistent with SARS-CoV-2 comorbidities, including heart

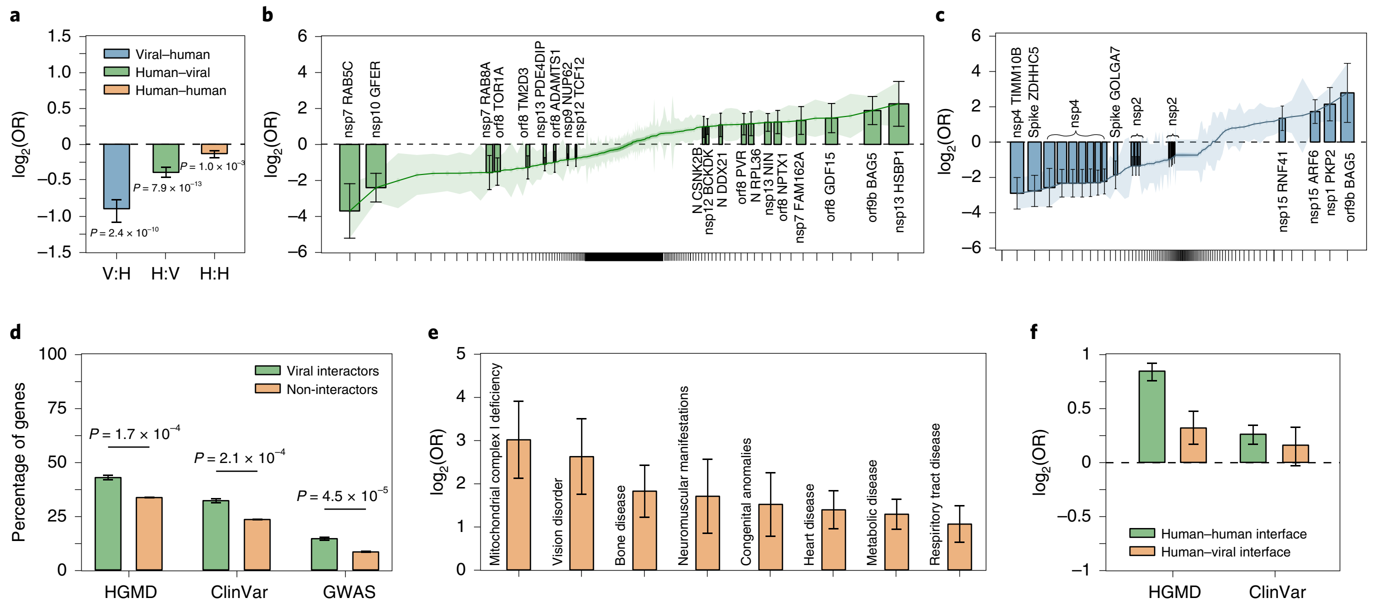


Fig. 3 | Enrichment of sequence divergences and disease mutations across all SARS-CoV-2–human interaction interfaces. a, Enrichment across 332 human genes interacting with SARS-CoV-2 for viral sequence divergence or human population variants along viral–human (V:H, $\log_2$OR = −0.91, $P = 2.41 \times 10^{-10}$ by two-sided z-test) human–viral (H:V, $\log_2$OR = −0.38, $P = 7.92 \times 10^{-13}$ by two-sided *z*-test) or human–human (H:H, $\log_2$OR = −0.14, $P = 9.98 \times 10^{-4}$ by two-sided *z*-test) interfaces. Data are presented as $\log_2$OR ± s.e.m. **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 $\log_2$OR (by two-sided *z*-test) are labeled and shown as bars, the remainder are plotted as a line. Data are presented as $\log_2$OR ± s.e.m. 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 ($\log_2$OR = 0.57, $P = 1.70 \times 10^{-4}$ by two-sided *z*-test), ClinVar ($\log_2$OR = 0.64, $P = 1.05 \times 10^{-4}$ by two-sided *z*-test) and GWAS ($\log_2$OR = 0.89, $P = 4.54 \times 10^{-5}$ 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 ± s.e.m. **e**, Sample of individual disease terms enriched in human genes targeted by SARS-CoV-2. Full results are reported in Supplementary Table 6. Data are presented as $\log_2$OR ± s.e.m. **f**, Comparison of the enrichment of HGDM- or ClinVar-annotated mutations on human–viral interfaces or human–human interfaces for 332 genes interacting with SARS-CoV-2. Disease mutations were enriched on human–human interfaces (HGMD, $\log_2$OR = 0.84, $P < 1 \times 10^{-20}$ by two-sided *z*-test; ClinVar, $\log_2$OR = 0.25, $P = 2.9 \times 10^{-3}$ by two-sided *z*-test), whereas human–viral interfaces showed no or marginal enrichment (HGMD, $\log_2$OR = 0.31, $P = 0.048$ by two-sided *z*-test; ClinVar, $\log_2$OR = 0.15, $P = 0.39$ by two-sided *z*-test). GWAS category was excluded from this analysis because most lead GWAS single-nucleotide polymorphisms occur in noncoding regions. Data are presented as $\log_2$OR ± s.e.m.

disease, respiratory tract disease and metabolic disease[12,71] (Fig. 3e and Supplementary 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[72,73]. 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[74] and PolyPhen[75] (Extended Data Fig. 3). However, while we showed that annotated disease mutations were significantly enriched along known human–human interfaces, enrichment was drastically reduced (HGMD) or insignificant (ClinVar) on human–viral interfaces (Fig. 3f). 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.

Binding affinity changes from SARS-CoV-1 to SARS-CoV-2. Using a PyRosetta pipeline[47,59,60] 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 toward a decreased binding energy (that is more stable interaction) (Fig. 4a and Supplementary Table 8). The 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 six of these interactions (20%) using the SARS-CoV-2 bait. Extensive previous 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%)[76–79]. 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 six 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. 4c). 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. 4c).

We additionally performed co-immunoprecipitation (co-IP) assays for the interaction between human DNA Primase Subunit 2 (PRIM2) and SARS-CoV-2 nsp1 (Fig. 4b and Source Data Fig. 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 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[80] included five interactions that we predicted to be more stable in SARS-CoV-2. Consistent with our predictions three 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 modeling 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[81] and that underlying genetic variation can explain up to 15% of variation in patient response and viral load in other viruses, including HIV[82]. To explore this hypothesis we employed a previously benchmarked scanning mutagenesis protocol provided through PyRosetta[47,59] 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. 4d).

To validate our predictions for the impact of population variants, we generated a Ras GTPase-activating protein-binding protein 2 (G3BP2) variant, G3BP2_P121T (rs1185000405) using site-directed mutagenesis as described previously[83]. We annotated this variant as strongly disruptive (predicted $\Delta\Delta G = 10.3$ REU) and had confirmed earlier that the interaction between N and wild-type G3BP2 could be recapitulated using Y2H. Comparing the Y2H results between wild-type and mutant G3BP2 confirmed complete disruption of the G3BP2–N protein interaction by G3BP2_P121T (Fig. 4e). Analysis of the docked models suggests that 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[84], 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[85,86]. 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 modeling 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 Supplementary Table 9 and may inform future studies investigating genetic contribution to COVID-19.

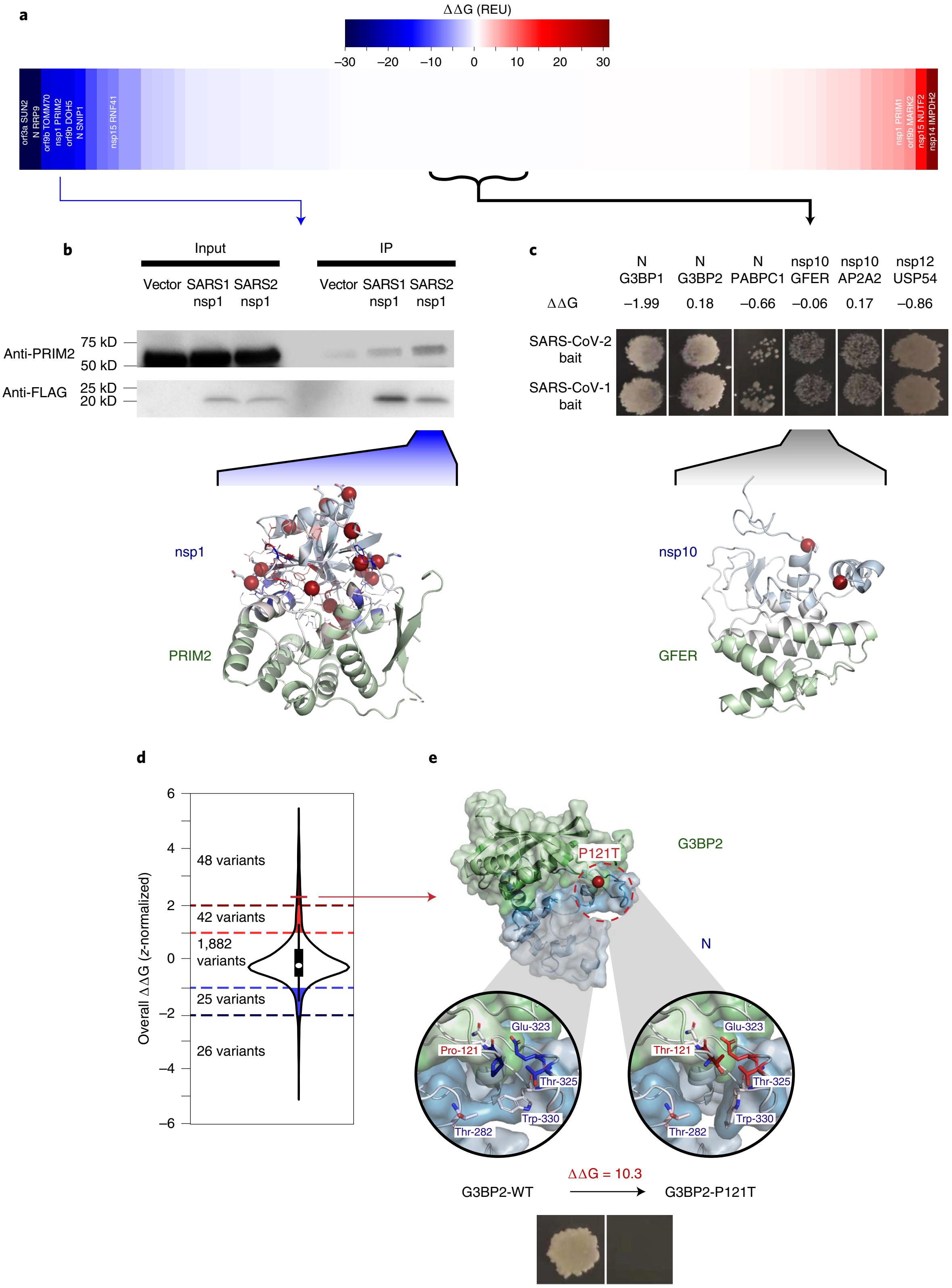
Comparing binding sites of drugs and SARS-CoV-2 proteins. Drugs that directly interfere with viral–host interactions (for instance by competing for the same binding site) could provide promising clinical leads to target viral infection or replication. On this basis we consider potential for our 3D interactome modeling 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 (ref. [20]) on the basis of the potential for competitive binding. We performed protein–ligand docking using smina[87] to identify drug binding sites for 30 out of 76 candidate drug–target pairs that have available human receptor structures (Supplementary Table 10). Smina, a fork of the widely used AutoDock Vina, competes competitively in pose prediction challenges[87] and 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. 5a)[88].

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. 5b; $\log_2\mathrm{OR} = 1.38$, $P = 2.1 \times 10^{-7}$).

Fig. 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, s.d. = 5.78 REU). Interaction labels shown wherever predicted $\Delta\Delta G$ exceeds mean ± (1 × s.d.). **b**, Representative cropped western blots (among three replicates) from co-IP comparing the interaction between human PRIM2 with SARS-CoV-1 or SARS-CoV-2 nsp1. More efficient PRIM2 pulldown with SARS-CoV-2 bait validates the PRIM2-nsp1 $\Delta\Delta G$ prediction ($\Delta\Delta G = -17.3$ REU, *z* score = −2.9). Docked structure for PRIM2 with SARS-CoV-2 nps1 (green and blue cartoon, respectively) (bottom). SARS-CoV-1 to SARS-CoV-2 sequence divergences are represented as spheres. Interface residues are colored relative to overall $\Delta\Delta G$ contribution ranging from blue (more stabilizing in SARS-CoV-2) to white (little impact on $\Delta\Delta G$), to red (more stabilizing in SARS-CoV-1). Residue side chains are shown as sticks in regions with high local $\Delta\Delta G$. **c**, Representative Y2H results (among three replicates) confirming that six interactions with no predicted $\Delta\Delta G$ 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 with SARS-CoV-2 nsp10 ($\Delta\Delta G = -0.06$) is shown to highlight that sequence divergences in these six 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 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 ($1 \leq z$ score <2, 42 total variants), stabilizing ($-2 < z$ score ≤−1, 25 total variants) or strongly stabilizing (*z* score ≤−2, 26 total variants). Interior box plot 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 wild-type interface (left) with a predicted strongly disruptive ($\Delta\Delta G = 10.3$ REU, *z* score = 2.3) population variant, G3BP2_P121T (right). Yeast two-hybrid results confirmed that the G3BP2_P121T variant completely disrupts the G3BP2–N interaction (bottom).

Individually, we further prioritized eight drugs that exhibited significant overlap between the drug- and viral-protein-binding sites (Fig. 5c), several of which have been explored by recent independent studies. A retroactive association study identified previous treatment with metformin as an independent factor associated with reduced mortality in diabetic patients[89], although a precise mechanism was not explored at the time. Ongoing phase 2 and phase 4 clinical trials are being conducted or are planned for silmitasertib and valproic acid, respectively (ClinicalTrials.gov identifiers NCT04668209 and NCT04513314).

As an example, we highlight orf9b-MARK3 interaction whose interface we predicted could be blocked by ZINC95559591



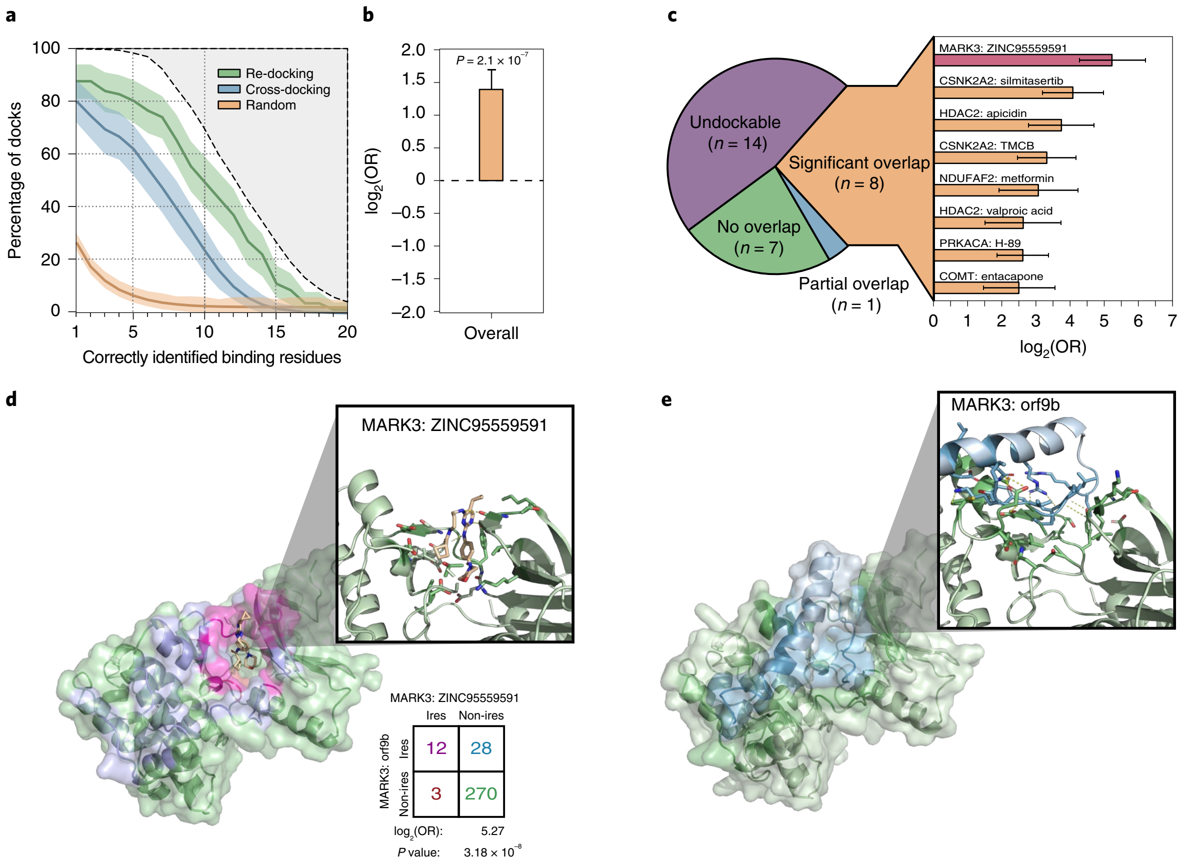
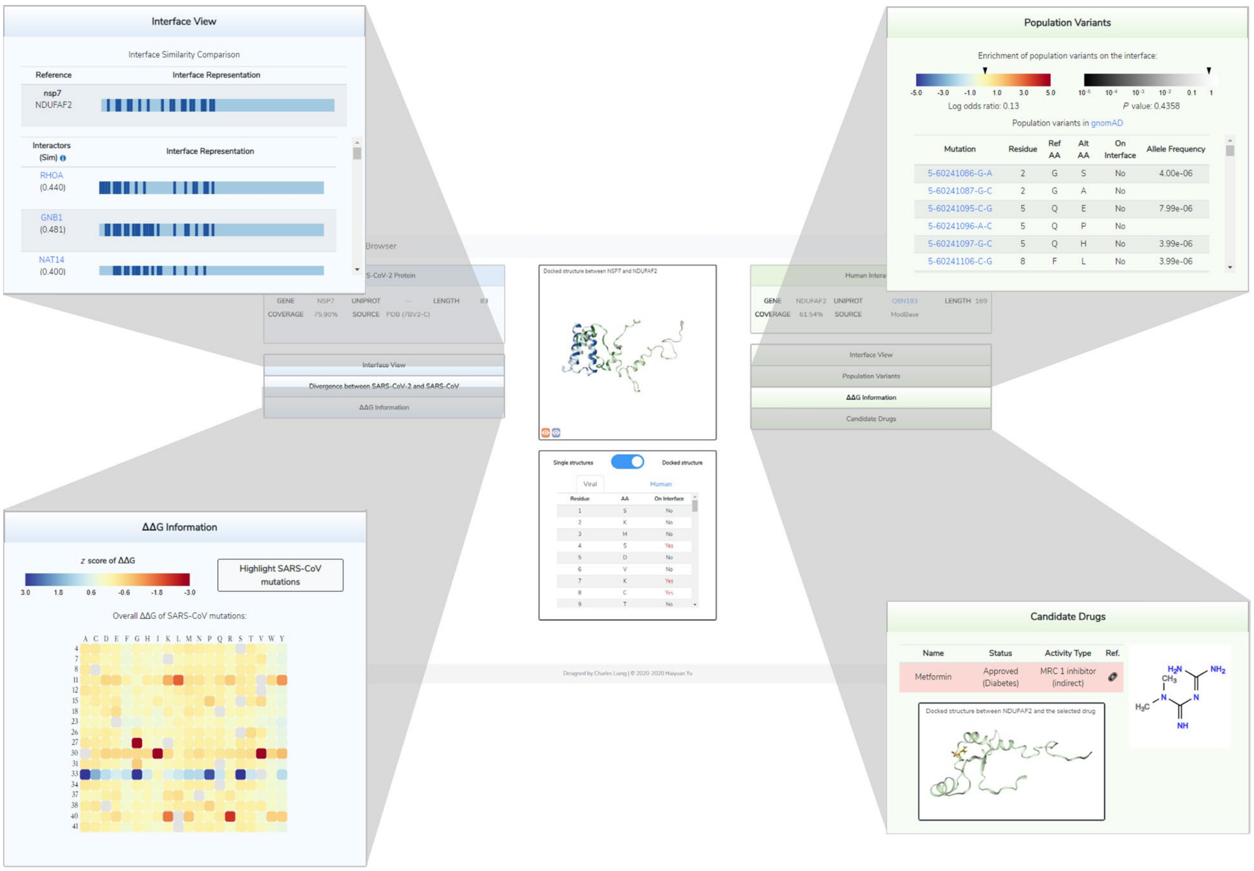


Fig. 5 | Drug docking and prioritization of SARS-CoV-2–human interaction inhibitors. a, Validation of smina's ability to identify the correct binding site from the full protein surface based on 4,399 drug–ligand pairs across 95 protein targets. Docking was carried out either by re-docking each ligand 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 ± s.d. as estimated by 1,000-fold bootstrapping sampling 95 drug–target pairs with replacement each iteration. The gray shaded area (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 ($\log_2$OR = 1.38, $P = 2.1 \times 10^{-7}$ by two-sided *z*-test). Data are presented as $\log_2$OR ± s.e.m. **c**, Individual breakdown of the overlap between the each of the protein–protein and protein–drug binding sites as either undockable (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 $\log_2$OR for each of the significant drug target pairs are shown. Data are presented as $\log_2$OR ± s.e.m. The MARK3–ZINC95559591 pair (shown in **d**) is highlighted in 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 are shown as dashed lines. **e**, Corresponding docked structure for SARS-CoV-2 orf9b bound to human MARK3.

(MRT-68601 hydrochloride) (Fig. 5d,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 that both ZINC95559591 and orf9b bind and make several polar contacts with MARK3 (for example one with Tyr-134) near its active ATP site. Consistent with its known role as an inhibitor of MARK3 (ref. [90]) our model shows that 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 that 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 versus disrupts native human protein function or whether interaction is part of an immune response against the virus) is needed to inform potential clinical utility of drug repurposing. As 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 to 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 modeling 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



Interface View

Interface Similarity Comparison

Reference	Interface Representation
nsp7 NDUFAF2	

Interactors (Sim)	Interface Representation
RHOA (0.440)	
GNB1 (0.481)	
NAT14 (0.400)	

Browser

S-CoV-2 Protein

GENE NSP7 UNIPROT -- LENGTH 83
COVERAGE 75.90% SOURCE PDB (7BV2-C)

Interface View

Divergence between SARS-CoV-2 and SARS-CoV

ΔΔG Information

Docked structure between NSP7 and NDUFAF2

Single structures / Docked structure

Viral / Human

Residue	AA	On Interface
1	S	No
2	K	No
3	M	No
4	S	Yes
5	D	No
6	V	No
7	K	Yes
8	C	Yes
9	T	No

Designed by Charles Liang | © 2020-2020 Haiyuan Yu

Human Intera

GENE NDUFAF2 UNIPROT Q8N183 LENGTH 169
COVERAGE 61.54% SOURCE ModBase

Interface View

Population Variants

ΔΔG Information

Candidate Drugs

Population Variants

Enrichment of population variants on the interface:

Log odds ratio: 0.13

P value: 0.4358

Population variants in gnomAD

Mutation	Residue	Ref AA	Alt AA	On Interface	Allele Frequency
5-60241086-G-A	2	G	S	No	4.00e-06
5-60241087-G-C	2	G	A	No	
5-60241095-C-G	5	Q	E	No	7.99e-06
5-60241096-A-C	5	Q	P	No	
5-60241097-G-C	5	Q	H	No	3.99e-06
5-60241106-C-G	8	F	L	No	3.99e-06

ΔΔG Information

z score of ΔΔG

Highlight SARS-CoV mutations

Overall ΔΔG of SARS-CoV mutations:

Candidate Drugs

Name	Status	Activity Type	Ref.
Metformin	Approved (Diabetes)	MRC 1 inhibitor (indirect)	🔗

Docked structure between NDUFAF2 and the selected drug

Fig. 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 a single structure 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. The interface view shows a linear representation of the protein sequence with interface residues annotated in dark blue or dark green (top left). Interfaces for other interactors of the protein are shown underneath for easy comparison. The mutations panel summarizes either human population variants or viral sequence divergences on the protein (top right). Mutations on the interface are labeled. The ΔΔG information panel summarizes the results from in silico mutagenesis scanning along the interface (bottom left). Results for each mutation are *z* score-normalized relative to the rest of the interface and colored blue (negative ΔΔG, stabilizing) to yellow (minimal impact) to red (positive ΔΔG, destabilizing). The heat map can be filtered to show only values corresponding to known mutations on the interface. The candidate drugs panel shows docking information for any known drug targets of the human protein (bottom right).

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 bottom 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 bottom 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[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[76]. Further, while structural coverage from SARS-CoV-2 proteins was robust, per-residue coverage of the human proteome is less complete (Extended Data Fig. 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[91–93] 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 $\Delta\Delta G$ 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 substantially impact binding affinity through refolding or other allosteric effects exist but cannot be captured by our method.

Notably, 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[94] with the aim of designing new protein–protein interaction inhibitors.

Overall, we believe that 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 web server will continue to grow as we incorporate the results of ongoing and future interactome screens between SARS-CoV-2 and human data. Finally, we note our 3D structural interactome framework can be rapidly deployed to analyze future viruses.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-021-01318-w.

Received: 18 September 2020; Accepted: 5 October 2021;
Published online: 29 November 2021

References

1. *COVID-19 Dashboard* (Johns Hopkins University, 2020); https://coronavirus.jhu.edu/map.html
2. Fehr, A. R. & Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol. Biol.* **1282**, 1–23 (2015).
3. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
4. McIntosh, K. & Perlman, S. in *Mandell, Douglas and Bennett's Principles and Practice of Infectious Diseases* 8th edn (eds Bennett, J. E. et al.) 1928–1936 (2015).
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.* **30**, 2196–2203 (2020).
6. Gupta, A. et al. Extrapulmonary manifestations of COVID-19. *Nat. Med.* **26**, 1017–1032 (2020).
7. Wang, D. et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus-infected pneumonia in Wuhan, China. *JAMA* **323**, 1061–1069 (2020).
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.* **8**, 475–481 (2020).
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* **395**, 1054–1062 (2020).
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* **108**, 154262 (2020).
11. Ferdinand, K. C. & Nasser, S. A. African-American COVID-19 mortality: a sentinel event. *J. Am. Coll. Cardiol.* **75**, 2746–2748 (2020).
12. Killerby, M. E. et al. Characteristics associated with hospitalization among patients with COVID-19 - Metropolitan Atlanta, Georgia, March–April 2020. *MMWR Morb. Mortal. Wkly Rep.* **69**, 790–794 (2020).
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* **42**, 451–460 (2020).
14. Moore, J. T. 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.* **69**, 1122–1126 (2020).
15. Mahajan, U. V. & Larkins-Pettigrew, M. Racial demographics and COVID-19 confirmed cases and deaths: a correlational analysis of 2886 US counties. *J. Public Health* **42**, 445–447 (2020).
16. Pfefferle, S. et al. The SARS-coronavirus-host interactome: identification of cyclophilins as target for pan-coronavirus inhibitors. *PLoS Pathog.* **7**, e1002331 (2011).
17. Jager, S. et al. Global landscape of HIV–human protein complexes. *Nature* **481**, 365–370 (2011).
18. Batra, J. et al. Protein interaction mapping identifies RBBP6 as a negative regulator of ebola virus replication. *Cell* **175**, 1917–1930 (2018).
19. Shah, P. S. et al. Comparative flavivirus-host protein interaction mapping reveals mechanisms of dengue and zika virus pathogenesis. *Cell* **175**, 1931–1945 (2018).
20. Gordon, D. E. et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
21. Niemann, H. H. et al. Structure of the human receptor tyrosine kinase met in complex with the *Listeria* invasion protein InlB. *Cell* **130**, 235–246 (2007).
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* **181**, 271–280 (2020).
23. Xu, G. G., Guo, J. & Wu, Y. Chemokine receptor CCR5 antagonist maraviroc: medicinal chemistry and clinical applications. *Curr. Top. Med. Chem.* **14**, 1504–1514 (2014).
24. Hayouka, Z. et al. Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. *Proc. Natl Acad. Sci. USA* **104**, 8316–8321 (2007).
25. Peat, T. S. et al. Small molecule inhibitors of the LEDGF site of human immunodeficiency virus integrase identified by fragment screening and structure based design. *PLoS ONE* **7**, e40147 (2012).

26. Maginnis, M. S. Virus–receptor interactions: the key to cellular invasion. *J. Mol. Biol.* **430**, 2590–2611 (2018).
27. Daczkowski, C. M. et al. Structural insights into the interaction of coronavirus papain-like proteases and interferon-stimulated gene product 15 from different species. *J. Mol. Biol.* **429**, 1661–1683 (2017).
28. Yao, J. et al. Mechanism of inhibition of retromer transport by the bacterial effector RidL. *Proc. Natl Acad. Sci. USA* **115**, E1446–E1454 (2018).
29. Zhang, L. et al. Solution structure of the complex between poxvirus-encoded CC chemokine inhibitor vCCI and human MIP-1β. *Proc. Natl Acad. Sci. USA* **103**, 13985–13990 (2006).
30. Jonker, H. R. et al. Structural properties of the promiscuous VP16 activation domain. *Biochemistry* **44**, 827–839 (2005).
31. Card, G. L. et al. Crystal structure of a γ-herpesvirus cyclin-cdk complex. *EMBO J.* **19**, 2877–2888 (2000).
32. Smith, M., Honce, R. & Schultz-Cherry, S. Metabolic syndrome and viral pathogenesis: lessons from influenza and coronaviruses. *J. Virol.* **94**, e00665-20 (2020).
33. Vidal, M. A unifying view of 21st century systems biology. *FEBS Lett.* **583**, 3891–3894 (2009).
34. Robinson, C. V., Sali, A. & Baumeister, W. The molecular sociology of the cell. *Nature* **450**, 973–82. (2007).
35. Barabasi, A. L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
36. Scott, D. E. et al. Small molecules, big targets: drug discovery faces the protein–protein interaction challenge. *Nat. Rev. Drug Discov.* **15**, 533–50. (2016).
37. Arkin, M. R., Tang, Y. & Wells, J. A. Small-molecule inhibitors of protein–protein interactions: progressing toward the reality. *Chem. Biol.* **21**, 1102–1114 (2014).
38. Rooklin, D. et al. AlphaSpace: fragment-centric topographical mapping to target protein–protein interaction interfaces. *J. Chem. Inf. Model.* **55**, 1585–1599 (2015).
39. Lampson, B. L. & Davids, M. S. The development and current use of BCL-2 inhibitors for the treatment of chronic lymphocytic leukemia. *Curr. Hematol. Malig. Rep.* **12**, 11–19 (2017).
40. Schormann, N. et al. Identification of protein–protein interaction inhibitors targeting vaccinia virus processivity factor for development of antiviral agents. *Antimicrob. Agents Chemother.* **55**, 5054–5062 (2011).
41. White, P. W. et al. Inhibition of human papillomavirus DNA replication by small molecule antagonists of the E1–E2 protein interaction. *J. Biol. Chem.* **278**, 26765–26772 (2003).
42. 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.* **15**, 2690–2700 (2007).
43. Brito, A. F. & Pinney, J. W. Protein–protein interactions in virus–host systems. *Front. Microbiol.* **8**, 1557 (2017).
44. Meyer, M. J. et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nat. Methods* **15**, 107–114 (2018).
45. Dominguez, C., Boelens, R. & Bonvin, A. M. HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
46. van Zundert, G. C. P. et al. The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
47. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).
48. Kirchdoerfer, R. N. et al. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci. Rep.* **8**, 15701 (2018).
49. Wang, Q. et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **181**, 894–904 (2020).
50. Wrobel, A. G. et al. SARS-CoV-2 and bat RaTG13 spike glycoprotein structures inform on virus evolution and furin-cleavage effects. *Nat. Struct. Mol. Biol.* **27**, 763–767 (2020).
51. Walls, A. C. et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. *Cell* **181**, 281–292 (2020).
52. Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).
53. Shang, J. et al. Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
54. Wrapp, D. et al. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
55. Jordan, R. E. & Adab, P. Who is most likely to be infected with SARS-CoV-2? *Lancet Infect. Dis.* **20**, 995–996 (2020).
56. Cao, Y. et al. Comparative genetic analysis of the novel coronavirus (2019-nCoV/SARS-CoV-2) receptor ACE2 in different populations. *Cell Discov.* **6**, 11 (2020).
57. Darbeheshti, F. & Rezaei, N. Genetic predisposition models to COVID-19 infection. *Med. Hypotheses* **142**, 109818 (2020).
58. Zhao, Y. et al. Single-cell RNA expression profiling of ACE2, the receptor of SARS-CoV-2. *Am. J. Respir. Crit. Care Med.* **202**, 756–759 (2020).
59. Kortemme, T. & Baker, D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proc. Natl Acad. Sci. USA* **99**, 14116–14121 (2002).
60. Shulman-Peleg, A. et al. Spatial chemical conservation of hot spot interactions in protein–protein complexes. *BMC Biol.* **5**, 43 (2007).
61. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
62. Suryamohan, K. et al. Human ACE2 receptor polymorphisms and altered susceptibility to SARS-CoV-2. *Commun. Biol.* **4**, 475 (2021).
63. Chan, K. K. et al. Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science* **369**, 1261–1265 (2020).
64. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
65. Pieper, U. et al. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* **42**, D336–D346 (2014).
66. Guharoy, M. & Chakrabarti, P. Conserved residue clusters at protein–protein interfaces and their use in binding site identification. *BMC Bioinf.* **11**, 286 (2010).
67. Gupta, R. et al. SARS-CoV-2 (COVID-19) structural and evolutionary dynamicome: insights into functional evolution and human genomics. *J. Biol. Chem.* **295**, 11742–11753 (2020).
68. Stenson, P. D. et al. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* **21**, 577–581 (2003).
69. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
70. Buniello, A. et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
71. 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.* **94**, 91–95 (2020).
72. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
73. Wang, X. et al. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30**, 159–64. (2012).
74. Sim, N. L. et al. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
75. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
76. Yu, H. et al. High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
77. Braun, P. et al. An experimentally derived confidence score for binary protein–protein interactions. *Nat. Methods* **6**, 91–97 (2009).
78. Vo, T. V. et al. A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. *Cell* **164**, 310–323 (2016).
79. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
80. Gordon, D. E. et al. Comparative host-coronavirus protein interaction networks reveal pan-viral disease mechanisms. *Science* **370**, eabe9403 (2020).
81. Fragoza, R. et al. Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. *Nat. Commun.* **10**, 4141 (2019).
82. Fellay, J. et al. A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944–947 (2007).
83. Wei, X. et al. A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. *PLoS Genet.* **10**, e1004819 (2014).
84. Hong, H. Q. et al. G3BP2 is involved in isoproterenol-induced cardiac hypertrophy through activating the NF-κB signaling pathway. *Acta Pharmacol. Sin.* **39**, 184–194 (2018).
85. Lu, S. et al. The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat. Commun.* **12**, 502 (2021).
86. 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.* **10**, 168 (2020).
87. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).

88. Wierbowski, S. D. et al. Cross-docking benchmark for automated pose and ranking prediction of ligand binding. *Protein Sci.* **29**, 298–305 (2020).
89. Crouse, A. B. et al. Metformin use is associated with reduced mortality in a diverse population with COVID-19 and diabetes. *Front Endocrinol.* **11**, 600439 (2020).
90. McIver, E. G. et al. Synthesis and structure–activity relationships of a novel series of pyrimidines as potent inhibitors of TBK1/IKKepsilon kinases. *Bioorg. Med. Chem. Lett.* **22**, 7169–7173 (2012).
91. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
92. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
93. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
94. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Generation and validation of SARS-CoV-2 homology models. Homology-based modeling of all 29 SARS-CoV-2 proteins was performed in Modeller[95] using a multiple template modeling procedure consistent with previous high-profile homology modeling resources[96]. In brief, candidate template structures for each query protein were selected by running BLAST[97] against all sequences in the PDB[64] retaining only templates with at least 30% identify. Remaining templates were ranked using a weighted combination of percent identity and coverage described previously[96]. 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 75% the identity of the initial seed template (that is, if the template seed showed 80% identity, additional templates with percent identity as low as 60% could be included). Query-template pairwise alignments were generated in Modeller using default settings and were manually trimmed to remove large gaps (five or more gaps in a ten-residue window). Finally, modeling 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 RMSD following alignment and refinement in PyMol[98]. Visual representations of these alignments between modeled and solved structures are provided in Extended Data Fig. 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 previously[20] were made in two phases. In phase one, we leveraged our previously validated ECLAIR framework[44] 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[99], conservation features are derived from the Jensen–Shannon divergence[100,101] from known homologs for each protein, coevolution features between interacting proteins are derived from direct coupling analysis[102] and statistical coupling analysis[103] among paired homologs, structural features are obtained by calculating the solvent-accessible surface area of available PDB[64] or ModBase[65] models using NACCESS[104] and docking features are the average interchain distance and surface occlusion per residue from a consensus of independent Zdock[105] trials.

Slight alterations were made to accommodate SARS-CoV-2–human predictions. First, construction of multiple sequence alignment (MSA) for statistical coupling analysis and direct coupling analysis calculations require at least 50 species containing homologs of both interacting proteins. Therefore, coevolution features could not be calculated for interspecies 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 Supplementary Table 1.

Interface prediction using guided HADDOCK docking. Interface predictions for all 332 interactions reported previously[20] were made in two phases. In phase two, we leveraged high-confidence interface predictions from ECLAIR to perform guided docking in HADDOCK[45,46]. An introduction to protein–protein docking in HADDOCK is provided at 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 (for example 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 (AIRs) that evaluate the distances of each active residue to the active or passive residues on the other protein. The approach ensures that 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% solvent-accessible surface area (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 provides consistency with our definition of interface residues, whereas the 40% SASA cutoff provides consistency with the typical recommendation in HADDOCK. All SASA calculations were carried out using NACCESS[104] and neighboring residues were selected using PyMol[98]. Following HADDOCK recommendations to reduce computational burden from using many restraints, we defined our AIRs using only the α-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, 1,000 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 that HADDOCK performs well using homology models, but that performance drops off for models produced from low sequence identity templates[106]. 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 Supplementary Table 2.

Definition of interface residues. We annotated interface residues from atomic-resolution docked models, using an established definition for interface residues[44]. The SASA for both bound and unbound docked structures was calculated using NACCESS[104]. We defined an interface residue as 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 interspecies interactions (Fig. 2a). 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[74] 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 intraspecies interactions and 1,482 represented interspecies interactions between humans and some other species. Finally, to provide the most relevant set of interactions that would be biologically similar to SARS-CoV-2–human interactions, we considered only 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 Supplementary 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 performed by AUROC in Python using scikit-learn and was compared against ECLAIR's original test set containing 200 intraspecies interactions[44]. Precision and recall metrics were calculated based on ECLAIR's binary definition for high-confidence versus 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 (AIRs) 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 that the two protein chains slide together during docking. Overall performance of protocols was evaluated based on precision and recall of the true interface (Fig. 2c). Secondary evaluation was conducted based on RMSD in PyMol before refinement between the docked and co-crystal structures (Fig. 2d). 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)[107,108]. Sequences for 29 SARS-CoV-2 proteins were reported previously[20] and based on GenBank accession code MN985325 (refs. [109,110]). Notably, UniProt accession codes 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[111,112] (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 Supplementary Table 5.

We obtained human population variants for all 332 human proteins interacting with SARS-CoV-2 proteins from gnomAD[61]. 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 at https://github.com/broadinstitute/gnomad-browser/tree/master/projects/gnomad-api. We used the Ensembl Variant Effect Predictor[113] to map gnomAD DNA-level single-nucleotide polymorphisms (SNPs) to equivalent protein-level UniProt annotations. After Variant Effect Predictor 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 Supplementary Table 4.

Log odds enrichment calculations. To determine enrichment or depletion, ORs were calculated as described previously.[114]

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

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 population variants (case, variant versus nonvariant) along predicted interaction interfaces (exposure, interface versus non-interface), we would have:

$$a = \text{number of variant interface residues}$$

$$b = \text{number of nonvariant interface residues}$$

$$c = \text{number of variant noninterface residues}$$

$$d = \text{number of nonvariant noninterface residues}$$

Statistical tests for enrichment or depletion were performed by calculating the z-statistic and corresponding two-sided P value for the OR (unadjusted for multiple hypothesis testing).

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

All reported ORs were $\log_2$ transformed to maintain interpretable symmetry between enriched and depleted values. To avoid arbitrary OR inflation or depletion from missing data, in all cases where the interface residues were predicted by molecular docking, the OR 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: HGMD[68], ClinVar[69] and the NHGRI-EBI GWAS catalog[70]. 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. A meaningful subset of significantly enriched terms were reported using the most general term with no more significant ancestor term (Supplementary Table 7, sheet 1). Raw enrichment values for all terms are also provided (Supplementary Table 7, sheet 2).

For curation of disease and trait associations from the NHGRI-EBI GWAS catalog (http://www.ebi.ac.uk/gwas/)[70], lead SNPs (P value $<5\times10^{-8}$) for all diseases/traits were retrieved on 16 June 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[115], 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 retain only missense variants.

In silico scanning mutagenesis and ΔΔG estimation. To explore the 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 used 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[59]. 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 considered all interface residue positions and began by estimating the wild-type 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 wild-type (WT) structure.

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

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[59] 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 for stochasticity of the PackRotamersMover optimization between trials, all ΔΔG values are reported from an average of ten independent trials. To test whether a mutation had a significantly nonzero impact on binding energy, a two-sided z-test between the ten 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 × s.d. away from the mean. Mutations that passed both criteria were identified as 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 modeled 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 values were calculated considering the SARS-CoV-1 as the wild-type such that a negative ΔΔG indicates that 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. 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[116] and the command:

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

Protein–ligand docking was executed using smina[87] 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 40 and the num_ modes was set to retain the top 1,000 ranked models. To reduce real wall time, each docking process was run using five 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 ten times (essentially the same as one trial with exhaustiveness set to 400) with a unique seed 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 applications[87]. To retain candidate poses covering different low-energy binding sites, a final set of up to ten 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 shows competitive performance across numerous Community Structure-Activity Resources[87,88]. However, traditional 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 manuscript) we re-benchmarked smina's performance using an established drug docking benchmark set containing 4,399 protein–ligand complexes representing 95 protein targets[88]. 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 re-docking (ligand docked back into the exact receptor structure it came from) and cross-docking (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, cross-docking 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[117]. 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-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 screens. Y2H experiments were carried out as previously described[76,81,118] to (1) confirm that SARS-CoV-2–human interactions previously detected by IP–mass spectrometry could be recapitulated in Y2H, (2) compare the occurrence of interactions using SARS-CoV-1 versus 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 an empty pDEST-AD vector to screen for autoactivators. Mated transformants were incubated overnight at 30 °C, before being plated onto selective Synthetic Complete agar medium 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 3 d 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 supplemented with 10% FBS. Cells were seeded onto six-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 16,000*g* at 4 °C. For co-IP, 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-IP samples were then treated in 6× SDS protein loading buffer (10% SDS, 1 M Tris-Cl, pH 6.8, 50% glycerol, 10% β-mercaptoethanol and 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. Mutant clones containing human population variants were generated using site-directed mutagenesis as described previously[83]. In brief, wild-type *G3BP2* was picked from the hORFeome 8.1 library[117] 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 (and no other mutations) had been introduced.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Protein–protein interaction sets and drug–target candidate pairs were obtained from data reported previously[20]. 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 (for example, HGMD is restricted access) the formatted data are provided in the Supplementary 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 modeling were obtained from the 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 have been described; 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 S.D.W. (sdw95@cornell.edu) and H.Y. (haiyuan.yu@cornell.edu). Source data are provided with this paper.

Code availability

Scripts used for guided docking and relevant analyses are available on GitHub (https://github.com/hyulab/3D_SARS2) and Zenodo (https://zenodo.org/record/4987957#.YXAaZRzTVQI). Source data are provided with this paper.

References

95. Eswar, N. et al. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinforma.* **5**, 5–6 (2006).

96. Mosca, R., Ceol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat. Methods* **10**, 47–53 (2013).
97. Altschul, S. F. et al. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
98. *The PyMOL Molecular Graphics System*, Version 2.0 (Schrödinger, LLC).
99. Gasteiger, E. et al. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press, 2005).
100. Lin, J. H. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**, 145–151 (1991).
101. Capra, J. A. & Singh, M. Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882 (2007).
102. Morcos, F. et al. Direct coupling analysis for protein contact prediction. *Methods Mol. Biol.* **1137**, 55–70 (2014).
103. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286**, 295–299 (1999).
104. Lee, B. & Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400 (1971).
105. Pierce, B. G., Hourai, Y. & Weng, Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PLoS ONE* **6**, e24657 (2011).
106. Rodrigues, J. P. et al. Defining the limits of homology modeling in information-driven protein docking. *Proteins* **81**, 2119–2128 (2013).
107. UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
108. He, R. et al. Analysis of multimerization of the SARS coronavirus nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **316**, 476–483 (2004).
109. Wu, F. et al. A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
110. Chan, J. F. et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* **9**, 221–236 (2020).
111. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
112. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA* **89**, 10915–10919 (1992).
113. McLaren, W. et al. The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
114. Szumilas, M. Explaining odds ratios. *J. Can. Acad. Child Adolesc. Psychiatry* **19**, 227–229 (2010).
115. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
116. O'Boyle, N. M. et al. Open Babel: an open chemical toolbox. *J. Cheminform* **3**, 33 (2011).
117. Yang, X. et al. A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).
118. Das, J. et al. Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. *Sci. Signal* **6**, ra38 (2013).

Acknowledgements

This work was supported by grants from The National Institute of General Medical Sciences (R01 GM124559 and R01 GM125639), The National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK115398), The Eunice Kennedy Shriver National Institute of Child Health and Human Development (R01 HD082568), The National Human Genome Research Institute (UM1 HG009393), NSF (DBI-1661380), The Simons Foundation Autism Research Initiative (575547) and Cornell Rapid Research Response to SARS-CoV-2 Seed Grant to H.Y., National Institutes of Health (R01AI35270) to G.R.W. and William Strong COVID-19 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 P. H. Wang from Shandong University.

Author contributions

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. Experiments were carried out by Y.L. The development and maintenance of the 3D-SARS2 website was performed 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

The authors declare no competing interests.

Additional information

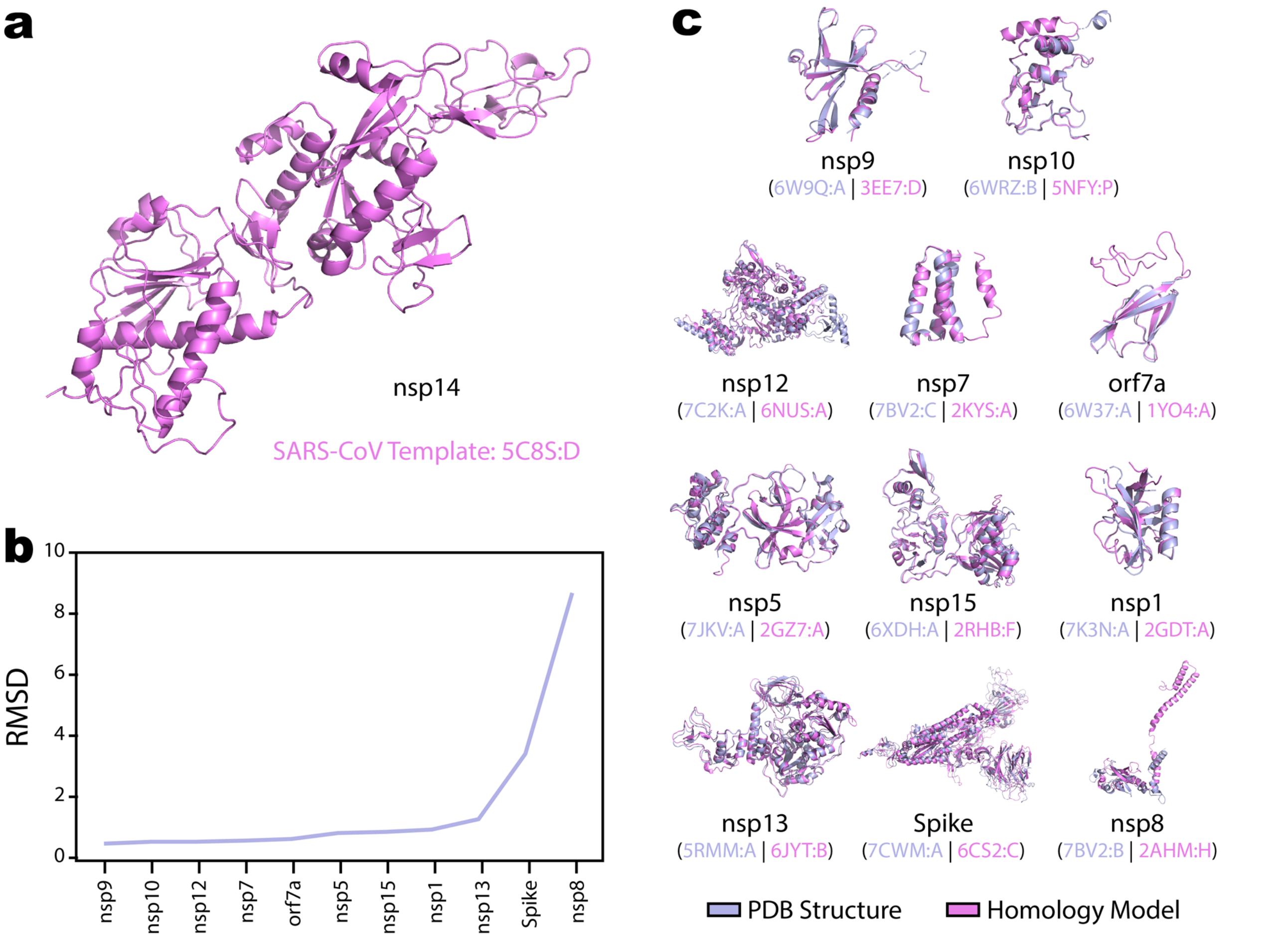
Extended data is available for this paper at https://doi.org/10.1038/s41592-021-01318-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-021-01318-w.

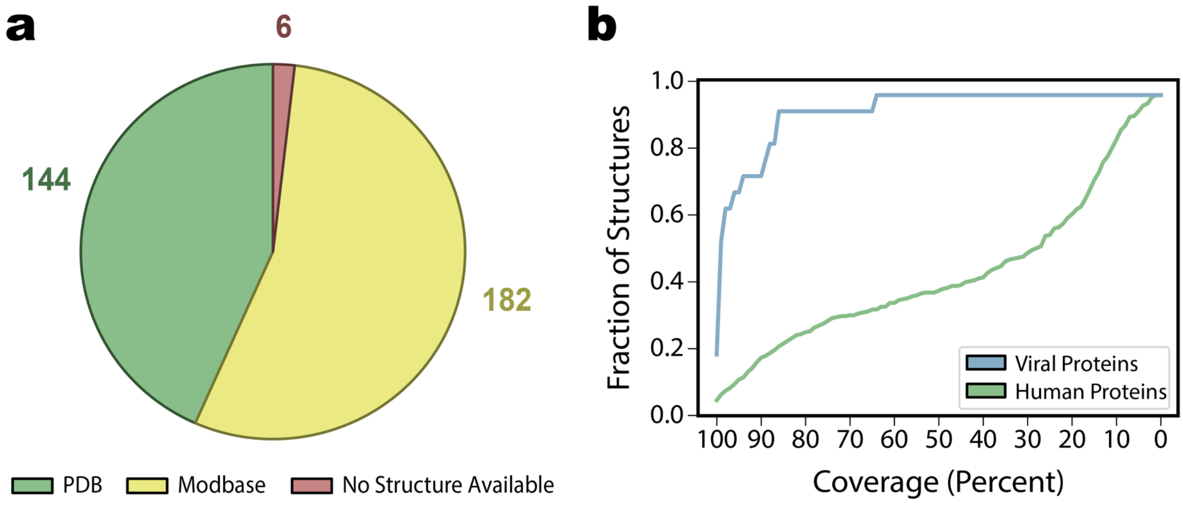
Correspondence and requests for materials should be addressed to Haiyuan Yu.

Peer review information Arunima Singh was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

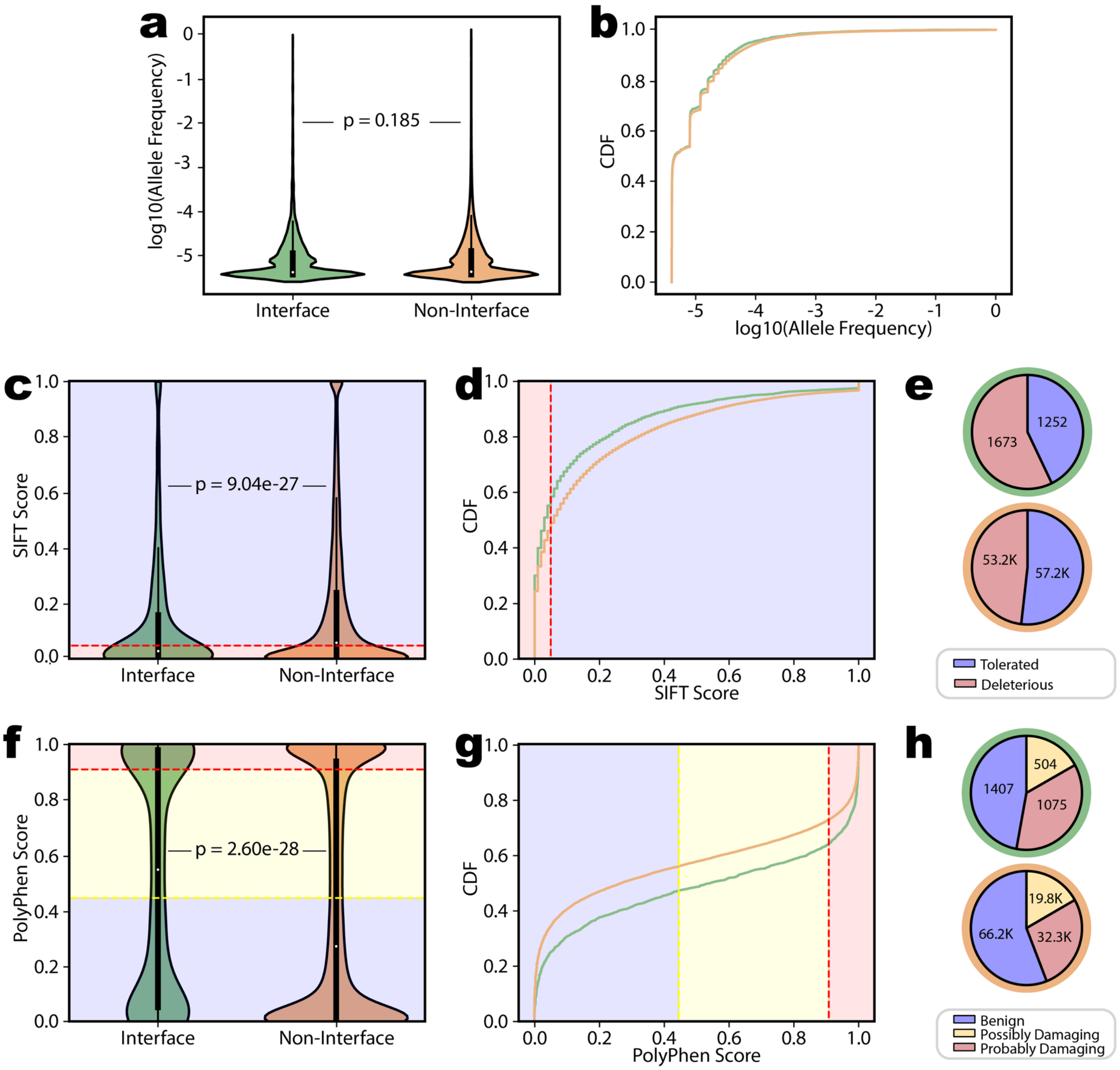
Reprints and permissions information is available at www.nature.com/reprints.



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.

nature research

Corresponding author(s): Haiyuan Yu

Last updated by author(s): Jun 16, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	none
Data analysis	Interface predictions were done using our published ECLAIR framework. Protein-protein docking was performed in HADDOCK 2.4 (September 2020) using ECLAIR-predicted interface residues as restraints (for a tutorial describing a comparable HADDOCK protocol see https://www.bonvinlab.org/education/HADDOCK-protein-protein-basic/). In-silico mutagenesis / energy function modeling were done using PyRosetta 4 (python 2.7 linux release number 249). Specifically our implementation was based on the https://graylab.jhu.edu/pyrosetta/downloads/scripts/demo/D090_Ala_scan.py demo provided in the documentation with minor modifications to fit our purposes). Protein-ligand docking was performed using smina (version from November 9, 2017). Homology modeling was done using the 9.16 version of modeller released for python 2.7. Data parsing and analysis was done using a series of custom python scripts and existing python libraries. All statistical tests were performed using existing libraries, except for the log-odds enrichment test which was manually coded in python. Relevant custom script for reproducing our work are provided on github https://github.com/hyulab/3D_SARS2. All python library dependencies are reported in the README for our repository. Protein structure visualizations were completed using the Open-source version of PyMol (version 2.2.2) and accompanying python API (version 2.1.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

The study includes an exploration of the structural interactome between SARS-CoV-2 and humans. The study is broken down into two sections; a dataset generation / presentation section and a data analysis section.

For dataset generation, we make predictions for the interface residues for 332 SARS-CoV-2 human protein-protein interactions. These are reported both as a residue-level and atomic-level resolution when appropriate structures were available to perform guided protein-protein docking. We additionally perform in-silico mutagenesis experiments to predict the impact of all mutations on the docked interfaces where applicable. Finally we performed protein-ligand docking for 30 drugs known to target human interactors of SARS-CoV-2. All results are available through our accompanying web server (http://3d-sars2.yulab.org/).

For the data analysis section, we perform additional analyses to describe these interfaces. These analyses are primarily based on two statistical tests. First, log odds ratios were used to explore the enrichment for certain biological features (e.g. population variants, disease terms, drug-binding sites) along the predicted interfaces. Second, an energy hotspot identification approach was used to normalize all comparable mutations relative to the rest of the interface and identify mutations with the most extreme predicted impact on interaction binding energy.

Research sample

The work presented herein is computational work based on publicly-accessible datasets and simulation. The relevant research sample here is a set of 332 SARS-CoV-2-human protein-protein interactions reported by Gordon et al. 2020. Working with these interactions we predict the interface residues for each interaction either at a residue- (ECLAIR) or atomic- (docking) level resolution. We consider the enrichment of sequence variation along these interfaces based on data pulled from UniProt or gnomAD. Impact of mutations along the predicted interfaces are estimated using a mutation scanning approach in PyRosetta. Individual energy hotspot mutations are identified by comparing same amino-acid mutations across the rest of the interface. Finally, potential drug-target candidates are considered through protein-drug docking experiments and comparison of overlap in protein- and drug- binding sites compared to random expectation.

Sampling strategy

Our work did not include any experiments for which a sample size could be selected. Insofar as sample sizes were relevant to our experiments they were determined based on availability of relevant information (e.g. number of population variants in a gene).

Data collection

None

Timing and spatial scale

No relevant time-based data collection was performed.

Data exclusions

No data was excluded from our results or analyses.

Reproducibility

Reproducibility for all computational simulations (docking or in-silico mutagenesis) was accounted for by running multiple independent trials with random initial parameters. For docking this was done to ensure adequate sampling of the docking space and that an optimal conformation was achieved. For energy function / mutation simulations, multiple trials were used to determine the consistency of the estimated binding affinity. Cases that did produce a significantly non-zero outcome were not considered.

Experimental results for yeast two-hybrid (Y2H) and co-immunoprecipitation (co-IP) were replicated successfully 3 times.

Randomization

Our work is primarily a computational analysis of existing data and new predictions. It included no experiments for which randomization or blinding was applicable. Small-scale experiments were carried out on an individual basis only to validate several predictions.

Blinding

Our work is primarily a computational analysis of existing data and new predictions. It included no experiments for which randomization or blinding was applicable. Small-scale experiments were carried out on an individual basis only to validate several predictions.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Monoclonal anti-FLAG M2 (Sigma, F1804) and anti-PRIM2 (abcam, ab241990) antibodies at 1:3,000 dilutions were used for co-IP experiments.

Validation

All primary antibodies used in this study were validated by the manufacturers. Validation included western blotting and immunoprecipitation from HEK-293T whole cell lysate. For details please see the manufacturer's website.

anti-FLAG - https://www.sigmaaldrich.com/US/en/product/sigma/f1804
anti-PRIM2 - https://www.abcam.com/prim2-antibody-ab241990.html?productWallTab=ShowAll

We trusted the quality of these antibodies. Additional validation for our experiments was done by confirming correct molecular weight in identified bands, and based on negative control input lane for anti-FLAG.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HEK293T cells were obtained from ATCC.

Authentication

HEK293T cells have been thoroughly tested and authenticated by ATCC. We trusted the quality of the provided product and did not explicitly perform any additional validation in house.

Mycoplasma contamination

Cell lines used tested negative for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used.