

# Identifying Human Interactors of SARS-CoV-2 Proteins and Drug Targets for COVID-19 using Network-Based Label Propagation

Jeffrey N. Law<sup>1</sup>, Kyle Akers<sup>1</sup>, Nure Tasnina<sup>2</sup>, Catherine M. Della Santina<sup>3</sup>, Meghana Kshirsagar<sup>4</sup>, Judith Klein-Seetharaman<sup>5</sup>, Mark Crovella<sup>6</sup>, Padmavathy Rajagopalan<sup>7</sup>, Simon Kasif<sup>3</sup>, and T. M. Murali<sup>2,\*</sup>

<sup>1</sup>Interdisciplinary Ph.D. Program in Genetics, Bioinformatics, and Computational Biology, Blacksburg, VA, USA

<sup>2</sup>Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

<sup>3</sup>Department of Biomedical Engineering, Boston University, Boston, MA, USA

<sup>4</sup>AI for Good Lab, Microsoft, Redmond, WA, USA

<sup>5</sup>Department of Chemistry, Colorado School of Mines, Golden, CO USA

<sup>6</sup>Department of Computer Science, Boston University, Boston, MA, USA

<sup>7</sup>Department of Chemical Engineering, Virginia Tech, Blacksburg, VA, USA

\*Corresponding author. Email: murali@cs.vt.edu

**Motivated by the critical need to identify new treatments for COVID-19, we present a genome-scale, systems-level computational approach to prioritize drug targets based on their potential to regulate host-virus interactions or their downstream signaling targets. We adapt and specialize network label propagation methods to this end. We demonstrate that these techniques can predict human-SARS-CoV-2 protein interactors with high accuracy. The top-ranked proteins**

that we identify are enriched in host biological processes that are potentially coopted by the virus. We present cases where our methodology generates promising insights such as the potential role of HSPA5 in viral entry. We highlight the connection between endoplasmic reticulum stress, HSPA5, and anti-clotting agents. We identify tubulin proteins involved in ciliary assembly that are targeted by anti-mitotic drugs. Drugs that we discuss are already undergoing clinical trials to test their efficacy against COVID-19. Our prioritized list of human proteins and drug targets is available as a general resource for biological and clinical researchers who are repositioning existing and approved drugs or developing novel therapeutics as anti-COVID-19 agents.

## Introduction

The COVID-19 pandemic has created many clinical, economic, and societal challenges world-wide. It has galvanized scientists to develop vaccines and drugs for the disease (1,2). Existing antiviral agents such as remdesivir to treat COVID-19 are already in clinical trials (3). These drugs interfere with different aspects of the viral life cycle, including fusion with the cell membrane, proteolysis, translation, and RNA replication (4). Since a virus must necessarily co-opt host cellular processes in order to replicate, an alternative attractive approach is to develop or repurpose drugs that target human proteins that the virus requires. To this end, a global, whole-genome view of host-pathogen interactions is likely to be valuable (5), especially in the case of SARS-CoV-2, given the size of its genome and the complexity of the observed clinical and epidemiological manifestations of COVID-19 (6).

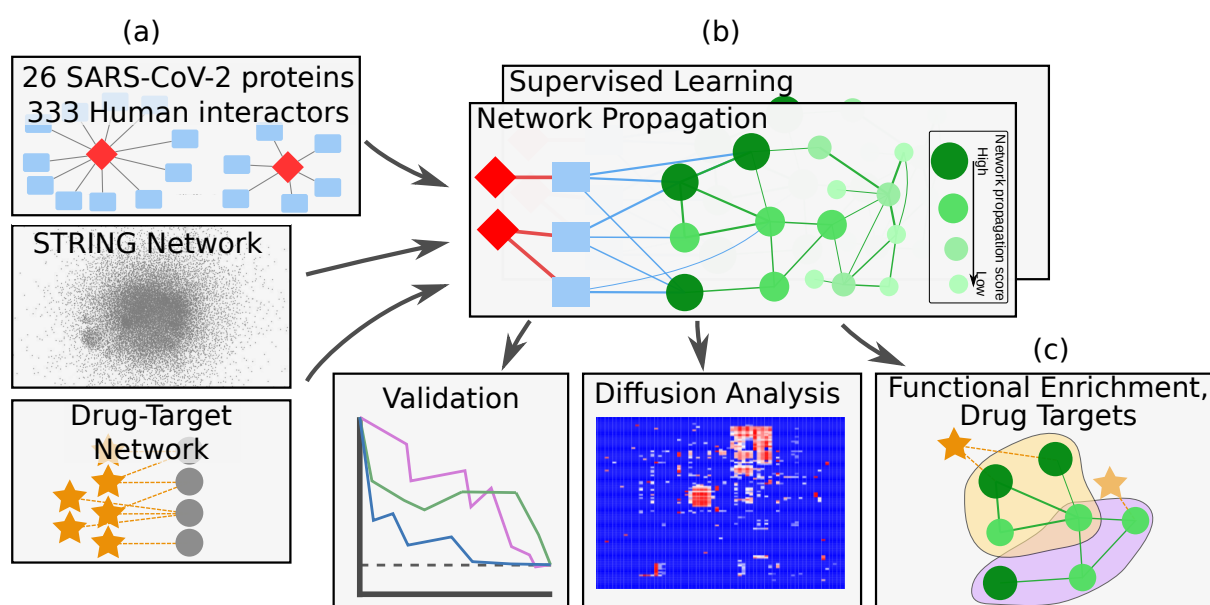


Figure 1: Our analysis framework comprising of (a) data integration: experimentally determined host-pathogen network, genome-scale functional linkage network, function ontologies, and drug-target databases; (b) computational analysis: network propagation, validation and statistical analysis, and network diffusion tracking; and (c) functional enrichment and literature-based examination of promising protein targets and drugs.

In this work, we present a strategy based on network-based functional label prediction to prioritize human proteins as drug targets and to highlight existing, approved drugs as anti-COVID-19 agents (Figure 1). We base our computational approach on the hypothesis that human proteins belonging to complexes or signaling pathways that are proximal to human interactors of viral proteins are potentially good targets for inhibition. Accordingly, we take advantage of a recently published dataset of human proteins that physically interact with SARS-CoV-2 (7). These SARS-CoV-2 interactors are entry points to host cellular processes that may be hijacked by viral infection.

Although these viral interactors are crucial starting points, we note that the proteomics pipeline used to discover them (7) may not capture *in vivo* conditions and tissue-specific interactions, leading to false negatives. Therefore, to prioritize additional human proteins, we formulate identification of downstream interactors of SARS-CoV-2 as a network labeling problem capable of tolerating uncertainty in its inputs. Specifically, given the known human protein interactors of SARS-CoV-2 proteins and a whole-genome protein interaction network, we use network propagation algorithms to predict other potential SARS-CoV-2 interactors. We first demonstrate with that this approach predicts known SARS-CoV-2 interactors with high accuracy. We further analyze highly-ranking proteins computed by these methods to identify statistically-enriched biological cellular processes and pathways that may be impacted by SARS-CoV-2. Additionally, we integrate drug-protein interactions into this framework to propose drugs that may be repositioned to treat COVID-19. We present several case studies that illustrate how drugs that target highly-ranked proteins may inhibit the virus in different stages of its life cycle. Some of these drugs are already in clinical trials for COVID-19. We emphasize that the drugs discussed in this work are *in silico* predictions that require further experimental and clinical validation before they can be used as treatments for COVID-19.

## Results

We took inspiration from the success of network propagation in diverse applications in systems biology (8). Our underlying hypothesis was that network propagation was a reasonable mechanism for predicting SARS-CoV-2 interactors. To test this hypothesis, we conducted the following validation. For each interactor of a SARS-CoV-2 protein (7), we computed the propagation score (also called “diffusion score” or “heat” in the literature) treating all the other interactors of SARS-CoV-2 as sources. To this end, we applied the Regularized Laplacian (RL) (9), a widely-used technique for semi-supervised learning (“Methods”). These scores were significantly larger than the scores of randomly-selected nodes (Figure 2(a);  $p$ -value of  $2.1 \times 10^{-78}$  for the Kolmogorov-Smirnov test, Table S1). This trend held true for other viruses as well, e.g., SARS-CoV, HIV-1, and HSV-1 (Figure 2(a);  $p$ -values of  $3.7 \times 10^{-28}$ ,  $2.1 \times 10^{-106}$ , and  $1.1 \times 10^{-49}$ , respectively), indicating that network proximity may be a general property of human proteins that bind to a virus.

### Prioritization of Potential SARS-CoV-2 Interactors

These results encouraged us to prioritize putative human protein interactors of SARS-CoV-2. In addition to RL, we used GeneMania (GM) (10), a variant of RL that has been used for finding associations between GO terms and proteins, and SinkSource (SS) (11), a related approach previously used to prioritize human proteins that are dependency factors for HIV. We also used two off-the-shelf classifiers: a Support Vector Machine with a linear kernel (SVM), and Logistic Regression (LR). For these classifiers we used the adjacency vector of each protein in the human interaction network as its feature vector. Finally, we tested Local, a method which sets each node’s score to be the weighted average of the scores of its neighbors.

We evaluated the performance of these algorithms using 5-fold cross validation of the

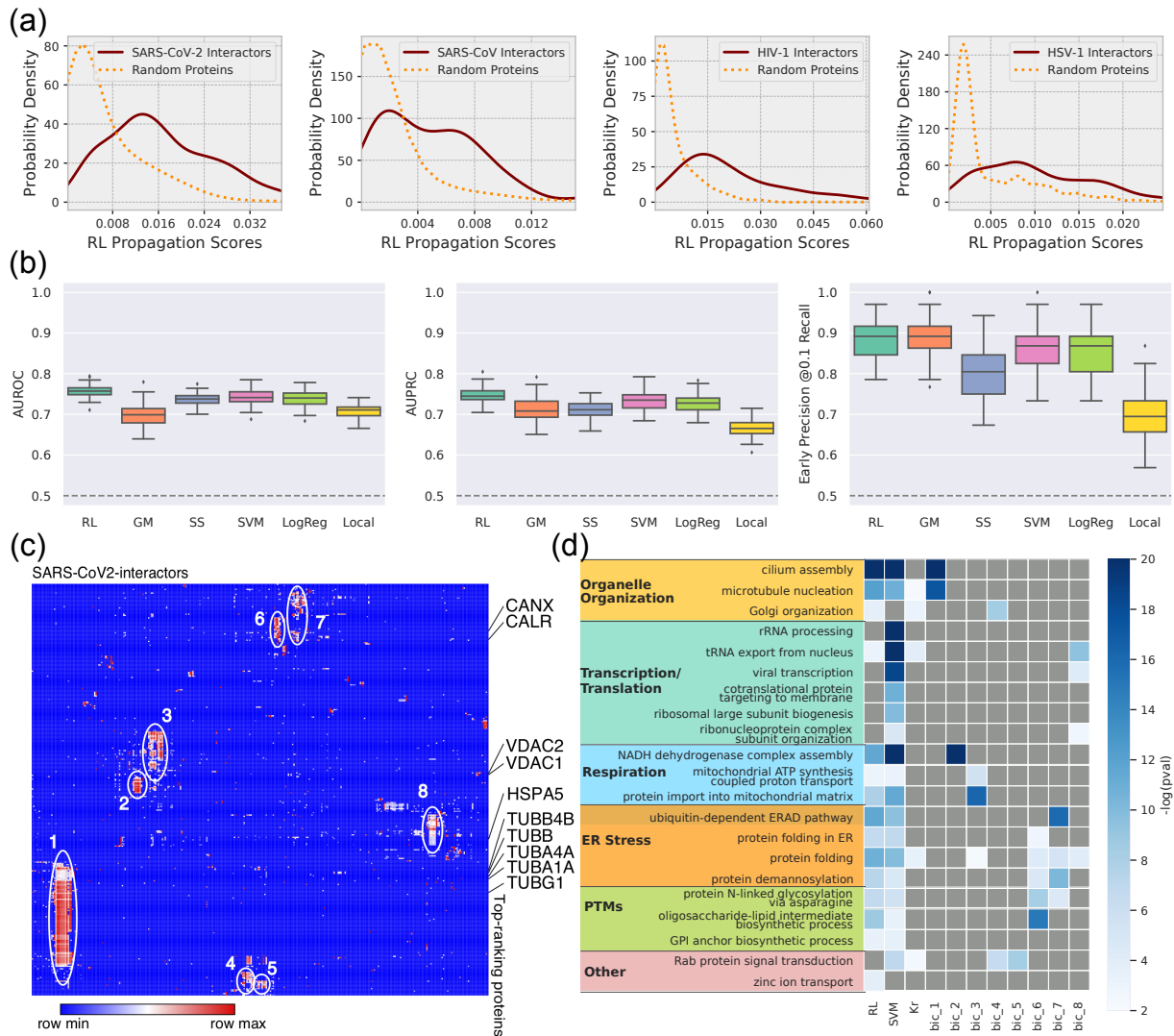


Figure 2: Network propagation results. (a) Distribution of network propagation scores among human interactors of a virus (solid curves) compared to distribution of scores of random proteins (dotted curves). The four plots are for SARS-CoV-2, SARS-CoV, HIV-1, and HSV-1, respectively. Each curve is a Gaussian kernel density estimate of the distribution. (b) Comparison of AUPRC, AUROC, and precision at 0.1 recall across six algorithms. The positive:negative ratio is one. The dashed line indicates the score for a random predictor. ‘LogReg’ is an abbreviation for logistic regression. (c) Heat map displaying contribution of each SARS-CoV-2 interactor to top-ranking proteins from RL. Ellipses highlight manually selected biclusters. Each bicluster includes a set of SARS-CoV-2 interactors that contribute high scores to a set of top-ranking proteins. We used Morpheus <https://software.broadinstitute.org/morpheus>, to generate this figure. (d) Heat map summarizing GO biological process terms enriched in top ranking proteins from RL and SVM, the biclusters highlighted in (c), and human interactors of SARS-CoV-2 proteins (indicated as ‘Kr’). We manually grouped the terms into broader categories shown in bold text. The color of a group matches the colors in Figure 3. A gray cell indicates a  $p$ -value larger than 0.01.

positive examples (human interactors of SARS-CoV-2 proteins). We used three measures of performance: the area under the receiver-operator characteristic curve (AUROC), the area under the precision-recall curve (AUPRC), and the precision at a recall of 0.1 (early precision). The third measure permitted us to estimate the accuracy of the methods for high-confidence proteins, which are likely to be the basis for experimental validation.

RL achieved a median AUROC of 0.76, a median AUPRC of 0.75, and an early precision of 0.89, values which were approximately 1.5–1.8 times superior to those of a random predictor. SVM and logistic regression achieved somewhat lower scores than RL. We obtained these results when we sampled as many negative examples as positive (see “Methods”). When we increased the sample size to five times or ten times the number of positives, RL continued to achieve the highest measures of performance (Figure S1). Local was consistently the worst-performing method, confirming the relative superiority of network propagation for this problem.

The strength of this cross-validation performance encouraged us to apply these methods to the full set of positive examples to rank the remaining proteins in the STRING network. We then used a stratified sampling approach to estimate the statistical significance of the resulting node scores (see “Methods”). For subsequent analyses, we selected one network propagation method (RL) and one supervised classifier (SVM). We chose RL because it performed the best in our evaluations, particularly with respect to early precision (Figure 2(b)). We chose SVM since it also had very good performance in cross-validation. Moreover, the protein rankings computed by the two methods were fairly dissimilar (Spearman’s correlation of 0.4 and Jaccard index of 0.3 for the top-300 ranks; Section S2.1). We considered the top 332 predictions of RL and SVM that were statistically significant at  $p \leq 0.05$  (Table S2), i.e., we selected as many top-ranking proteins as the number of experimentally-determined human proteins that interact with SARS-CoV-2

proteins (7). We selected this many proteins reasoning that that each of the SARS-CoV-2 interactors would themselves interact with or otherwise affect at least one other human protein.

We refer to these as “top-ranking proteins” below. We reasoned that due to their low overlap, using the two sets of proteins identified by RL and by SVM would result in broader coverage of human cellular processes and provide additional insights into the potential effects of SARS-CoV-2 as compared to using the results of just one method.

## **Enriched Biological Processes**

We tested for enrichment of Gene Ontology (GO) biological processes (Benjamini-Hochberg corrected  $p$ -value  $\leq 0.01$ ) among the top-ranking proteins from RL and from SVM, as well as in the interactors of SARS-CoV-2 (“Methods”). Since parent-child relationships in the GO cause many closely related terms to be enriched, we used a heuristic to select a non-redundant set of terms that were enriched in at least one of these sets of proteins (“Methods” and Section S2.4). While some terms were common to all three sets of proteins, there were many that were enriched only in our predictions (Figure 2(d), Table S3), indicating that network propagation successfully identified specific cellular processes involving proteins proximal to, but not directly interacting with, viral proteins. We observed that GO terms related to transcription and translation (second group in Figure 2(d)) were enriched in the SVM’s top-ranking proteins but not in RL’s list.

We sought to gain a deeper understanding of the trace of evidence linking a top-ranked protein to an experimentally-determined SARS-CoV-2 interactor. To this end, we computed the contribution of each SARS-CoV-2 interactor to every top-ranking prediction computed by RL (“Methods”). Visual examination of this matrix (Figure 2(c), Table S4) revealed the presence of numerous biclusters (numbered ellipses in Figure 2(c)). Each



bicluster contained a subset of SARS-CoV-2 interactors that made high contributions to the scores of a subset of top-ranking proteins. The GO biological processes enriched in the top-ranking proteins in each bicluster had a high correspondence with the terms enriched in the full set of top-ranking proteins (Figure 2(d)). Thus, these biclusters provide useful information on which human proteins the virus may use as “gateways” to target host cellular processes. We further study these biclusters and the largest contributors to individual top-ranking proteins in Section S2.3.

Our top-ranking proteins were enriched in five broad categories of GO biological processes: organelle organization, transcription and translation, respiration, ER stress, and post-translational modifications (Figure 2(d)). In addition, two processes did not belong to any of these categories. Figure 3 provides a high-level view of these GO terms and their connections in the STRING network to SARS-CoV-2 interactors and SARS-CoV-2 proteins. In the rest of this section, we examine the relevance of some of these processes to the viral cell cycle. In subsequent sections, we consider other enriched processes from the viewpoint of identifying promising drug targets.

One group of enriched terms are related to protein translation (green nodes in Figure 3) including “rRNA processing” ( $p$ -value  $3.96 \times 10^{-26}$  for SVM and 0.05 for SARS-CoV-2 interactors), “ribonucleoprotein complex subunit organization” ( $p$ -value  $1.75 \times 10^{-5}$  for SVM and 0.48 for SARS-CoV-2 interactors), and “ribosomal large subunit biogenesis” ( $p$ -value  $1.12 \times 10^{-10}$  for SVM and 0.69 for SARS-CoV-2 interactors). RNA viruses appear to reduce protein synthesis in host cells, including in the case of SARS-CoV (12). A recent proteomics study that infected Caco-2 cells with SARS-CoV-2 (13) revealed that while global translation rates in host cells exhibited only minor changes, the levels of a significant number of human proteins involved in translation were highly positively correlated with levels of key viral proteins. This study further showed that translation

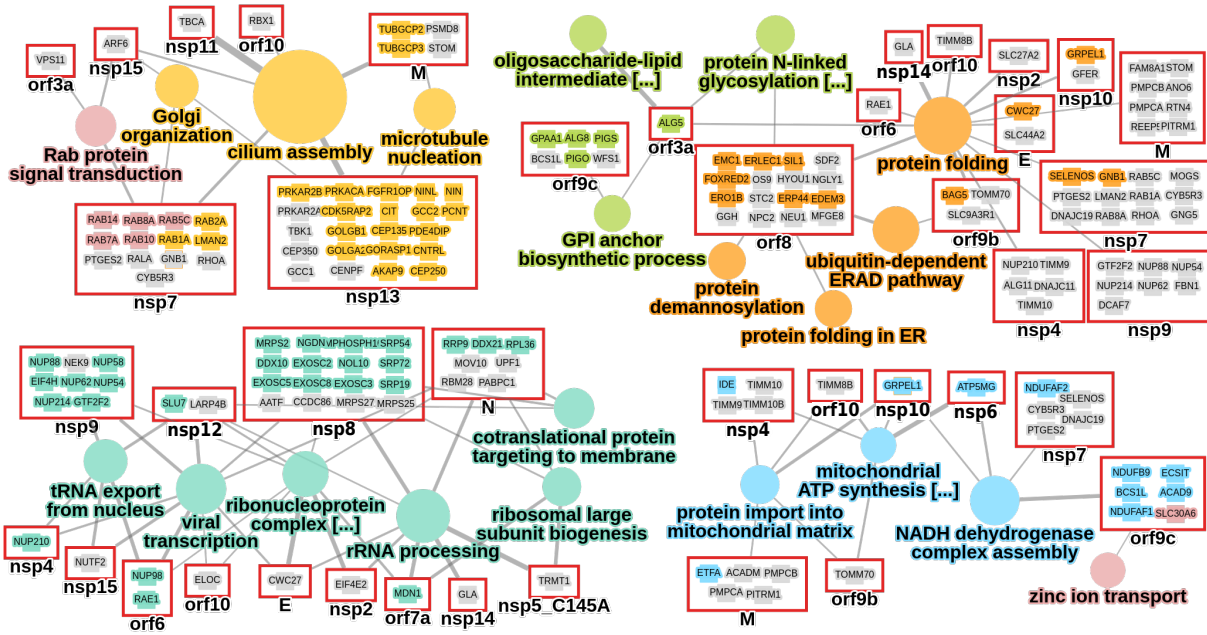


Figure 3: Overview of connections between SARS-CoV-2 interactors and the GO biological process terms enriched in the top-ranking proteins. Each circle is a GO term shown in Figure 2(d); the color of the GO term corresponds to its group in Figure 2(d) and its size corresponds to the number of annotated top-ranking proteins. Each box with a red border is a SARS-CoV-2 protein (the box labels). The box contains one or more human proteins that interact with that viral protein. A viral protein may label more than one box. A gray edge connects a red box to a GO term. The width of the edge is proportional to the fraction of the human interactors of the viral protein that are connected in the STRING network to the top-ranking proteins annotated to the term (considering only STRING edge weights  $\geq 900$ ). We do not show edges with a fraction  $< 0.05$ , except for the edge between orf9c and “zinc ion transport” (0.026). If a viral interactor is annotated to a GO term, its color is the same as that of the GO term.

inhibitors prevented SARS-CoV-2 replication. The authors concluded that SARS-CoV-2 may increase production of proteins involved in the translation machinery components so as to compensate for inhibition of translation in host cells.

The GO term “viral transcription” is enriched in top-ranking SVM proteins ( $p$ -value  $3.44 \times 10^{-19}$ ) but not in SARS-CoV-2 interactors ( $p$ -value 0.05). The proteins involved are primarily nucleoporins and members of the large ribosomal subunit. An overwhelming fraction of the annotations to this GO term come from the Reactome pathway “Viral Messenger RNA Synthesis” (R-HSA-168325). This pathway describes how the influenza A virus co-opts the host machinery to transcribe its mRNAs. Thus the enrichment of this GO term in the top-ranking SVM proteins may indicate a similar exploitation of the host by SARS-CoV-2.

N-linked glycosylation is an important post-translational protein modification that is conserved across eukaryotes (14). It involves the attachment of an oligosaccharide to an asparagine residue of a protein (14). The process takes place in two major stages. The first step involves the biosynthesis of lipid-linked oligosaccharides (14) corresponding to the GO term “oligosaccharide-lipid intermediate biosynthetic process” (Figure 4(a)), which is enriched in top-ranking RL proteins ( $p$ -value  $1.77 \times 10^{-9}$ ) but not in SARS-CoV-2 interactors ( $p$ -value 0.35). In the second step, oligosaccharyltransferases enable the transfer of the oligosaccharide to the asparagine residue. These proteins are annotated to the GO term “protein N-linked glycosylation via asparagine”, which is enriched in top-ranking RL proteins ( $p$ -value  $2.06 \times 10^{-7}$ ) but not in SARS-CoV-2 interactors ( $p$ -value 0.2).

Several viruses can exploit the host glycosylation pathway to modify viral proteins in order to enter host cells and evade the immune response (15). An extensive glycan shield that covers the S protein in human coronavirus NL63 (HCoV-NL63) masks the protein

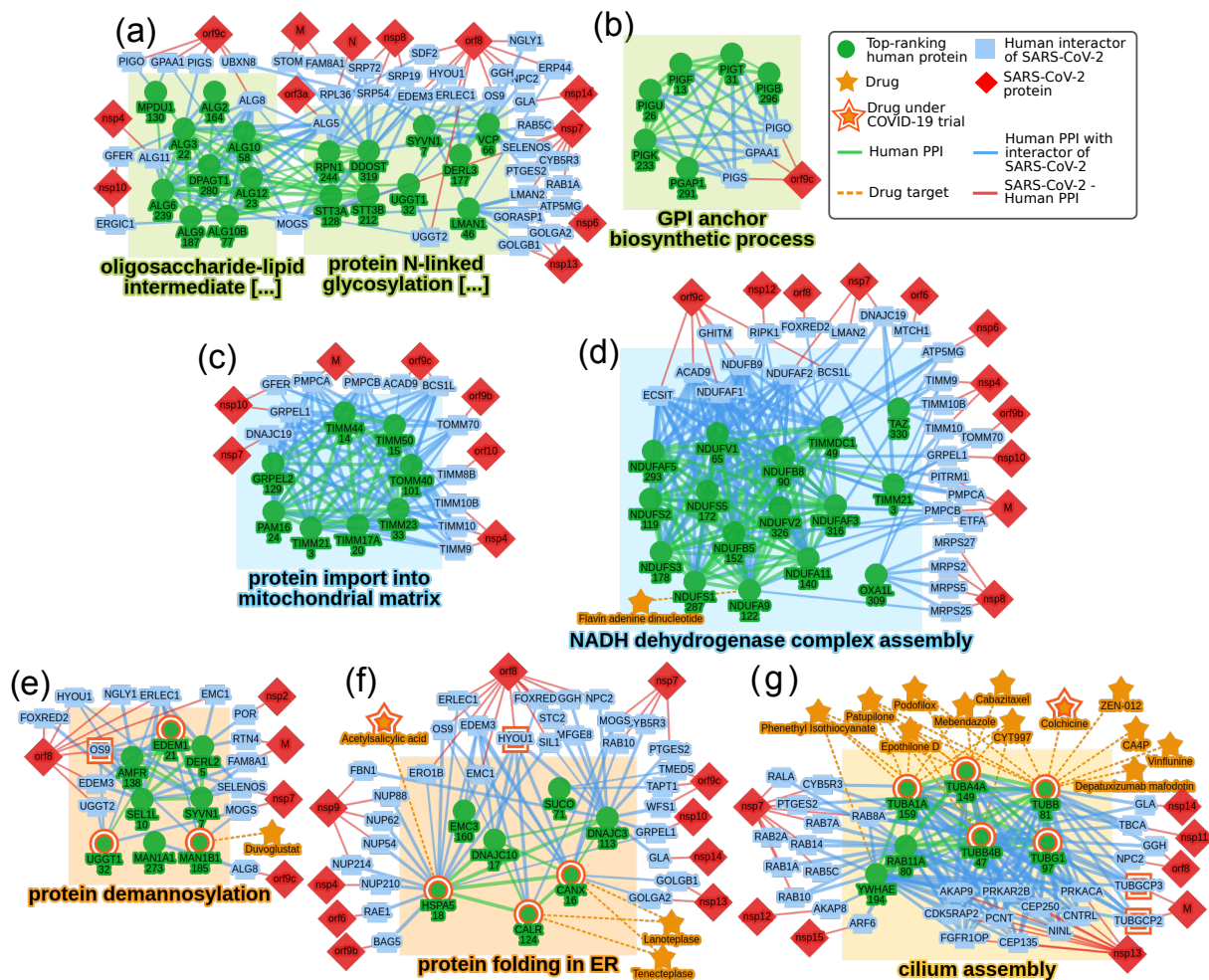


Figure 4: Networks of the top-ranking proteins for RL (green nodes) that are annotated to the enriched terms (a) “oligosaccharide-lipid intermediate biological process” and “protein N-linked glycosylation via asparagine,” (b) “GPI anchor biosynthetic process,” (c) “protein import into mitochondrial matrix,” (d) “protein demannosylation,” (e) “protein folding in ER,” (f) “cilium assembly,” or (g) “NADH dehydrogenase complex assembly.” The number below the name of a green protein is its rank in RL. A rectangle encompasses proteins (top-ranking or SARS-CoV-2 interactors) annotated to the respective term. Proteins discussed in the text are highlighted with a red border. For each top-ranking protein, we also display its connections with neighboring SARS-CoV-2 interactors. We removed STRING edges with weight  $< 700$  to simplify the visualization. In (d), we limited proteins to those that are the target of an approved or investigational drug. In (f), we removed drugs that promote clotting.

surface, thus limiting access to neutralizing antibodies (16). The receptor-binding and fusion-peptide domains of the S protein (fusion protein) in SARS-CoV-2, SARS-CoV, and MERS-CoV contain N-glycans (17,18). It appears that the N-linked glycan modifications of SARS and MERS CoV S proteins are extensive but do not form as effective a shield as in viruses such as HIV-1 (17). Drugs that target host N-linked glycosylation pathways have been proposed as treatments for COVID-19 (19).

The term “(GPI)-anchor biosynthetic process” is significantly enriched in the top-ranking proteins from RL ( $p$ -value  $4 \times 10^{-4}$ ) but not in the human interactors of SARS-CoV-2 ( $p$ -value 0.21). The proteins prioritized by our methods (Figure 4(b)) are either components of the Glycosylphosphatidylinositol (GPI) transamidase complex or transfer GPI to proteins during the synthesis of GPI anchors; these anchors tether proteins to lipid bilayers (20). GPI-anchored proteins are often associated with lipid rafts, which are microdomains in plasma membranes that are enriched with cholesterol and sphingolipids (20). Lipid rafts play a major role in viral entry, assembly, replication, and budding. They are known to be involved in the entry of SARS-CoV into host cells (21). Thus, the proteins involved in (GPI)-anchor biosynthesis that we prioritize may provide a deeper understanding on SARS-CoV-2 entry into host cells.

Several GO terms related to cellular respiration (blue nodes in Figure 3) were enriched in the top-ranking proteins, e.g., “protein import into mitochondrial matrix” (Figure 4(c);  $p$ -value of  $1.21 \times 10^{-8}$  for RL and 0.3 for SARS-CoV-2 interactors). Many of the proteins annotated to this term are translocases of the inner and outer membranes of the mitochondrion. These proteins comprise complexes that transport proteins across the mitochondrial membranes. “NADH dehydrogenase complex assembly” is a related GO biological process (Figure 4(d)), which has a  $p$ -value of  $2.8 \times 10^{-12}$  for RL in contrast to a  $p$ -value of 0.035 in SARS-CoV-2 interactors. NADH dehydrogenase is the first complex

in the mitochondrial respiratory chain. Viral infections have been proposed to cause the Warburg effect that is frequently associated with cancer (22). Due to oxygen deprivation, cells switch from respiration to glycolysis. We note that a recent genome-wide CRISPR screen for human genes that regulate SARS-CoV-2 infection (23) identified several members of the respiratory chain as “anti-viral”, i.e., the knockout of the gene sensitizes the host cell to virus-induced cell death. Thus, the GO terms enriched in our results may suggest mechanisms by which SARS-CoV-2 modulates cellular respiration in the host to promote infection.

## **The Role of Endoplasmic Reticulum Stress, HSPA5, and Anti-Clotting Drugs**

We now describe the connection unearthed by our analysis between interactors of SARS-CoV-2, proteins involved in endoplasmic reticulum (ER) stress (orange nodes in Figure 3, and anti-clotting drugs . The GO biological process “protein demannosylation”, the removal of a mannose group from a protein, is enriched in top-ranking RL proteins ( $p$ -value  $4.48 \times 10^{-8}$ ) but not in SARS-CoV-2 interactors ( $p$ -value 0.08). Key top-ranking proteins in this network (Figure 4(e)) include MAN1B1 and UGGT1, which are localized to the ER. Enveloped viruses have been shown to hijack the host cell’s ER for survival and replication, thereby causing ER-associated degradation (ERAD) (24). MAN1B1 plays a role in protein quality control as well as in their degradation (25). The utilization of UGGT1 by enteroviruses leads to increased viral replication and higher pathogenicity (26).

Viral protein Orf8 interacts with human OS9 (an ER degradation protein), which in turn interacts with top-ranking protein EDEM1 (ER degradation-enhancing alpha-mannosidase-like protein 1). EDEM1 controls the degradation of mis-folded glycoproteins and is a key player in ERAD (27) When EDEM1 is cleared from the ER, it is transported

out of the organelle in nanoscale vesicles that are called EDEMosomes. Coronaviruses such as the mouse hepatitis virus subsequently hijack these organelles and remodel them to form double membrane vesicles (DMVs) (28). DMVs are widely used by viruses, including coronaviruses, to promote viral replication (29).

The GO biological process “protein folding in endoplasmic reticulum” was also enriched in the top-ranking proteins ( $p$ -value  $2.29 \times 10^{-7}$  for RL and 0.28 for interactors of SARS-CoV-2). Viral protein nsp9 interacts with several human nucleoporins (NUP54, NUP62, NUP88 and NUP214) (7), which in turn interact with heat shock protein HSPA5 in the STRING network (Figure 4(f)). HSPA5, also referred to as glucose regulated protein (GRP78) or immunoglobulin binding protein (BiP) in the literature, is evolutionarily conserved from prokaryotes to humans (30). It has a repertoire of functions associated with ER stress response. HSPA5 is usually localized in the ER. When the ER is stressed, HSPA5 can translocate to the cell surface, the nucleus and mitochondria (31, 32). On the cell surface, HSPA5 plays a multi-functional role in cell proliferation, cell viability, apoptosis, and regulation of innate and adaptive immunity (32, 33).

HSPA5 has been proposed as a universal target for human diseases (34). It has increasingly well-documented essential interactions and activities during viral infections. In particular, the role of HSPA5 in viral entry and pathogenesis has been widely investigated. As a cell surface protein, HSPA5 has been reported to play an important role in viral entry (35, 36). SARS-CoV-2 infection has been shown to lead to ER stress and the up-regulation of HSPA5 (35, 36). The S protein of SARS-CoV-2 can induce transcriptional activation of HSPA5 (36). This protein can serve as a point of attachment for both MERS-COV and bat coronavirus (bCOV HKU9) (37). Both Zika virus and Japanese encephalitis virus use HSPA5 to prevent apoptosis and to help in viral replication (38). A recent molecular docking study has predicted HSPA5 as a potential receptor

for the SARS-CoV S protein (39). Based on its observed expression *in vitro* in airway epithelial cells, it has been suggested that HSPA5 may serve as an additional receptor for SARS-CoV-2 in these cells (40). Based on our network-based analysis and support in the literature, we hypothesize that HSPA5 may serve as a co-receptor, a point of viral attachment, or aid in viral entry of SARS-CoV-2.

Blood hypercoagulability is reported to be common among COVID-19 patients (41). We now present the linkage between the “protein folding in ER” network (Figure 4(f)) and coagulation. We also suggest how anti-coagulant drugs may function within the context of COVID-19. SARS-CoV-2 protein Orf8 interacts with human hypoxia up-regulated protein 1 (HYOU1). HYOU1 (also known as GRP170) plays a cytoprotective role in response to oxygen deprivation (30). In addition to HSPA5, HYOU1 interacts with calnexin (CANX) and calreticulin (CALR), both of which are chaperone proteins found in the ER (42). HSPA5, CANX and CALR act as protein chaperones for pro-coagulant proteins such as Factor V and Factor VIII. Once Factor VIII is secreted, it binds to another pro-coagulant protein von Willebrand factor (vWF) to prevent degradation of clots (43). Although Factor V, Factor VIII, and vWF are not among the top-ranking proteins and thus do not appear in Figure 4(f), this network is suggestive of mechanisms that SARS-CoV-2 may use to cause abnormal blood coagulation.

Anti-coagulant drugs that interact with HSPA5, CANX or CALR include Tenecteplase, a third generation plasminogen activating enzyme and the investigational drug Lanoteplase, which is a serine protease that binds to fibrin leading to the formation of plasmin (44), an enzyme that breaks clots. Lanoteplase is a second-generation derivative of alteplase, and a third generation derivative of recombinant plasminogen. It is notable that there is a clinical trial for Alteplase (ClinicalTrials.gov, NCT04357730) to test its effectiveness in treating COVID-19. Calcium citrate, a known anti-coagulant agent also interacts



with CANX and CALR (45). Aspirin, also present in (Figure 4(f)), binds to and inhibits the ATPase activity of HSPA5 (46). A clinical trial (ClinicalTrials.gov, NCT04363840) is testing whether early treatment of COVID-19 patients with aspirin and vitamin D can inhibit the production of blood clots and decrease rates of hospitalization.

## Cilium Assembly and Tubulin-Modulating Drugs

GO biological processes related to cilia were significantly enriched in the top-ranking predictions RL and SVM predictions (yellow nodes in Figure 4(g), e.g., “cilium assembly,”  $p$ -value  $9.16 \times 10^{-52}$  for RL) but not in the human interactors of SARS-CoV-2 ( $p$ -value 0.31). Many proteins annotated to this term belong to the tubulin family, which are components of microtubules. The SARS-CoV-2 M protein binds to two  $\gamma$ -tubulins (TUBGCP2 and TUBGCP3), which interact with several  $\alpha$ - and  $\beta$ -tubulins among the top 332 predictions (Figure 4(g)). Microtubules are polymers that provide shape and structure to eukaryotic cells and are necessary in cell transport and cell division, among other functions (47).  $\alpha$ - and  $\beta$ -tubulins compose microtubule filaments, while  $\gamma$ -tubulins connect them to the microtubule organizing center.

Viruses commonly utilize microtubules for cellular entry, intra-cellular trafficking, and exit from cells (48). For instance, the S protein of human  $\alpha$ -coronavirus interacts with tubulin  $\alpha$  and  $\beta$  chains (49), suggesting that tubulin may be involved in the transport and localization of the S protein and its assembly into virions (49). Relevant to SARS-CoV-2, microtubules are the primary structural component of cilia, which line epithelial cells in the respiratory tract and are responsible for the transport of mucus out of cells (50). The ACE2 receptor that SARS-CoV-2 uses to enter cells appears to be expressed primarily on the cilia of respiratory tract epithelial cells (51, 52), further implicating microtubules in viral infection. The combination of high expression levels of ACE2 and the presence of

cilia may also explain the detection of the virus in multiple organs (53) and the deleterious effect of COVID-19 on the renal, gastrointestinal, and olfactory systems (54).

Next, we turned our attention to drugs that targeted Tubulin proteins (Figure 4(g)). Our methods highlighted the drug Colchicine since it had the smallest  $p$ -value (0.001) among the approved or investigational drugs that target proteins annotated to “ciliary assembly.” Colchicine is FDA-approved for treating gout. Its effectiveness against COVID-19 is being tested in thirteen ongoing clinical trials (clinicaltrials.gov).

The drugs that target proteins involved in ciliary assembly (Figure 4(g)) are mostly anti-mitotic agents, which are also being investigated as anti-cancer therapeutics. These drugs fall into three broad classes depending on how they affect the dynamic equilibrium between unpolymerized and polymerized tubulin (55). Colchicine, Podofilox, Vinchristine, Vinflunine, and Vinblastine destabilize microtubules by binding to specific sites of  $\alpha$ - and  $\beta$ -tubulins, thereby preventing their assembly. Albendazole, Mebendazole, and Oxibendazole, Cabazitaxel, Milataxel, and CYT99 (a synthetic drug) inhibit tubulin polymerization thereby affecting the assembly and dynamics of microtubules and slowing cell growth. Two other drugs in this network, Patupilone and Epithilone-D, also cause cell-cycle arrest but by a different mechanism: they stabilize the tubulin network. In the case of COVID-19, we hypothesize that drugs that destabilize the microtubule network might prove to be more efficacious since they may prevent viral proteins from using microtubules for intra-cellular trafficking. In addition, as in the case of Colchicine, they may act as anti-inflammatory agents and provide additional benefits to COVID-19 patients.

## Discussion

Drug development is widely acknowledged to be one of the most challenging industrial processes requiring billions of dollars and high levels of diverse expertise to overcome pro-

found scientific and logistic barriers. There remains a high failure rate in the steps that lie between discovering a drug target and manufacturing a drug for market (56). The COVID-19 pandemic and its medical and economic impact created a new and urgent challenge for research and pharmaceutical communities to develop a therapeutic response (4). As one manifestation of this community response, a recent proteomics effort by a consortium of scientists generated a first-of-a-kind interactome associated with the SARS-CoV-2-human interface (7). This interactome inspired a large-scale drug repositioning effort targeting the human proteins that directly interact with the virus.

This important and timely advance signals a very promising direction in drug development for COVID-19. The set of human proteins reported to interact with SARS-CoV-2 is likely to have both false positives and false negatives due to the properties of the proteomic screening pipeline used. Thus, we sought to further extend the results of this study and significantly expand the resources available to the drug development and repositioning community through the use of state-of-the-art network prediction algorithms. We also aimed to demonstrate that the area of function prediction using network propagation has achieved significant technological maturity since its introduction over 15 years ago (57, 58, 59, 60). In particular, network propagation-based approaches to implement label propagation have evolved to be relatively easy to implement and validate statistically (8), thereby making them amenable for translational research.

In this work, we take full advantage of these mature predictive platforms to significantly expand the resources available to the COVID-19 community by producing a larger set of putative SARS-CoV-2 interactors. Many of these proteins could serve as drug targets and several are already under development or in clinical trials. Our methods predict interactors with high accuracy. We expect this performance to be further improved by the computational biology community building on our and other resources.

Our initial findings suggest existing as well as new drug targets. Several targets are associated with ER stress, a well-documented initial response to viral infections with lethal or severe clinical outcomes. These ER stress response proteins include a number of well-studied heat shock proteins (e.g., HSPA9, rank 9,  $p$ -value 0.031 and HSPA5, rank 18,  $p$ -value 0.035 for RL, Table S1). We have discussed one of them, HSPA5, as a promising drug target. Other authors have also noted the potential for HSPA5 in the context of COVID-19 (61). Other HSPs in our prediction list are also natural targets for drug repositioning. More broadly, targeting ER stress is a promising direction to reduce cell death (62).

In addition to the primary drug targets we have identified in “Results”, we point out VDAC proteins, which are voltage-dependent anion channels (e.g., VDAC2, rank 99,  $p$ -value 0.035 and VDAC1, rank 182,  $p$ -value 0.037 for RL) that have been associated with mitochondria-triggered pro-apoptotic processes. We speculate that several recently identified inhibitors of VDAC1 may be considered as therapeutic interventions to inhibit anion transport channels. An example is DIDS, which leads to a potential decrease in ROS-associated cell death (63).

We have noted that a number of dehydrogenases appear among our top ranking proteins. This finding naturally promotes NADH as a promising drug target to treat SARS-CoV-2 infection. NADH is also associated with metabolic health and has been studied as an energy currency, suggesting a number of promising biological follow-ups. The Warburg effect mentioned earlier is targeted by a number of metabolic drugs, which offer new directions for viral treatments. We also note the classical connection between elevated inflammation and metabolic slowdown associated with both aging and viral infections (64). Thus, a combined immuno-metabolic combinatorial drug regimen might be worth considering for COVID-19 patients to prevent severe outcomes.

We have already begun to integrate our current techniques both with other omics data and with orthogonal methods to predict more biologically meaningful networks and processes impacted by the virus. In particular, single-cell RNAseq data offer many opportunities to examine cellular heterogeneity and context-specific interactions. In this context, we note complementary efforts to repurpose drugs for SARS-CoV-2 that are based on protein structures (65), observational studies of treatments being administered to patients (66), and shortest paths in protein networks (67).

In summary, this paper provides a significant new resource for drug repurposing and development to the COVID-19 community. The relatively quick turn-around of this project and the host-virus protein interaction resource on which we have built (7), demonstrates that biological network science and network propagation (8) have achieved significant maturity, as further evidenced by their strong presence in cancer and chronic diseases. These efficient computational methods and rapidly generated data have allowed us to develop high accuracy predictions of proteins involved in specific functional activities as an expeditious response to this global pandemic.

## References

1. J. M. Sanders, M. L. Monogue, T. Z. Jodlowski, J. B. Cutrell, *JAMA - Journal of the American Medical Association* (2020).
2. M. P. Lythgoe, P. Middleton, *Trends in Pharmacological Sciences* (2020).
3. G. Li, E. De Clercq, *Nat Rev Drug Discov* **19**, 149 (2020).
4. R. K. Guy, R. S. DiPaola, F. Romanelli, R. E. Dutch, *Science* **368**, 829 (2020).
5. C. V. Forst, *Infectious Disease Informatics* (Springer, 2010), pp. 123–147.

6. M. Z. Tay, C. M. Poh, L. Rénia, P. A. MacAry, L. F. Ng, *Nature Reviews Immunology* pp. 1–12 (2020).
7. D. E. Gordon, *et al.*, *Nature* (2020).
8. L. Cowen, T. Ideker, B. J. Raphael, R. Sharan, *Nature Reviews Genetics* **18**, 551 (2017).
9. D. Zhou, B. Schölkopf, *ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields* (2004), pp. 132–137.
10. S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, Q. Morris, *Genome Biology* **9**, S4 (2008).
11. T. M. Murali, M. D. Dyer, D. Badger, B. M. Tyler, M. G. Katze, *PLoS computational biology* **7**, e1002164+ (2011).
12. W. Kamitani, C. Huang, K. Narayanan, K. G. Lokugamage, S. Makino, *Nat. Struct. Mol. Biol.* **16**, 1134 (2009).
13. D. Bojkova, *et al.*, *Nature* (2020).
14. J. Breitling, M. Aebi, *Cold Spring Harb Perspect Biol* **5**, a013359 (2013).
15. D. J. Vigerust, V. L. Shepherd, *Trends Microbiol.* **15**, 211 (2007).
16. A. C. Walls, *et al.*, *Nat. Struct. Mol. Biol.* **23**, 899 (2016).
17. Y. Watanabe, *et al.*, *Nat Commun* **11**, 2688 (2020).
18. Y. Watanabe, J. D. Allen, D. Wrapp, J. S. McLellan, M. Crispin, *bioRxiv* (2020).
19. S. J. Williams, E. D. Goddard-Borger, *Biochem. Soc. Trans.* (2020).

20. C. Metzner, B. Salmons, W. H. Günzburg, J. A. Dangerfield, *Virology* **382**, 125 (2008).
21. Y. Lu, D. X. Liu, J. P. Tam, *Biochem. Biophys. Res. Commun.* **369**, 344 (2008).
22. E. L. Sanchez, M. Lagunoff, *Virology* **479-480**, 609 (2015).
23. J. Wei, *et al.*, *bioRxiv* (2020).
24. S. E. Bettigole, L. H. Glimcher, *Annu. Rev. Immunol.* **33**, 107 (2015).
25. Y. Wu, M. T. Swulius, K. W. Moremen, R. N. Sifers, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 8229 (2003).
26. P. N. Huang, *et al.*, *PLoS Pathog.* **13**, e1006375 (2017).
27. Y. Yoshida, K. Tanaka, *Biochim. Biophys. Acta* **1800**, 172 (2010).
28. F. Reggiori, *et al.*, *Cell Host Microbe* **7**, 500 (2010).
29. G. Wolff, C. E. Melia, E. J. Snijder, M. Bárcena, *Trends Microbiol.* (2020).
30. A. S. Lee, *Nat. Rev. Cancer* **14**, 263 (2014).
31. Y. Zhang, R. Liu, M. Ni, P. Gill, A. S. Lee, *J. Biol. Chem.* **285**, 15065 (2010).
32. Y. L. Tsai, *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4245 (2018).
33. M. Ni, Y. Zhang, A. S. Lee, *Biochem. J.* **434**, 181 (2011).
34. L. Booth, *et al.*, *J. Cell. Physiol.* **230**, 1661 (2015).
35. M. L. DeDiego, *et al.*, *PLoS Pathog.* **7**, e1002315 (2011).
36. C. P. Chan, *et al.*, *J. Virol.* **80**, 9279 (2006).

37. H. Chu, *et al.*, *J. Biol. Chem.* **293**, 11709 (2018).
38. H. R. Lyoo, S. Y. Park, J. Y. Kim, Y. S. Jeong, *Viol. J.* **12**, 32 (2015).
39. I. M. Ibrahim, D. H. Abdelmalek, M. E. Elshahat, A. A. Elfiky, *J. Infect.* **80**, 554 (2020).
40. J. A. Aguiar, *et al.*, *bioRxiv* (2020).
41. E. Terpos, *et al.*, *Am. J. Hematol.* (2020).
42. D. B. Williams, *J. Cell. Sci.* **119**, 615 (2006).
43. R. J. Kaufman, S. W. Pipe, L. Tagliavacca, M. Swaroop, M. Moussalli, *Blood Coagul. Fibrinolysis* **8 Suppl 2**, 3 (1997).
44. M. Flemmig, M. F. Melzig, *J. Pharm. Pharmacol.* **64**, 1025 (2012).
45. K. G. Mann, M. F. Whelihan, S. Butenas, T. Orfeo, *J. Thromb. Haemost.* **5**, 2055 (2007).
46. W. G. Deng, K. H. Ruan, M. Du, M. A. Saunders, K. K. Wu, *FASEB J.* **15**, 2463 (2001).
47. E. Nogales, *Annu. Rev. Biochem.* **69**, 277 (2000).
48. U. F. Greber, M. Way, *Cell* **124**, 741 (2006).
49. A. T. Rüdiger, *et al.*, *Virology* **497**, 185 (2016).
50. P. Satir, S. T. Christensen, *Annu. Rev. Physiol.* **69**, 377 (2007).
51. I. T. Lee, *et al.*, *medRxiv* (2020).



52. W. Sungnak, *et al.*, *Nat. Med.* **26**, 681 (2020).
53. V. G. Puelles, *et al.*, *N. Engl. J. Med.* (2020).
54. C. Huang, *et al.*, *Lancet* **395**, 497 (2020).
55. Y. Lu, J. Chen, M. Xiao, W. Li, D. D. Miller, *Pharmaceutical Research* (2012).
56. S. Pushpakom, *et al.*, *Nat Rev Drug Discov* **18**, 41 (2019).
57. M. Deng, K. Zhang, S. Mehta, T. Chen, F. Sun, *J Comput Biol* **10**, 947 (2003).
58. A. Vazquez, A. Flammini, A. Maritan, A. Vespignani, *Nat Biotechnol* **21**, 697 (2003).
59. S. Letovsky, S. Kasif, *Bioinformatics* **19 Suppl 1**, i197 (2003).
60. U. Karaoz, *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2888 (2004).
61. D. P. Ha, R. Van Krieken, A. Carlos, A. S. Lee, *J. Infect.* (2020).
62. J. D. Mallotra, *et al.*, *Proceedings of the National Academy of Sciences* **105**, 18525 (2008).
63. D. Ben-Hail, V. Shoshan-Barmatz, *Biochim. Biophys. Acta* **1863**, 1612 (2016).
64. G. Pawelec, D. Goldeck, E. Derhovanessian, *Curr. Opin. Immunol.* **29**, 23 (2014).
65. C. Wu, *et al.*, *Acta Pharmaceutica Sinica B* (2020).
66. M. Vaduganathan, *et al.*, *New England Journal of Medicine* **382**, 1653 (2020).
67. Y. Zhou, *et al.*, *Cell discovery* **6**, 1 (2020).

## Data and Software Availability

Our software is available under the GNU Public License version 3 at <https://github.com/Murali-group/SARS-CoV-2-network-analysis>. The networks in Figures 3 and 4 are available for visualization and download on GraphSpace at <http://graphspace.org/graphs/?query=tags:2020-sarscov2-network-analysis>.

We used publicly available datasets for our analysis.

## Acknowledgments

TMM acknowledges support from NSF grants DBI-1759858 and MCB-1817736. KA acknowledges support from the Genetics, Bioinformatics, and Computational Biology program at Virginia Tech. JK acknowledges support from NSF grant CCF-2029543. MC acknowledges support from NSF grant CNS-1618207. CMDS acknowledges support from the Hariri Institute and the Department of Biomedical Engineering at Boston University. PR acknowledges support from NSF grant CBET-1510920 and USDA-NIFA grant 2018-07578. PR and TMM acknowledge support from the Computational Tissue Engineering Graduate Education Program at Virginia Tech.

The authors wish to thank S. Alabdullatif, S. Alshuaib, M. Iennaco, M. Kouzminov, S. Murthy, S. Makwana, N. Naguib, C. Taglietti, and M. Zanna for exploratory research on this data and insightful and thought-provoking analysis. We also thank Roded Sharan, Noga Alon, Dan Lancour and Rich Roberts for discussions that helped formulate the techniques and ideas we used in this paper.

# Supplementary materials

## Methods

### Algorithms

We describe each of the algorithms we use for label propagation and prediction. We are given a weighted, undirected network  $G = (V, E, w)$  and a set  $P$  of positive examples. Each node in  $G$  is a human protein and each edge represents a physical or functional interaction between two proteins. We describe the network we use in more detail in “Datasets”. The set  $P$  comprises of human proteins that interact with SARS-CoV-2 proteins (7). We seek to compute a score  $s(v) \geq 0$  for every node in  $G$  that indicates our confidence that  $v$  either physically interacts with or is functionally linked to a SARS-CoV-2 protein.

**Regularized Laplacian (RL) (9).** We first provide some intuition behind this method. We seek to compute a diffusion score that represents the probability that a given node in the network will be reached by a random walk of a given length from a set of seed nodes (the human interactors of SARS-CoV-2); the parameter  $\alpha$  introduced below governs the length of the random walk. Biologically, this model captures the stochasticity in protein interactions in signaling networks or complexes that depends on many factors such as concentrations, conditions, co-factors, chaperones, ATP levels, etc. Thus, any given signaling cascade may be modelled as a random walk that will reach its target with some probability. In our context, the higher the diffusion score our algorithms compute for a node, the more likely it is that the node may be involved in the host response to the virus or be used by the virus in its life cycle.

Formally, this method defines a label vector  $\vec{y}$  over the nodes in  $G$  where  $y(u) = 1$  if

node  $u$  is a SARS-CoV-2 interactor and  $y(u) = 0$ , otherwise. It computes a score  $s(u)$  between 0 and 1 for each protein  $u$  in  $G$ . Let  $W \in \mathbb{R}^{n \times n}$  denote the adjacency matrix of  $G$  and  $\tilde{W} = D^{-1/2}WD^{-1/2}$  denote the normalized network, where  $D$  is a diagonal matrix with  $D_{uu} = \sum_v a_{uv}$ , for every node  $u$  in  $G$ . Then, to compute the scores for each node, we minimize the following sum, where  $\alpha > 0$  is a parameter:

$$\sum_{u \in V} (s(u) - y(u))^2 + \alpha \sum_{(u,v) \in E} \tilde{w}_{uv} (s(u) - s(v))^2,$$

where  $n$  is the number of nodes in the graph and the minimization ranges over all vectors  $s \in \mathbb{R}^n$ . We note that  $\tilde{L} = I - \tilde{W}$  is the *normalized Laplacian* of  $G$ . To minimize the above expression, we solve the system of linear equations  $(I + \alpha\tilde{L})\vec{s} = \vec{y}$ . It can be shown that for any connected graph  $G$  this system always has a unique solution, and that  $(I + \alpha\tilde{L})$  is always invertible. The matrix  $(I + \alpha\tilde{L})^{-1}$  can be interpreted as the amount of diffusion that flows in the network between any two node pairs, and is termed the *Regularized Laplacian* (9).

**GeneMANIA (10).** This method is a variation of RL that can take negative examples into consideration. Starting with a label vector  $\vec{y}$  where  $y(u)$  represents the prior evidence for protein  $u$  being a SARS-CoV-2 interactor, this algorithm computes a score  $s(u)$  between  $-1$  and  $1$  for each protein  $u$  in  $G$ . The value of  $y(u)$  is  $1$  or  $-1$  if  $u$  is a positive or negative example, respectively. Let the number of positive examples be  $n^+$  and the number of negative example be  $n^-$ . If  $u$  is an unknown example, then  $y(u) = \frac{n^+ - n^-}{n^+ + n^-}$ , the mean of the labels of the labeled nodes. Apart from this definition of  $\vec{y}$ , this method is identical to RL. Note that the original version of GeneMANIA implicitly chose  $\alpha = 1$ . We introduce the parameter  $\alpha$  to allow a tradeoff in the importance given to the input node labels  $\vec{y}$  vis-a-vis the similarity of adjacent labels in the output  $\vec{s}$ .

**SinkSource (11).** This method fixes the score  $s(u) = 1$  for every positive example and  $s(u) = -1$  for every negative example. It computes the score of every other node in  $G$  by minimizing the following function:

$$\sum_{(u,v) \in E} w_{uv} (s(u) - s(v))^2.$$

Let  $U$  denote the set of nodes in  $G$  that are unlabeled, i.e., are not positive or negative examples. For every node  $u \in U$ , we define the set of its neighbours as  $N(u)$  and use

$$f(u) = \sum_{\substack{v \in N(u) \\ v \text{ is positive}}} w_{uv} - \sum_{\substack{v \in N(u) \\ v \text{ is negative}}} w_{uv}$$

to denote the sum of the weighted scores of its neighbors that are positive or negative examples. Let  $\vec{f}$  denote the vector containing these values and let  $W$  denote the adjacency matrix of the subgraph of  $G$  induced by  $U$ . Defining  $D$  as a diagonal matrix of  $W$  as in the case of GeneMANIA but only for the nodes in  $U$ , we compute  $\vec{s}$  as the solution of the linear system of equations  $(I + WD^{-1})s = f$ .

When negative examples are not available, as is the case here, SinkSource adds an artificial negative example to the network, and connects each node to the artificial negative with an edge of weight  $\lambda$ , where  $\lambda > 0$  is a tunable parameter.

**Local.** We set  $s(v) = 1$  for every node in  $P$ . For every other node  $u$ , we initialise  $s(u) = 0$  and then compute  $s(u)$  as the weighted average of the scores of its neighbors.

**Support Vector Machine (SVM).** We set each node’s feature vector to be its adjacency vector in the normalized network  $\tilde{W}$ . We trained a linear kernel using the `LinearSVC` function in the `scikit-learn` Python package with default parameters.

**Logistic Regression.** We set each node’s feature vector as in the case of SVM. We used the `LogisticRegression` function in the `scikit-learn` Python package with default parameters.

## Network Proximity Among Interactors of SARS-CoV-2

We sought to develop a statistical justification for the RL method that we use to prioritize additional viral interactors using their network propagation distance to the set  $P$  of human proteins that interact with SARS-CoV-2 proteins. we used the following procedure to this end. We applied the Kolmogorov-Smirnov test to evaluate whether the distribution of scores of any given protein in  $P$  from the other proteins in  $P$  is the same as the null distribution of scores of randomly selected proteins in the network from  $P$ .

In more detail, for every protein  $p$  in  $P$ , we computed the contribution of the other proteins in  $P$  to the score of  $p$  after running RL, akin to leave-one-out cross validation. Specifically, we set  $y(p) = 0$  and executed RL to compute the score  $s(p)$  of  $p$ . We also computed the scores in this manner for 1,000 proteins selected uniformly at random from the other nodes in  $G$ . These distributions appear in Figure 2(a).

## Statistical Significance

To estimate the statistical significance of each node’s scores, we adopted a null hypothesis corresponding to the distribution of scores obtained from a randomly chosen positive set  $P'$  where  $|P'| = |P|$ . (As an alternative to randomizing the positive set, randomizing the network (e.g., via a degree-preserving edge swap process) would destroy the correlations between adjacent nodes (homophily) that are important contributions to pathway and neighborhood structure in the network.) We note that the degrees of the nodes in  $P$  may have a strong effect on the resulting distribution of scores. For example, if many nodes

in  $P$  have high degree, then scores may tend to be larger overall than if there are few nodes in  $P$  with high degree. Thus if the degree distribution of  $P$  does not approximately match that of  $P'$ , the resulting  $p$ -values will be biased.

Had we selected each random sample uniformly at random from all nodes in  $G$ , then the degree distribution of the chosen nodes would not be ensured to match that of the nodes in  $P$ . Therefore, we implemented a stratified sampling approach, as follows: Given a number of bins  $b$ , we partitioned the nodes in  $G$  into  $b$  sets as follows:

1. We sorted the nodes by weighted degree, i.e., we computed the degree sequence of  $G$ .
2. We executed  $k$ -means clustering on the degree sequence of  $G$  to compute  $b$  clusters (i.e., we set  $k = b$  in the  $k$ -means algorithm).

This approach emphasizes nearly-equal-degree groups. Then, to generate a random sample  $P'$  having  $|P|$  nodes, for every positive example  $v$  in  $P$ , we determined the subset whose range endpoints contained  $v$  and sampled a node from that subset uniformly at random. After evaluating various values of  $b$ , we selected the second approach and  $b = 10$  for use in our results (see Section S2.2).

For each  $P'$  we designated these nodes to be the set of positive examples and executed each of the prediction algorithms, ensuring that the negative samples we selected for each  $P'$  did not intersect with the original set of positive examples  $P$ . Repeating this procedure 1,000 times, we constructed a distribution of scores for each node in  $P$ . We then estimated the  $p$ -value of a node's score as the fraction of values in this distribution that were at least as large as the score. We did not correct these scores for multiple hypothesis testing.

## Identification of Biclusters with High Diffusion from SARS-CoV-2 Interactors to Top-Ranking Proteins

Let  $K$  denote the regularized Laplacian matrix  $(I + \alpha\tilde{L})^{-1}$ . We remind the reader that the RL algorithm ranks proteins based on diffusion scores that associate a node  $u$  in the network with a diffusion score  $s(u)$ , where  $s(u) = \sum_{v \in P} K_{uv}$ , where  $v$  ranges over the set  $P$  of all SARS-CoV-2 interactors. We sought to identify the experimentally determined interactors that contributed the most to the predicted diffusion scores. This analysis is important for tracing the provenance of computational predictions to their experimental sources (68, 69).

We considered the sub-matrix of  $K$  where every column represented one of the 332 SARS-CoV-2 interactors and every row corresponded to one of the top 332 proteins identified by the RL algorithm. In this submatrix, each value  $K_{uv}$  quantifies the diffusion between a given SARS-CoV-2 interactor in column  $v$  and a top-ranking protein in row  $u$ . We replaced each element in row  $u$  and column  $v$  by the value  $K_{uv}/s(u)$ , i.e., the fractional contribution of a SARS-CoV-2 interactor to the overall score for top-ranking protein  $u$ . We performed hierarchical clustering on both the rows and columns of this matrix using Pearson’s correlation as the similarity measure and average linkage. We manually identified eight biclusters and validated this selection with tests for enrichment of GO biological processes.

### Datasets

**SARS-CoV-2–Human PPIs.** We obtained 332 human proteins that interact with SARS-CoV-2 (7) and treated them as positive examples for our analysis. We added the ACE2 receptor to this set.



**Virus–Human PPIs for network proximity computations.** We obtained 104, 283, and 296 human proteins that interact with SARS-CoV, HIV-1, and HSV-1, respectively, from the VirHostNet database (70). We used these sets of proteins for the analysis in Figure 2(a).

**Functional and protein interaction networks.** We started with the human functional interaction network in the STRING database (version 11) (71), comprising of 18,886 nodes and 977,789 edges after applying a “medium” score cutoff of 400. We used the interaction reliabilities provided by STRING as edge weights; we divided each value in STRING by 1,000 to scale them between 0 and 1. An edge in this network may be derived from experimental data or computational analysis. Thus, an edge may represent either direct physical binding or indirect functional interaction. Of the 332 viral interactors, 328 were present in this network; REEP6 (Q96HR9), PPIL3 (Q9H2H8), RAB18 (Q9NP72), and FKBP7 (Q9Y680) were missing.

**Negative Examples.** To evaluate the precision of our predictions, we needed negative examples. Since datasets of human proteins that are certain not to interact with SARS-CoV-2 proteins are not available, we took the simple expedient of sampling them uniformly at random from the STRING network. We considered three different numbers of negative examples: as many as, five times, and ten times the number of positive examples. When we applied the methods to the full set of positive examples, we averaged the results over 100 random samples, with a positive:negative ratio of 1:5.

**Drug-protein interactions.** We downloaded interactions among drugs and proteins from the DrugBank database (version 5.1.6) (72). This dataset contained 16,503 drug-protein target pairs among 5,665 drugs and 2,891 target proteins. Limiting the targets

to those in the STRING network reduced the number of drugs and targets to 5,589 and 2,769, respectively.

## Evaluation

We evaluated the prediction algorithms using 100 runs of five-fold cross validation. In each run we ensured that all algorithms saw identical partitions of the examples into folds. We computed three measures of performance: (a) the area under the receiver operator characteristic (ROC) curve. (b) the area under the precision-recall curve, and (c) precision at a recall of 0.1 (*early precision*).

## Functional Enrichment

We used the `clusterProfiler` package in *R* (73) to compute Gene Ontology terms, KEGG pathways, and Reactome pathways enriched in our predictions or in the human interactors of SARS-CoV-2 proteins. This package uses Fisher’s exact test to estimate the enrichment of an individual term or pathway and the method of Benjamini and Hochberg to correct for testing multiple hypotheses. We applied this correction for each database (GO, KEGG, Reactome) separately. We used a threshold of 0.01 to decide if a GO term or KEGG/Reactome pathway was significantly enriched.

The enrichment analysis yields many highly similar statistically significant GO terms and KEGG and Reactome pathways; by “similar”, we mean that two different terms or pathways may annotate many proteins in common. This problem is well-known with several approaches that have been proposed to mitigate it either by grouping similar terms and selecting a small subset of dissimilar terms (74, 75) or by directly computing a set of non-redundant GO terms (76, 77). As far as we can tell, these methods have been developed to consider the enriched GO terms for one set of terms. When we apply them

independently to different sets of proteins (e.g., predictions from RL and predictions from SVM), they may select one term for one set of proteins but a similar but not identical term for another set, making the distinctions in enrichment hard to discern.

Therefore, taking inspiration from previously developed methods (cited above), we developed a simple heuristic based on the weighted set cover algorithm that simultaneously simplifies multiple sets of enriched terms or pathways. For every term that is enriched in at least one protein set, we defined its *composite odds ratio* to be the product of the odds ratios for that term across the protein sets. We iteratively selected the term with the largest composite odds ratio, deleted the proteins annotated to this term from the annotations of all other enriched term, adjusted the odds ratio for every term, and recomputed the composite odds ratios. We stopped when the maximum composite odds ratio of the remaining terms became less than one. We demonstrate how this algorithm was successful in reducing the redundancy among enriched GO terms in Section S2.4.

In the case of two terms automatically selected by this algorithm, there was a different, highly overlapping term that we felt would be more interpretable in the context of SARS-CoV-2 and COVID-19. Therefore, we manually replaced the selected terms by their alternative choices. Specifically, we replaced “ciliary basal body-plasma membrane docking” and “mRNA transport” with “cilium assembly” and “viral transcription”, respectively. To substantiate our choices, we computed the Jaccard index between the original and replaced term. Here we considered the proteins annotated by a term across all the protein sets (i.e., top-ranking proteins from RL and SVM and human interactors of SARS-CoV-2 proteins). The Jaccard index value was 0.95 between “ciliary basal body-plasma membrane docking” and “cilium assembly” with all 88 proteins annotated by the first term being annotated by the second, which had 93 annotations. The Jaccard index between “mRNA transport” and “viral transcription” was 0.46. The terms annotated 40

and 43 proteins, respectively, with 26 proteins in common. In addition, we removed many terms with a small number of annotations ( $\leq 7$ ) that were fairly similar to other terms in the list. We used the final list of terms for further analysis.

## Supplementary Files

These two supplementary files are available at the GitHub site mentioned above.

**Table S1:** Network propagation score of each SARS-CoV-2 interactor using the remaining SARS-CoV-2 interactors as seeds. Each row contains a SARS-CoV-2 interactor, its score, the average score of 1,000 randomly-selected proteins, and a  $p$ -value. This  $p$ -value is the fraction of randomly-selected proteins whose score was as least as large as that of the protein in that row.

**Table S2:** The prediction rank and  $p$ -value computed by RL and SVM for each human protein on the STRING network, the list of drugs that target the protein (when this information is available in DrugBank), and the closest SARS-CoV-2 interactor and SARS-CoV-2 protein. For the last piece of information, we computed the shortest weighted path, where we defined the weight of a path to be the sum of the absolute value of the base-10 logarithm of the weights of the edges in the path.

**Table S3:** Enrichment results for RL, SVM and the viral interactors on GO biological processes.

**Table S4:** Matrix of contributions to the network propagation score from each SARS-CoV-2 interactor to every top-ranking protein.

**Table S5:** Sorted contributions from SARS-CoV-2 interactors to specific proteins discussed in the paper. The table is divided into pairs of columns: the first column

specifies the rank and the name of a top-ranked protein followed by the SARS-CoV-2 interactors. The second column shows the fractional contribution from each interactor to the top-ranked protein.

# Supplementary Text

## S1 Supplementary Figures

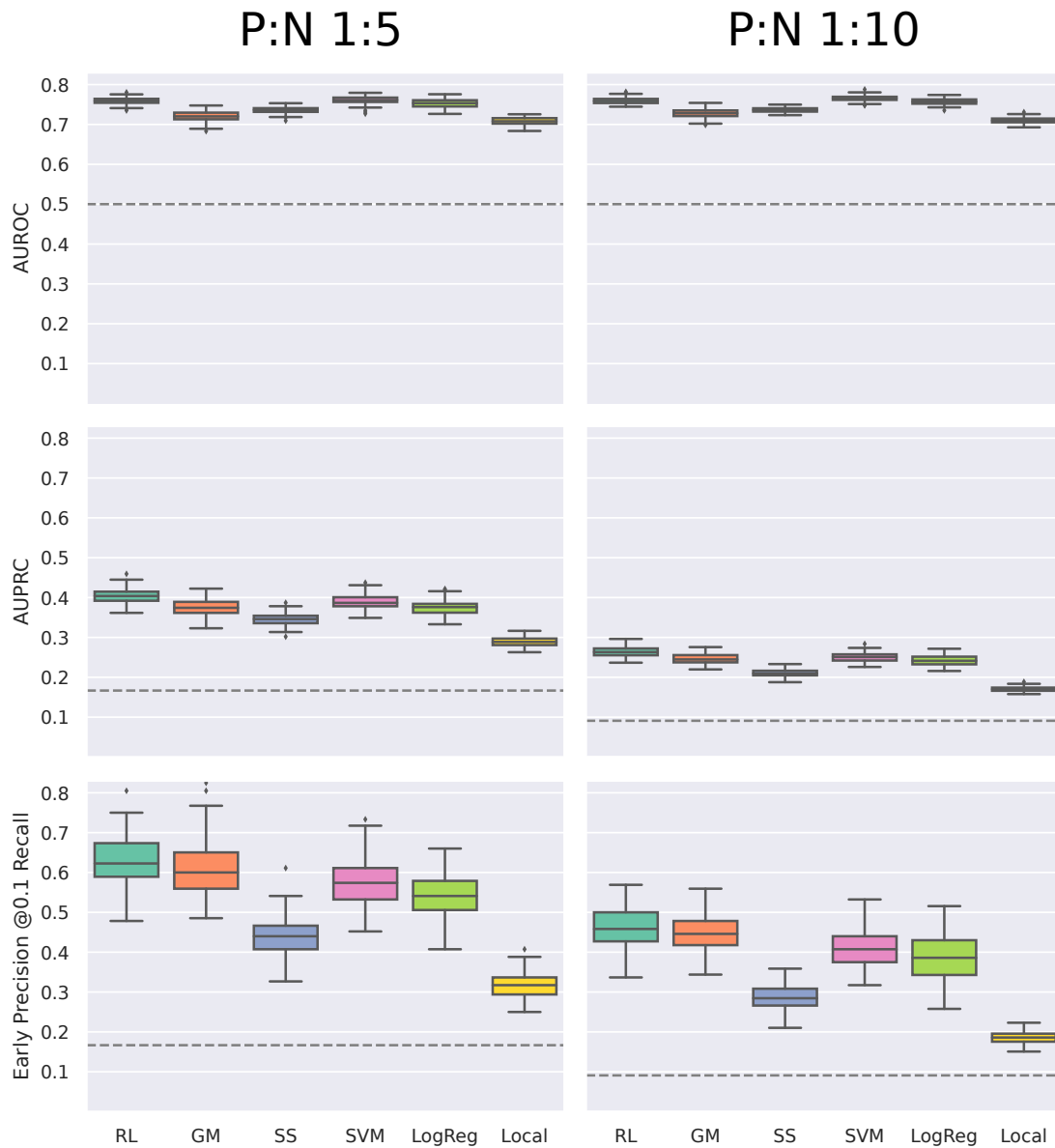


Figure S1: Cross validation results for positive:negative ratios of 1:5 and 1:10.

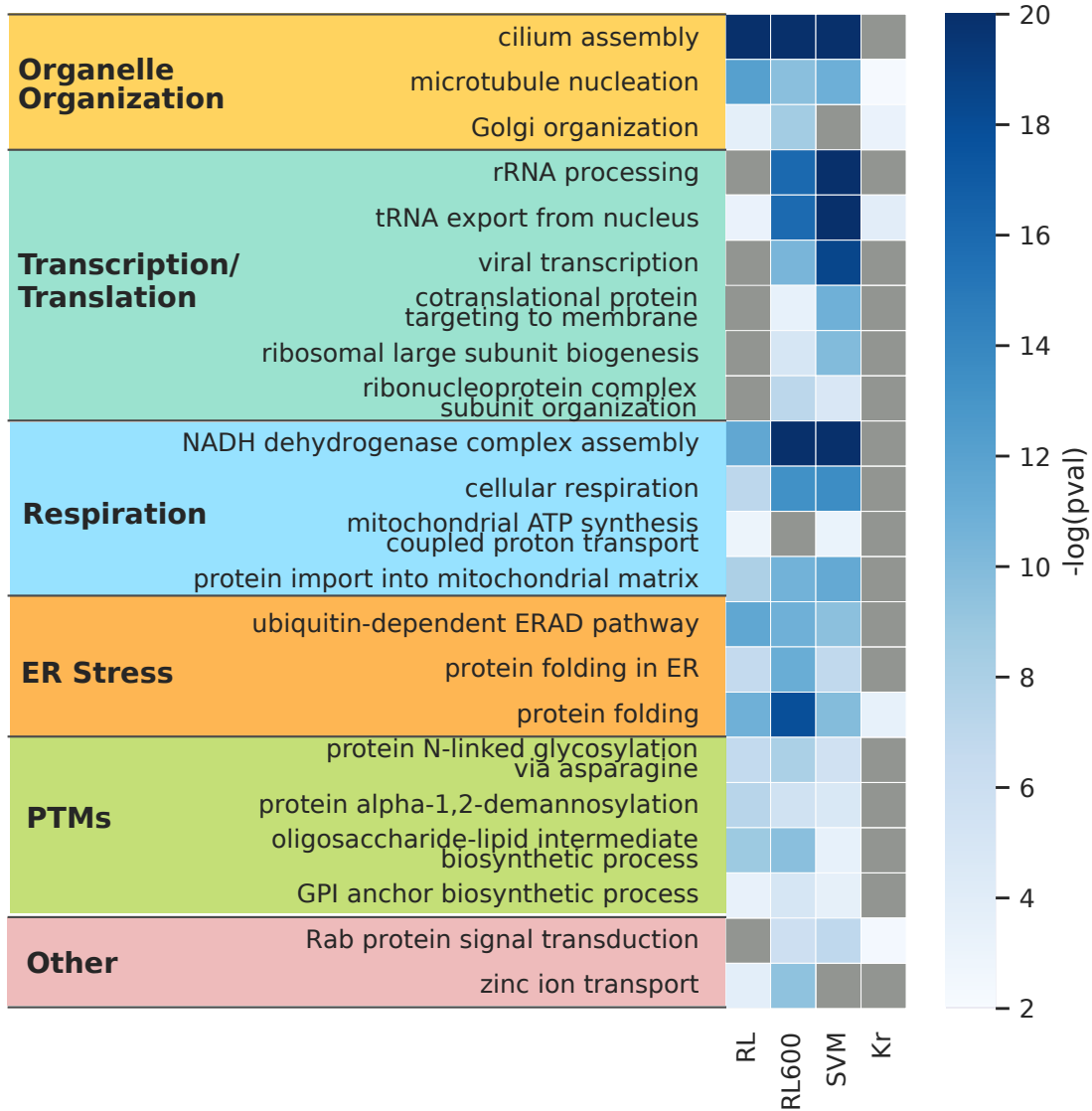


Figure S2: Comparison of GO terms enriched in two sets of top-ranking proteins in RL (top-332 and top-600), top-332 proteins from SVM, and SARS-CoV-2 interactors (abbreviated as 'Kr'). Terms related to transcription and translation are enriched in the top-600 RL proteins and the top-332 proteins in the SVM set.

## S2 Supplementary Text

### S2.1 Overlap Among Algorithms

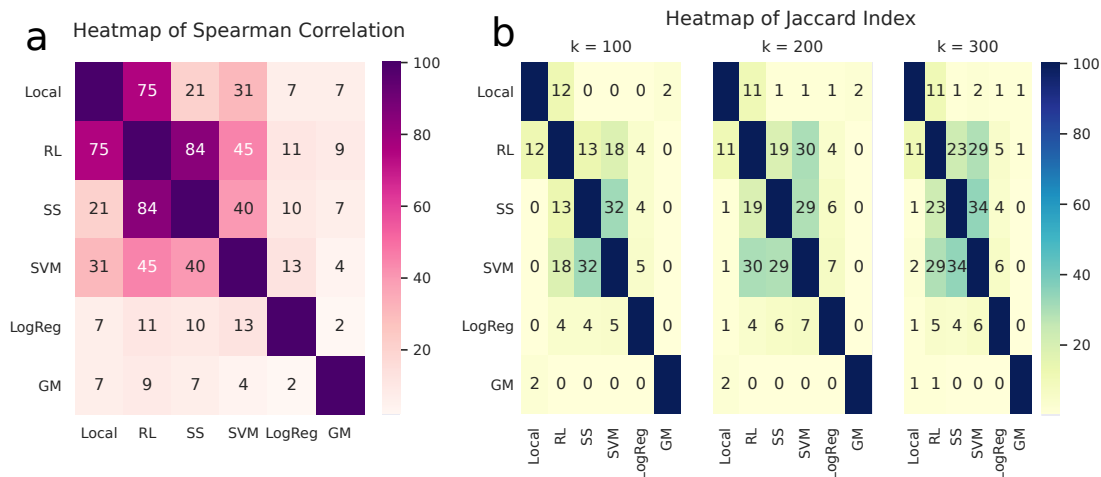


Figure S3: Similarity of predictions between every pair of methods. **(a)** Spearman correlations of node prediction scores. **(b)** Overlap of the top  $k$  predictions of each method, measured using the Jaccard index. The number in each cell is the value of the corresponding correlation or Jaccard index multiplied by 100.

To understand the diversity of the predictions across the prediction methods, between each pair of algorithms, we computed the Spearman's correlation of all scores. We also compared the top-ranking predictions between algorithms using the Jaccard index. RL had a high correlation with Local (0.75), which was surprising given the poor performance of Local. However, the Jaccard index for the two methods was around 0.1 (Figure S3(b)), suggesting that they shared very few top-ranking predictions and that the high correlation may be caused by lower-scoring proteins. SVM had moderate values of correlation and Jaccard's index with RL and SS (around 0.4 and 0.3, respectively). Logistic regression and GM both had low correlations and Jaccard's indices with all other methods ( $< 0.15$ ).



## S2.2 Parameter Selection

**Algorithms.** To tune the methods, we varied the parameter  $\alpha$  for RL, the weight of the edges  $\lambda$  connecting each node to the artificial sink for SS, and the parameter  $C$  controlling the inverse of the regularization strength for SVM and LR. For each setting of these parameters, we repeated 5-fold cross-validation with all three positive:negative ratios. We show the results for the ratios 1:1, 1:5, and 1:10 in Figure S4 and observed the results were fairly consistent across ratios. We focused on optimizing early precision values since we were interested in the analysis of top-ranking predictions.

For RL, GM, and SS, we found that in general, constraining the propagation locally around positive examples (i.e., small values of  $\alpha$ , large  $\lambda$ ) achieved higher early precision than more global propagation (i.e., large  $\alpha$ , small  $\lambda$ ). We chose the parameter values  $\alpha = 0.01$ ,  $\alpha = 0.1$ , and  $\lambda = 100$  for RL, GM, and SS, respectively.

For the supervised classifiers SVM and LR, we found that decreasing the regularization parameter (i.e., trying large values of  $C$ ) resulted in a slight increase in median early precision (about 0.05 for SVM, and 0.07 for LR) over the default  $C = 1$ . However, to avoid overfitting, we chose to use  $C = 1$  for both methods.

**Stratified sampling.** The number  $b$  of bins is a parameter. We tested  $b = 10, 20, 30$  and found that in each case, almost all of the top 332 nodes had a  $p$ -value  $< 0.05$  with the exception of the top 15 predictions for SVM, many of which had  $p$ -values slightly higher than the cutoff (Figure S5). In general, the  $p$ -values for SVM were slightly higher than those for RL. Since we did not observe much difference when varying  $b$ , we selected  $b = 10$ .

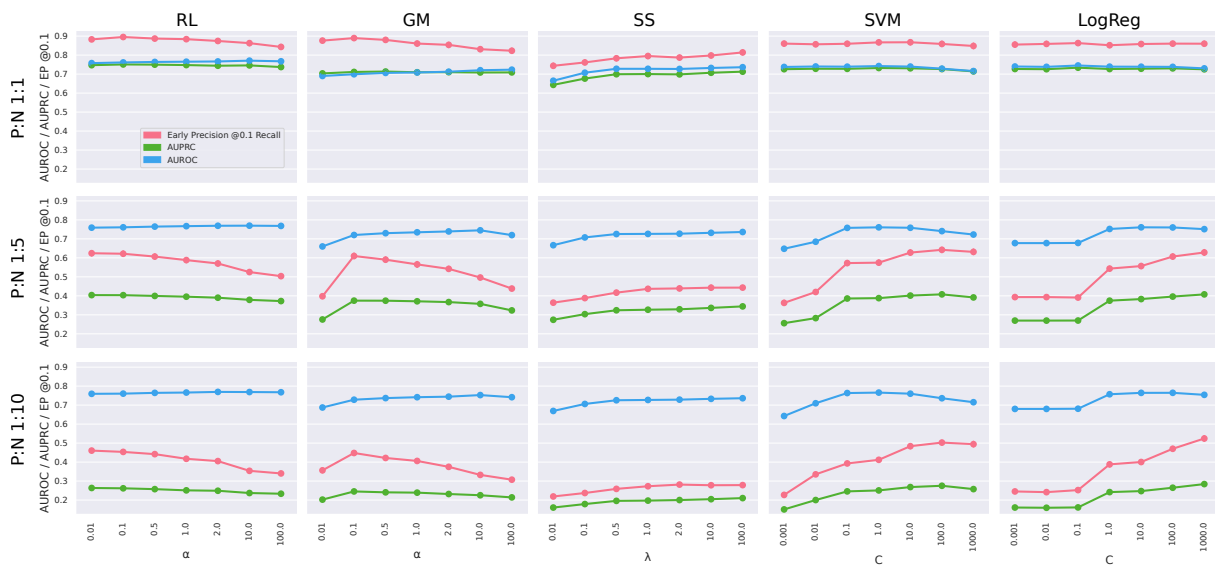


Figure S4: Parameter search results for each method, evaluated using AUROC, AUPRC, and early precision (at recall equal to 0.1) of 5-fold CV with a positive:negative ratio of 1:1, 1:5, and 1:10 on the STRING network. Each point shows the median value of 100 repetitions

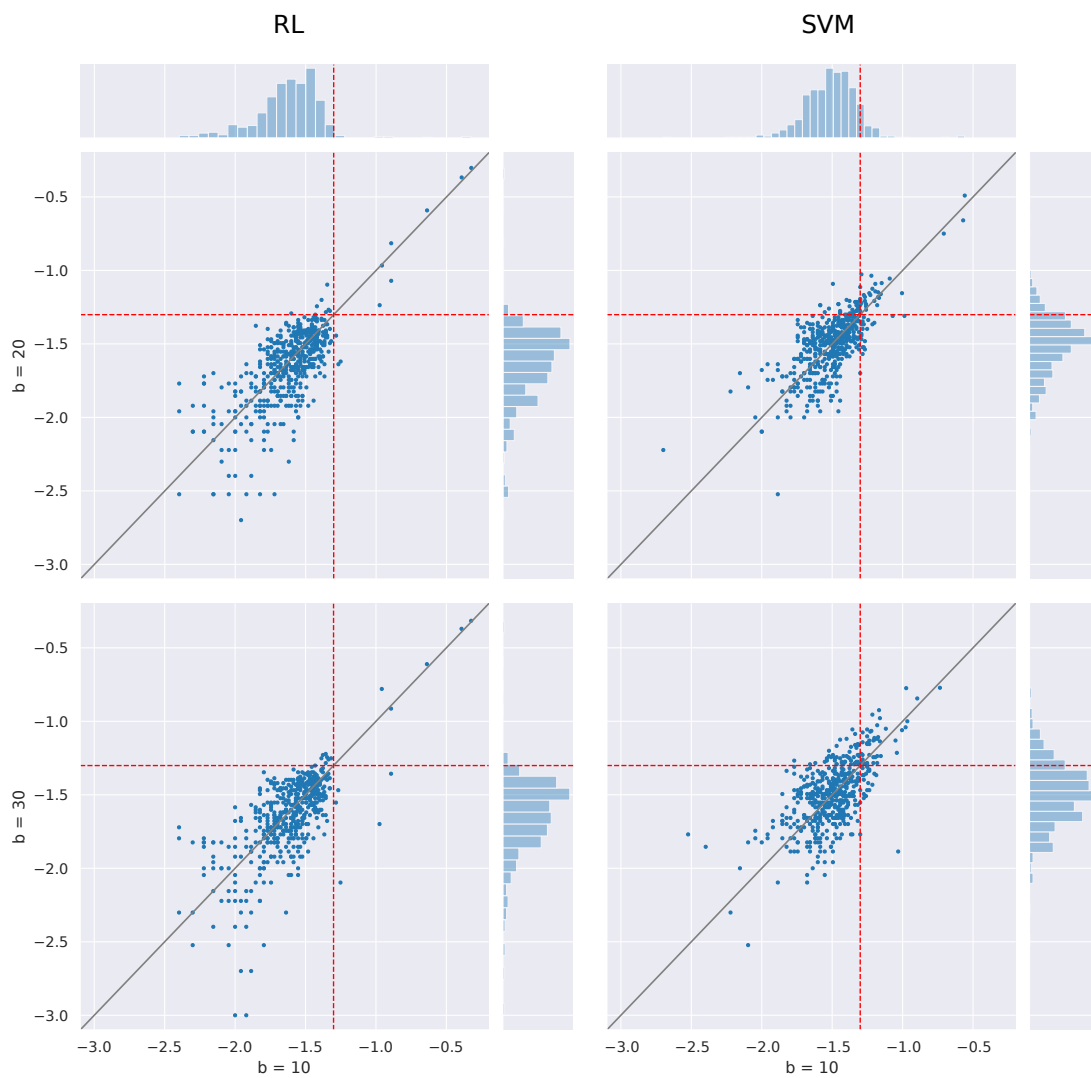


Figure S5: Base-10 logarithms of the  $p$ -values of node scores of the top 500 ranked proteins for RL and SVM for three values of  $b$ . The red dashed lines show the significance cutoff of 0.05, while the diagonal line shows  $x = y$ .

## S2.3 Analysis of Biclusters

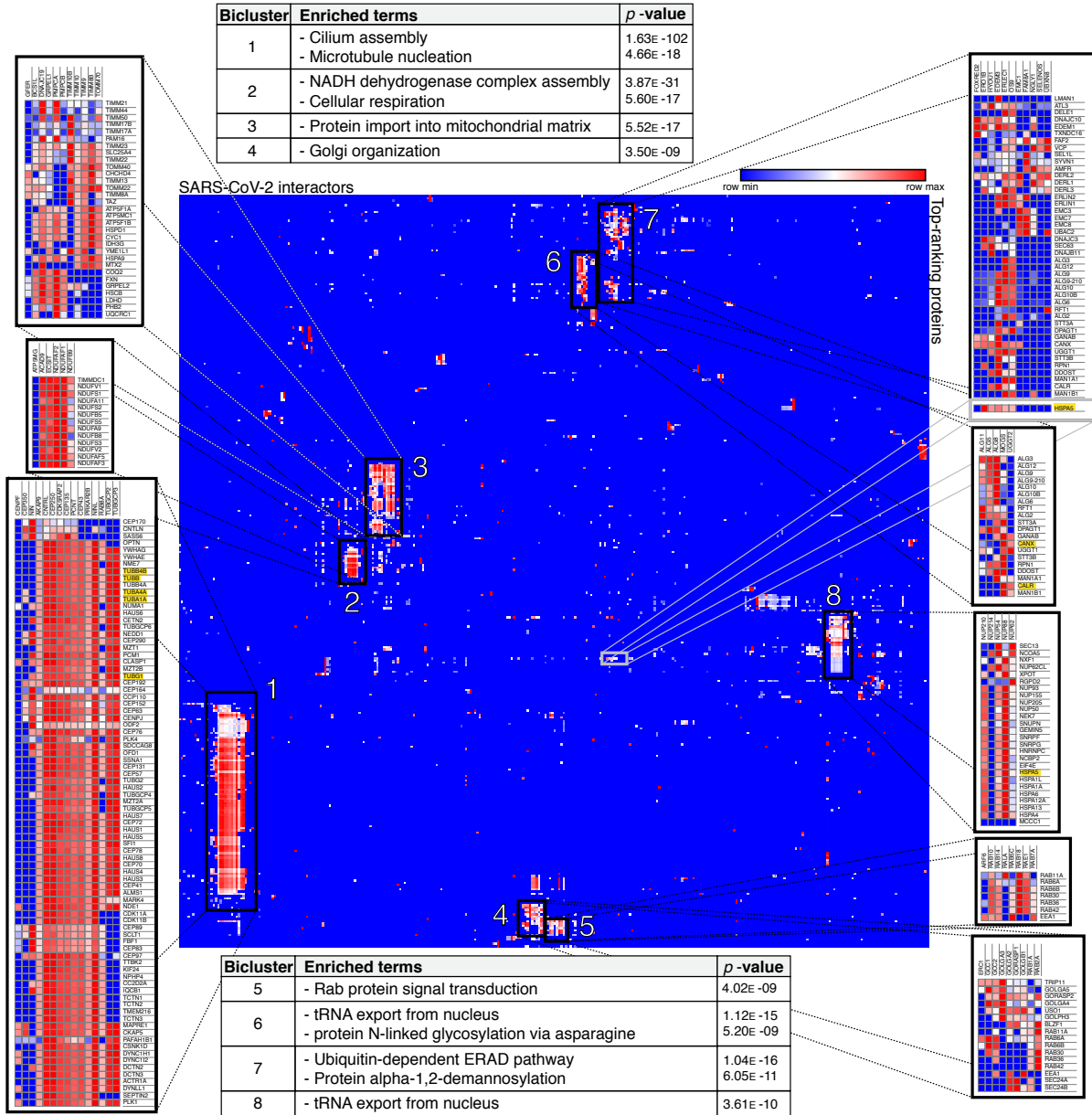


Figure S6: Figure 2(c) reproduced with individual biclusters visualized and GO terms denoted in additional tables. We generated this figure using MORPHEUS <https://software.broadinstitute.org/morpheus>.

Figure S6 shows a number of biclusters that represent groups of experimentally iden-

tified viral interactors (columns) that contribute substantially to diffusion scores of predicted high-ranking proteins in the network as described in “Methods”. High diffusion biclusters can naturally emerge either in long signaling chains or in dense subnetworks such as complexes. We hypothesize that these biclusters may arise from densely-connected subnetworks that are exploited by the virus to perturb host processes to promote viral infection.

One limitation of this approach arises when a protein matches well with multiple biclusters, in which case proteins with closely related functions may not appear in close proximity after we perform hierarchical clustering of the matrix of diffusion scores. In the case of HSPA5, we find that despite it having well-documented connections to CANX and CALR, it does not appear alongside them in bicluster 7 and instead is included in bicluster 8. This behavior could arise from the many additional functions HSPA5 has on the cell surface in addition to its role in blood coagulation along with CANX and CALR as discussed in “Results”. Despite not being included in bicluster 7, the HSPA5 score contributions that are attributed to the viral interactors of bicluster 7 are indicative of relatively high diffusion similar to that of the other rows in bicluster 7. Although HSPA5, CANX, and CALR are not directly adjacent in Figure S6, further observation reveals that there is notable overlap in the viral interactors that significantly contribute to their respective diffusion scores. We illustrate this point in Figure S6 in the form of the additional row shown in the breakout for bicluster 7.

To verify that these biclusters are associated with functionally coherent subnetworks implicated in viral pathogenesis, we performed gene set enrichment on the genes in each identified bicluster (“Methods”). The results indicated that each bicluster can be effectively mapped to one or more enriched GO terms (Figure S6). This type of biclustering analysis also serves as a nuanced visual representation of how the RL diffusion algorithm

works and its contrast to other network properties such as shortest path lengths or centrality. The biclustering analysis reveals that in several cases the majority of diffusion for a given protein can be attributed to a single or a small group of seed nodes. For example, 100% of the diffusion score for RTF1 originates from four viral interactors: (1) ALG11: 37%; (2) ALG5: 26%; (3) ALG8: 21%; and (4) MOGS: 16%. This analysis also reveals that in several cases the highest contributor to a given protein’s score may not be the nearest seed node by path length. For instance, 8.91% of the score for protein TUBB4B can be attributed to TBCA despite UPF1 being the nearest viral interactor and only contributing  $7.48 \times 10^{-4}\%$ . To further explore the trends among top contributors, we sorted the components contributing to each protein score in descending order (Table S5) and plotted the results. Figure S7 shows this analysis for a group of four tubulin proteins that all appear in bicluster 1 and in the network related to cilium assembly (Figure 4(g)).

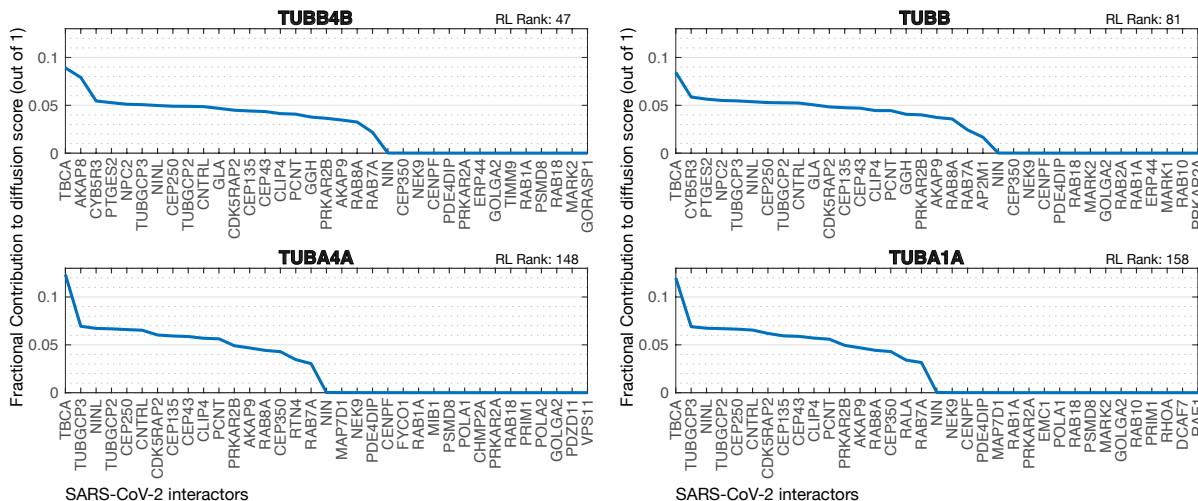


Figure S7: Top 35 SARS-CoV-2 interactors contributing to the diffusion score of 4 notable proteins: TUBB4B, TUBB, TUBA4A, and TUBA1A with RL ranks of 47, 81, 148, and 158 respectively. Each observation represents the fractional contribution by the viral interactor on the x-axis to the total predicted score of the protein in each title.

The four proteins depicted share the same highest contributor, TBCA, and all demon-

strate a sharp decrease from the first to second contributing node. TBCA is the gene that codes for the tubulin-specific chaperone A protein which is involved in beta-tubulin binding to facilitate tubulin complex assembly (78, 79). Within the first 25 highest ranked contributors in each plot, we observe another sharp decrease to near zero relative contribution from the remaining seed nodes. We view this steep drop-off as an indication that seed nodes could be clustered such that diffusion from one group of seed nodes to a given protein far outweighs that of another set of seed nodes. These lower contribution seed nodes could be configured in either a distant cluster or such that many individual nodes are distributed widely. Further analysis using diffusion mapping techniques could assist in distinguishing between these two proposed configurations.

## **S2.4 Evaluation of the Method for Simplifying Functional Enrichment Results**

The main goal of simplifying the result from enrichment analysis was to filter out the redundant terms, i.e., terms that have highly overlapping annotated proteins. To evaluate how our simplification method (“Methods”) performs in terms of reducing overlap, we analyzed the enrichment results from each protein set (i.e., top-ranking proteins from RL, top-ranking proteins from SVM, and human interactors of SARS-CoV-2 proteins) in two ways. We performed each analysis for the set of terms before simplification and for the set of terms after simplification. We show the results RL as an exemplar.

First, for each enriched term  $t$ , we computed the largest Jaccard index between  $t$  and every other enriched terms. We plotted the distributions of these quantities. We expected the Jaccard indices to be large in the ‘before’ set and small in the ‘after’ set of terms. Indeed, we observed that our results exhibited these trends for each protein set. We present the result from top-ranking RL proteins in Figure S8(a). For the top-ranking RL proteins,

there were 181 and 23 enriched terms before and after simplification respectively. The maximum Jaccard index was  $\geq 0.5$  for 78% of the terms before simplification. This percentage reduced to 4.5% of the terms after simplification. Moreover, after simplification, for 86% of the terms, the maximum Jaccard index was  $< 0.3$ , which indicates a considerable improvement in overlap. An exception was the pair of terms “NADH dehydrogenase complex assembly” and “cellular respiration”, which were both present after simplification yet had a Jaccard index of 0.51 and an overlap of 29 annotated proteins. Of these two terms, our method selected “NADH dehydrogenase complex assembly” first since it had the higher composite odds ratio. The second term still had 14 annotated proteins that were not covered by any other previously selected term. Its resulting composite odds ratio of  $1.85 > 1$  enabled its selection by the algorithm in the later iterations. Of these two highly overlapping terms, we only retained “NADH dehydrogenase complex assembly” (which had the higher composite odds ratio) for further analysis due to the low composite odds ratio of “cellular respiration”.

Second, we considered the set of proteins annotated by at least one enriched term before simplification. There were 260 such proteins from the top-ranking RL results and GO biological processes. For each such protein, we computed the number of enriched terms that contained the protein and plotted the distributions of these counts. With the same set of proteins, we repeated the analysis after simplification. We observed that before simplification, 72% of proteins were annotated by more than 5 enriched terms. This percentage reduced to 18% of proteins after simplification. Moreover, the percentage of proteins annotated by five or fewer terms increased dramatically after simplification. We observed that before simplification 7%, 8%, 5%, 2%, and 4% of proteins were covered by 1, 2, 3, 4, and 5 terms, respectively. In contrast, these values were 54%, 20%, 5%, 1%, and 3% after simplification. In addition, 18% of proteins were not covered by any terms after



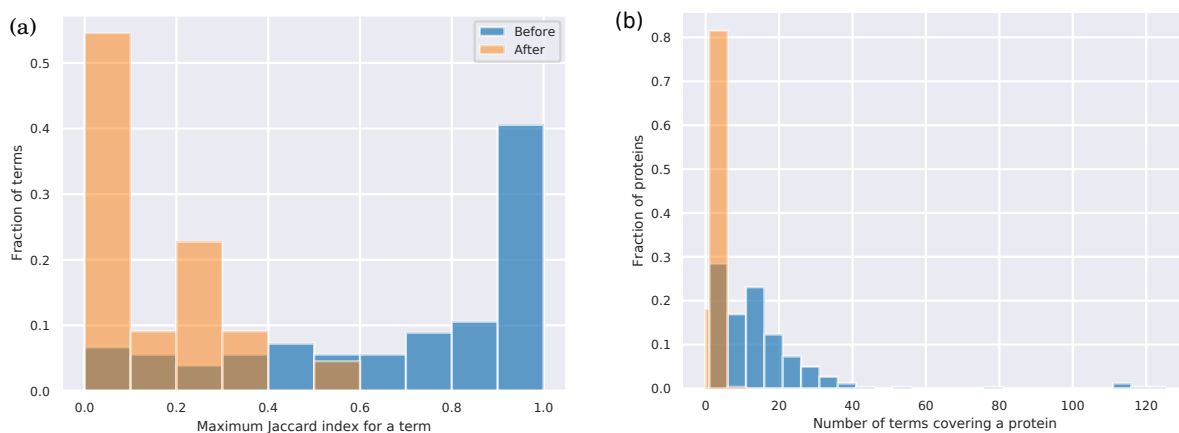


Figure S8: Results of simplification of enriched terms. (a) Distribution of the maximum Jaccard index values for the enriched GO biological processes in the top-ranking RL proteins, both before and after simplification. The  $x$ -axis corresponds to the maximum Jaccard index and the  $y$ -axis to the fraction of terms that have a maximum Jaccard index in a particular range. (b) Distribution of the number of enriched enriched GO biological processes containing each protein, considering the top-ranking RL proteins, both before and after simplification. The thin bar at 0 indicates the fraction of proteins in the protein universe left uncovered after simplification.

simplification. However, no GO terms were enriched in this set of proteins at the 0.01 level. Overall, these results confirmed the efficacy of our approach in computing non-redundant enriched terms.

## References

68. B. P. Anton, *et al.*, *PLoS Biol.* **11**, e1001638 (2013).
69. S. Kasif, R. J. Roberts, *To appear* (2020).
70. V. Navratil, *et al.*, *Nucleic Acids Research* **37**, D661 (2009).
71. D. Szklarczyk, *et al.*, *Nucleic Acids Research* **45**, D362 (2016).
72. D. S. Wishart, *et al.*, *Nucleic Acids Research* (2018).
73. G. Yu, L. G. Wang, Y. Han, Q. Y. He, *OMICS A Journal of Integrative Biology* **16**, 284 (2012).
74. Supek, F. and Bošnjak, M. and Škunca, N. and Tomislav, Š. , *PLoS One* **6**, e21800 (2011).
75. D. Merico, R. Isserlin, O. Stueker, A. Emili, G. D. Bader, *PLoS ONE* **5**, e13984 (2010).
76. Y. Lu, R. Rosenfeld, I. Simon, G. J. Nau, Z. Bar-Joseph, *Nucl. Acids Res.* **36**, e109+ (2008).
77. S. Bauer, J. Gagneur, P. N. Robinson, *Nucleic Acids Research* **38**, 3523 (2010).
78. The UniProt Consortium, *Nucleic Acids Research* **47**, D506 (2018).

79. P. Gaudet, M. S. Livstone, S. E. Lewis, P. D. Thomas, *Briefings in bioinformatics* **12**, 449 (2011). Edition: 2011/08/27 Publisher: Oxford University Press.