RESEARCH

Interpretable network propagation with application to expanding the repertoire of human proteins that interact with SARS-CoV-2

Jeffrey N. Law [1,2], Kyle Akers [1], Nure Tasnina[3], Catherine M. Della Santina [4], Shay Deutsch [5], Meghana Kshirsagar [6], Judith Klein-Seetharaman [7], Mark Crovella [8], Padmavathy Rajagopalan [9], Simon Kasif [4] and T. M. Murali [3,*]

[1]Interdisciplinary Ph.D. Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA 24061, USA; [2]Present address: Biosciences Center, National Renewable Energy Laboratory, 15013 Denver West Parkway Golden, CO 80401, USA; [3]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA; [4]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA; [5]Department of Mathematics, University of California, Los Angeles, CA 90095, USA; [6]AI for Good Lab, Microsoft, Redmond, WA 98052, USA; [7]Department of Chemistry, Colorado School of Mines, 1500 Illinois St, Golden, CO 80401, USA; [8]Department of Computer Science, Boston University, Boston, MA 02215, USA and [9]Department of Chemical Engineering, Virginia Tech, Blacksburg, VA 24061, USA

*Correspondence address. T. M. Murali, Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA.
E-mail: murali@cs.vt.edu http://orcid.org/0000-0003-3688-4672

Abstract

Background: Network propagation has been widely used for nearly 20 years to predict gene functions and phenotypes. Despite the popularity of this approach, little attention has been paid to the question of provenance tracing in this context, e.g., determining how much any experimental observation in the input contributes to the score of every prediction. **Results:** We design a network propagation framework with 2 novel components and apply it to predict human proteins that directly or indirectly interact with SARS-CoV-2 proteins. First, we trace the provenance of each prediction to its experimentally validated sources, which in our case are human proteins experimentally determined to interact with viral proteins. Second, we design a technique that helps to reduce the manual adjustment of parameters by users. We find that for every top-ranking prediction, the highest contribution to its score arises from a direct neighbor in a human protein-protein interaction network. We further analyze these results to develop functional insights on SARS-CoV-2 that expand on known biology such as the connection between endoplasmic reticulum stress, HSPA5, and anti-clotting agents. **Conclusions:** We examine how our provenance-tracing method can be generalized to a broad class of network-based algorithms. We provide a useful resource for the SARS-CoV-2 community that implicates many previously undocumented proteins with putative functional relationships to viral infection. This resource includes potential drugs that can be opportunistically repositioned to target these proteins. We also discuss how our overall framework can be extended to other, newly emerging viruses.

Background

Network propagation algorithms have been widely used for nearly 20 years for function and phenotype prediction in systems biology [1–7]. More recently, applications of these techniques have included determination of genes associated with cancers and complex diseases [8] and denoising single-cell gene expression data [9]. Nowadays, network-based algorithms facilitate large-scale and automated data analysis of such complexity that it can be difficult for humans to understand the rationale that underlies a prediction, leading to decreased transparency and interpretability.

In this work, we consider the fundamental problem of tracing the provenance of a prediction back to the experimental sources [10]. Given a protein interaction network and a set of "sources," e.g., the human proteins that physically interact with SARS-CoV-2 [11], suppose we apply a network-based algorithm to score and prioritize additional proteins that may directly or indirectly interact with the virus. Can we determine which source proteins make the highest contribution to the score computed for each prediction? Surprisingly, this question has been insufficiently studied in the field of network biology [10]. This aspect takes particular importance in the context of COVID-19 or other clinically or scientifically critical applications, where it may be important to understand the rationale behind the computational prediction of a new drug target before committing to expensive experimental validation.

We present a simple and direct method to solve this problem for a large class of network propagation algorithms. Specifically, for each protein u in the network, we compute the precise contribution of each source to the score of u. This calculation enables us to sort the sources by their relative contributions to u and to quantify the relative roles of sources at different distances from u.

To evaluate the effectiveness of this strategy, we apply it to prioritize host proteins that may "functionally" (directly or indirectly) interact with SARS-CoV-2 proteins and host cellular processes that may be hijacked by the virus (Fig. 1). To this end, we take advantage of a recently published dataset of human proteins that physically interact with SARS-CoV-2 [11]. Although these SARS-CoV-2 interactors are entry points to host cellular processes that may be hijacked by viral infection, the proteomics pipeline used to discover them [11] may not capture *in vivo* conditions and tissue-specific interactions, leading to false-negative results. Therefore, we apply network propagation algorithms to these known human protein interactors of SARS-CoV-2 proteins (sources) and a whole-genome human protein interaction network from the STRING database [12]. We identify statistically enriched host biological processes and pathways that include highly ranking proteins computed by our methods. We illustrate how our provenance analysis can simplify visualizations of these processes and assist in understanding how they may be affected by SARS-CoV-2.

Data description

Here, we detail the different viral-human and human protein and functional interaction networks that we used in our study.

SARS-CoV-2–human protein-protein interactions

We obtained 332 human proteins that interact with SARS-CoV-2 [11] and treated them as positive examples for our analysis. We added the ACE2 receptor to this set.

Functional and protein interaction networks

We used the human functional interaction network in the STRING database (version 11) [12], comprising 18,886 nodes and 977,789 edges after applying a "medium" score cut-off of 400 and mapping to UniProt IDs. 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.

We also computed results for protein-protein interaction (PPI) networks from 2 other sources: the BioGRID database [13], and the high-quality "HI-union" network published by Luck et al. [14]. For BioGRID, we considered 2 versions: (i) all PPIs (including protein complex membership) and (ii) only direct PPIs from yeast 2-hybrid (Y2H) screens. For each of these networks, we did not use edge weights and restricted the nodes and edges to those in the largest connected component. See Table 1 for statistics of the network size and density.

Drug-protein interactions

We downloaded interactions among drugs and proteins from the DrugBank database (version 5.1.6) [15]. 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.

SARS-CoV-2–human A549 AP-MS interactome

We obtained 882 human proteins determined to interact with SARS-CoV-2 proteins by affinity purification followed by mass spectrometry analysis (AP-MS) [16]. This dataset was generated in A549 lung carcinoma cells transduced with lentivirus vectors expressing HA-tagged SARS-CoV-2 proteins. The authors used affinity purification with anti-HA antibodies to isolate stable complexes of human proteins bound to SARS-CoV-2 proteins. Subsequently, they identified and quantified the purified proteins by mass spectrometry.

SARS-CoV-2–human HEK293 AP-MS interactome

We obtained a set of 225 human proteins determined to interact with SARS-CoV-2 by AP-MS [17]. This dataset was generated by analyzing HEK293 embryonic kidney cells transfected with plasmid vectors expressing FLAG-tagged SARS-CoV-2 proteins. Affinity purification with anti-FLAG antibodies was used to isolate stable complexes of human proteins bound to SARS-CoV-2 proteins, and the purified proteins were identified and quantified by mass spectrometry.
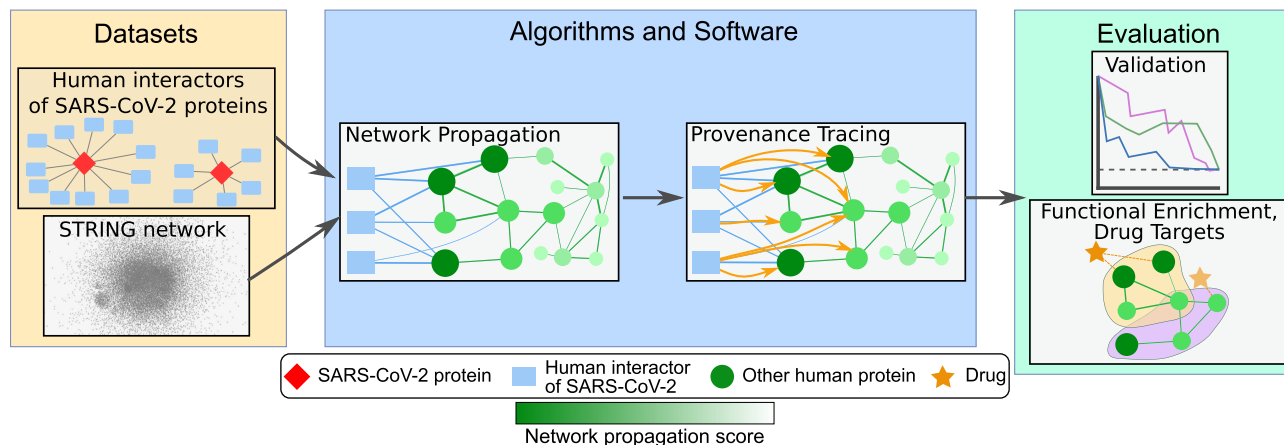
Figure 1: Overview of methodology. Algorithms and software for network propagation and provenance analysis take as input experimentally determined host-pathogen protein interactions and a human protein interaction network. Evaluation includes cross-validation, functional enrichment, and literature-based examination of promising protein targets and drugs.

Table 1: Network statistics.

Network	Nodes	Edges	Edge weights	Density	No. SARS-CoV-2 interactors (/333)	No. neighbors of sources
STRING (400)	18,886	977,789	Y	5.5×10^{-3}	329	12,480
BioGRID	16,595	488,787	N	3.6×10^{-3}	333	9,178
BioGRID-Y2H	12,582	87,801	N	1.1×10^{-3}	271	2,891
HI-union	9,053	64,193	N	1.6×10^{-3}	168	2,031

For STRING, the weight cut-off applied is in parentheses. The column titled "No. SARS-CoV-2 interactors (/333)" shows the number of sources that were in the network. The "No. neighbors of sources" column shows the number of neighbors of the human proteins that interact with SARS-CoV-2 proteins (i.e., sources) in the given network.

SARS-CoV-2–human BioID interactome

We obtained a set of 2,241 human proteins determined to interact transiently or weakly with SARS-CoV-2 proteins by using proximity-dependent biotinylation (BioID) [18]. This dataset was generated by analyzing A549 lung carcinoma cells transduced with lentivirus vectors expressing SARS-CoV-2 proteins fused with a bacterial biotin ligase. The addition of biotin resulted in the biotinylation of host proteins in the proximity of SARS-CoV-2 proteins. Biotinylated proteins were purified and then identified and quantified by mass spectrometry. Compared to interactomes identified by AP-MS, BioID is more capable of identifying weaker interactions in poorly soluble intracellular locations such as membranes and organelles.

Differential protein abundance in SARS-CoV-2–infected iAT2 cells

We obtained a set of 5,665 human proteins determined to have differential abundance in response to SARS-CoV-2 infection [19]. This dataset was generated by infecting induced pluripotent stem cell–derived alveolar epithelial Type 2 cells (iAT2) with SARS-CoV-2 and measuring protein abundance by quantitative mass spectrometry at 1, 3, 6, and 24 hours after infection. The authors compared protein abundance in infected iAT2 cells with that of the uninfected iAT2 controls to obtain differentially expressed proteins. In our analysis, we used the set of proteins with differential expression (false discovery rate [FDR] $P < 0.05$) at any of the times 1, 3, 6, and 24 hours after infection.

Differential gene expression in upper airway samples in SARS-CoV-2–infected patients

We obtained 3 sets of human proteins determined to have differential gene expression in cells infected with respiratory viruses [20]. To generate this dataset, the authors used metagenomic RNA-seq to identify and quantify both human and viral RNA expression in upper airway samples collected from patients with acute respiratory illness. They compared the gene expression values between samples that contained SARS-CoV-2 and uninfected samples to obtain differentially expressed genes. They also identified additional viral infections including SARS-CoV, human rhinovirus, influenza, human metapneumovirus, respiratory syncytial virus, and parainfluenza virus in patient samples. Comparing SARS-CoV-2 infections with other viral infections and other viral infections with uninfected samples yielded 2 additional sets of differentially expressed genes. In our analysis, we used the genes with differential expression (FDR $P < 0.05$) in these 3 sets, obtaining (i) 1,383 genes from SARS-CoV-2–infected cells compared with uninfected samples, (ii) 7,338 genes from SARS-CoV-2–infected cells compared with other viral infections, and (iii) 5,779 genes from other viral infections compared with uninfected samples.

From each of these interactome and differential expression datasets, we removed human proteins used as positive examples in our analysis and the proteins that were not present in the STRING network. This step resulted in 2,080, 807, and 212 proteins, respectively, from the interactome datasets and 5,447, 1,293, 6,940, and 5,472 proteins, respectively, from the differential expression datasets. We used the Fisher exact test to esti-

mate the statistical significance of the overlap between the remaining proteins and our top-ranking proteins.

Analyses

Various network propagation methods have been successfully used in diverse applications in systems biology [21]. In particular, we model network propagation using the regularized Laplacian (RL) [22]. As we describe in the Methods section, RL has the benefit of 2 mutually reinforcing interpretations. On one hand, it can be understood as an optimal labeling of network nodes, when some node labels are known a priori. On the other hand, it can be seen as the result of diffusion, i.e., a continuous-time random walk, on the network. Under this second interpretation, we derived a novel mathematical formula for the expected length of the path traversed in the network by the random walker, which we then used to characterize our top-ranking proteins.

Prioritization of potential SARS-CoV-2 interactors

Our underlying hypothesis was that network propagation via methods such as the RL yields a reasonable mechanism for predicting SARS-CoV-2 interactors. Therefore, we applied RL to the set of positive examples to rank the remaining proteins in the STRING network. We also ranked these proteins using multiple other network propagation methods and off-the-shelf classifiers [23–26]. We used a stratified sampling approach to estimate the statistical significance of the resulting node scores (see "Statistical Significance of Node Scores" in the Supplementary Methods). The sampling accounted for the possibility that if many sources have high degree, then scores may tend to be large overall in the network. Henceforth, for every method, we only considered proteins in the network that had $P < 0.05$.

To decide which methods to select for subsequent analyses, we compared them using 5-fold cross validation ("Comparison of Cross-Validation Results" in the Supplementary Results and Supplementary Fig. S1). RL, random walk with restarts (RWR) [23], and deepNF [26] had the highest values of area under the precision-recall curve (AUPRC) followed by SVM and logistic regression. RL achieved marginally worse values of AUPRC than RWR and deepNF. We selected 1 network propagation method (RL) and 1 supervised classifier (SVM) for the following reasons. We preferred RL over deepNF because the provenance-tracing method we developed for RL enabled its results to be more easily interpreted than those for deepNF. Because RL and RWR produced highly similar predictions with a very high Spearman correlation for the ranking of all proteins ("Overlap among algorithms" in the Supplementary Results and Supplementary Fig. S2), we selected RL as representative of the 2 methods. We chose SVM among the 2 off-the-shelf classifiers because it also had very good performance in cross-validation. We considered the top 332 predictions of RL and SVM that were statistically significant at $P < 0.05$ ("List of RL and SVM predictions, P-values, and top-2 contributors" [27]), which we refer to as "top-ranking proteins" below.

Three recent publications or preprints have independently discovered physical interactions between SARS-CoV-2 and human proteins [16–18]. These datasets differed in the type of host cell in which the viral proteins were expressed and the experimental methods used to determine whether 2 proteins interacted ("Datasets"). The top-ranking proteins for both RL and SVM had significant overlaps with each of the 2 new datasets, while the results for Local were not statistically significant ($P > 0.01$) (Fig. 2A). We observed an especially striking overlap with the "proximity interactome" [18]. Approximately one-third of the 332 and 1,000 top-ranking proteins computed by RL were present in this dataset of 2,080 interactions ($P = 8.2 \times 10^{-24}$ and 3.9×10^{-74}, respectively).

The corresponding publication used BioID with the fast-acting miniTurbo enzyme [18], a technique that is useful for discovering viral-host protein interactions that take place at intracellular membranes and poorly soluble organelles, which are difficult to profile using classical biochemical purification approaches used in the other publications [11,16,17]. Thus, our top-ranking proteins may be members of biological processes that occur in such locations in the cell. These 3 independent datasets provide strong support for our predictions. Our top-ranking proteins that do not overlap with these resources may interact with viral proteins indirectly and thus would not be captured by assays that test for direct protein-protein interactions.

We additionally tested the overlap between our top predictions and independent experimental datasets identifying differential expression of proteins in response to SARS-CoV-2 infection [19]. As in the previous analysis, we observed that the results for Local were not statistically significant ($P > 0.01$), while both RL and SVM had significant overlaps with differential protein abundance in SARS-CoV-2–infected cells compared with uninfected cells [19] (Fig. 2A). Approximately half of the 332 and 1,000 top-ranking proteins computed by RL were present in this dataset of 5,447 differentially expressed proteins ($P = 3.9 \times 10^{-14}$ and 9×10^{-62}, respectively). This high overlap may indicate that these proteins are involved with changes in host protein expression occurring in SARS-CoV-2–infected cells, via either direct or indirect virus-host protein interactions.

In contrast, when we analyzed gene expression measurements in response to SARS-CoV-2 infection [20], we did not observe a significant overlap between our top-ranking proteins and differentially expressed genes (Supplementary Fig. S8). This result may be attributed to a difference in cell types used for measuring gene expression data, including cells not directly infected by the virus. Moreover, the lack of edges connecting transcription factors to target genes in the PPI network we used may limit the size of the overlap between interactors predicted by RL and SVM with differentially expressed genes.

We tested for enrichment of Gene Ontology (GO) biological processes (Benjamini-Hochberg–corrected $P \leq 0.01$) among the top-ranking proteins from RL and from SVM, as well as in the interactors of SARS-CoV-2 ("Functional Enrichment" in the Supplementary Methods). Our top-ranking proteins were enriched in 5 broad categories of GO biological processes: organelle organization, transcription and translation, respiration, endoplasmic reticulum (ER) stress, and post-translational modifications (Fig. 2B, Supplementary Fig. S5, and "Enrichment results for RL, SVM, and viral interactors" [27]). We examine the relevance of these processes to the viral life cycle in more detail in "Discussion" and in "Enriched Biological Processes" in the Supplementary Results.

Tracing the provenance of top-ranking proteins

We can interpret the RL in terms of a continuous-time random walk over the network, which is governed by the internal parameter α. We are interested in the node reached by the walker after a random time that depends on α. The expected number of transitions made by the walker increases with the parameter α ("Analytical Perspective on the RL and Expected Path Length" in the Supplementary Methods). Hence for larger values of α, the "influence" of the sources is diffused more broadly across the
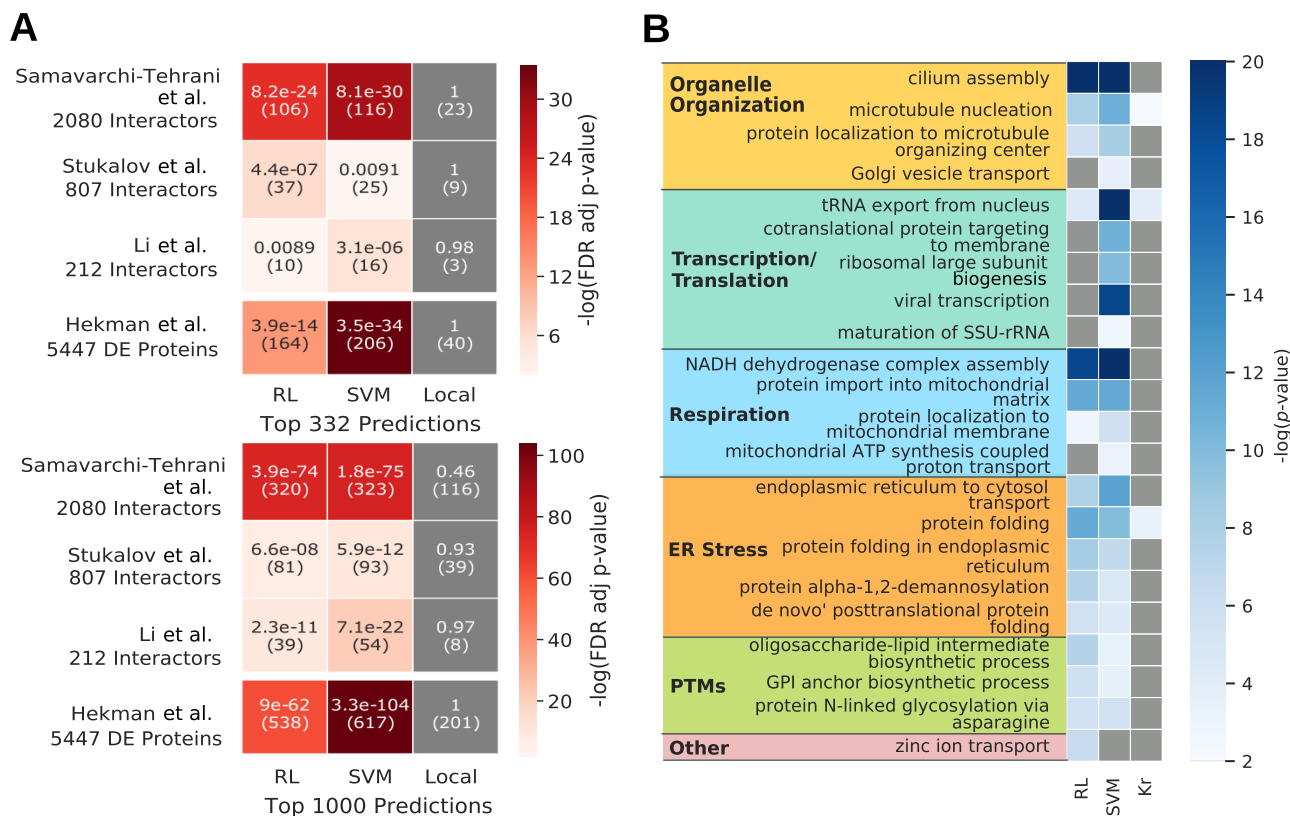
Figure 2: Network propagation results. (A) Heat map showing the FDR-adjusted *P*-value from the hypergeometric test for the overlap between the top-ranking predictions of RL, SVM, and Local and 3 new experimental datasets of SARS-CoV-2–human protein interactions [16–18] and 1 dataset of differentially expressed (DE) proteins after SARS-CoV-2 infection [19]. Each cell displays the FDR-adjusted *P*-value and the number of overlapping proteins in parentheses. A gray cell indicates *P* > 0.01. (B) Heat map summarizing GO biological process terms enriched in top-ranking proteins from RL and SVM and human interactors of SARS-CoV-2 proteins (indicated as "Kr"). We manually grouped the terms into broader categories shown in boldface. A gray cell indicates *P* > 0.01. We examine the relevance of these biological processes to SARS-CoV-2 and COVID-19 in "Enriched Biological Processes" in the Supplementary Results and in "Discussion." adj: adjusted; GPI: glycosylphosphatidylinositol; SSU-rRNA: short subunit ribosomal RNA; tRNA: transfer RNA.

network. To test how this spreading of influence affects our results, we varied α over 4 orders of magnitude from 0.01 to 100 and performed 2 analyses. First and most importantly, for each top-ranking protein computed by the RL, we developed a systematic procedure to determine the provenance of its score, i.e., which SARS-CoV-2 interactors made the greatest contributions to this score. For our second analysis, we developed a new methodology to select a value of α. We were motivated to do so because we could not use the common practice of choosing the parameter's value on the basis of maximization of cross-validation performance: the AUROC, AUPRC, and precision at 0.3 recall of the RL varied very little with α (Supplementary Fig. S3).

For provenance tracing, we took advantage of the fact that the score computed by the RL for each protein in the network is a linear combination of contributions from source proteins (see Methods). Therefore, for each protein u in the network, we sorted the source proteins by their relative contributions to the score of u ("Provenance tracing matrix" [27]). Figure 3A–D provides illustrative examples of the practical usefulness of provenance tracing. We used a value of $\alpha = 3.4$ to obtain these results. We present our method for selecting α at the end of this section.

In Fig. 3A, we display the top 332 ranking proteins computed by the RL that are annotated to the enriched GO term "protein folding in endoplasmic reticulum." For each such protein, we also show all the sources that interact with it, as well as the viral proteins that in turn interact with the sources. This network is

complex and difficult to understand. In contrast, in Fig. 3B, we connect each top-ranking protein only to the 2 source proteins that contribute the most to its score. This simplified network considerably facilitates the interpretation and rationalization of the RL's predictions. Figures 3C and D are similar in nature and correspond to the enriched term "cilium assembly." We return to the biological insights present in these networks in the Discussion.

Next, we considered the effect of α on the amount of diffusion in the network. When α was very small, e.g., 0.01, we expected the highest contributing sources to be direct neighbors of top-ranking proteins. As α increased, and the random walker traversed longer paths in the network, we expected more of the highest contributors to not be directly connected by an edge to top-ranking proteins. Contrary to our expectations, we found that for every value of α and for every top-ranking protein u (till a rank of 1,000), the source protein with the highest contribution to u's score was always a neighbor of u. Even when we considered the second and third highest contributors, we found that they were >1 edge away for as few as 2% of the top-ranking proteins for $\alpha = 0.01$. This number increased only to 25% for $\alpha = 100$.

The STRING network includes both direct, physical and indirect, functional PPIs. Therefore, we sought to see whether this trend in the provenance analysis held for networks with only physical interactions corresponding to direct binding and indi-
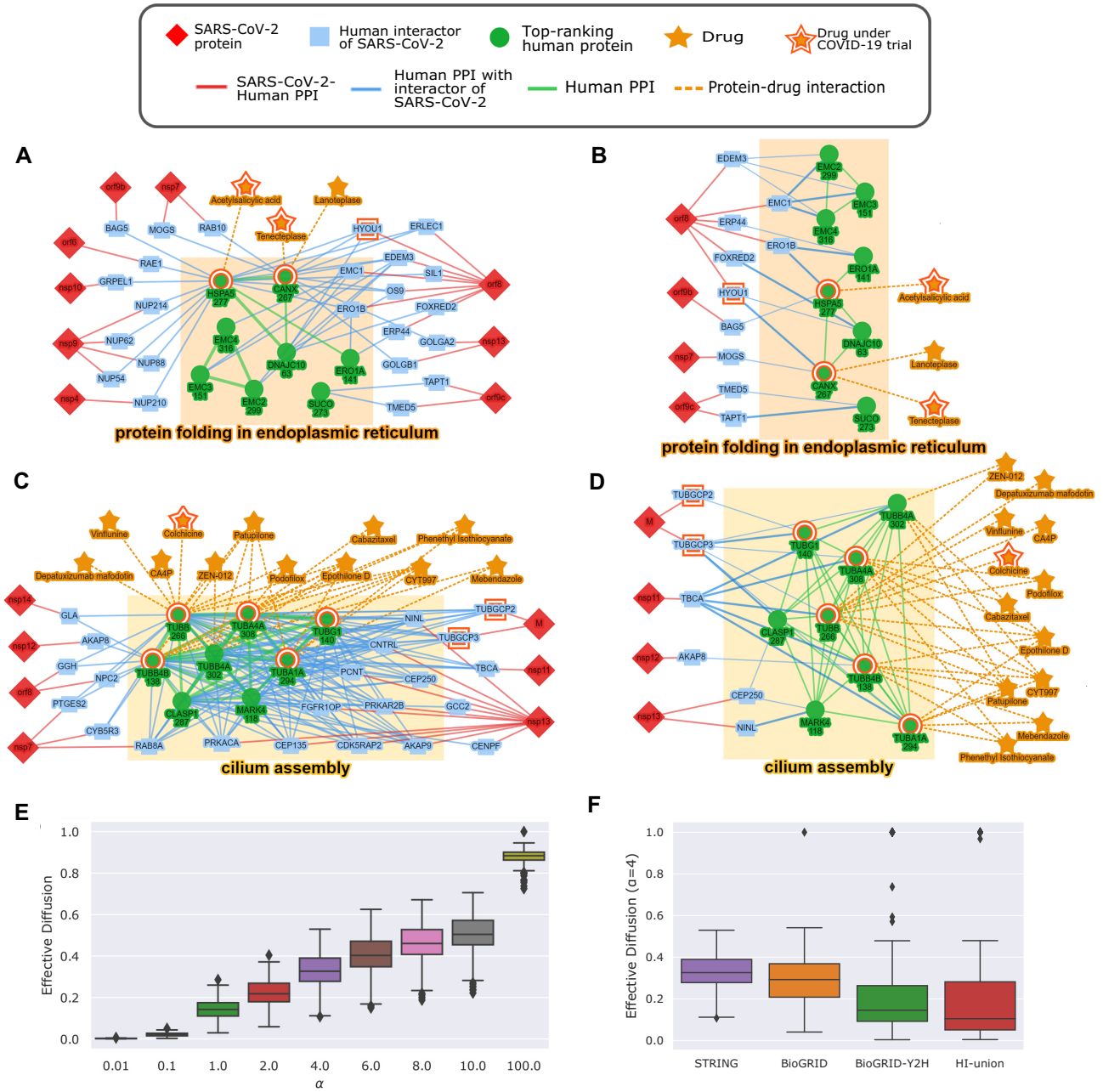
Figure 3: Provenance-tracing results and illustrative examples of networks. (A) Network of the top 332 ranking proteins for RL (green nodes) that are annotated to the enriched term "protein folding in ER." For each top-ranking protein, we display its connections with all neighboring SARS-CoV-2 interactors. (B) The same network as in (A) except that we display only the top 2 contributing SARS-CoV-2 interactors for each top-ranking protein. (C) Network of the top 332 ranking proteins for RL (green nodes) that are annotated to the enriched term "cilium assembly." (D) The same network as in (C) except that we display only the top 2 contributing SARS-CoV-2 interactors for each top-ranking protein. In all 4 network visualizations, the number below the name of a green protein is its rank as computed by the RL. Proteins discussed in the text are highlighted with a red border. In (A and C), we removed STRING edges with weight <700 to simplify the visualization. We retained this restriction in (B and D) as well to maintain consistency between the visualized networks. In (C and D), we removed drugs that promote clotting. (E) Distribution of effective diffusion for the top 332 ranking proteins for different values of α. (F) The same distribution as (E) except comparing different networks with $\alpha = 4.0$. In every boxplot, the box spans the interval between the 1st and 3rd quartiles, with whiskers extending to at most 1.5 times the interquartile range.

rect protein complex membership. We repeated the analyses up to this point on 3 other PPI networks: BioGRID, BioGRID-Y2H, and HI-union (see Methods). For BioGRID, the results were comparable to those for STRING. The highest contributor was always a neighbor, except for $\alpha \geq 10$, where up to 3% of nodes received most of their score from a source >1 edge away. The second and third highest contributor was >1 step away for as few as

8% of top-ranking nodes for $\alpha = 0.01$, and up to 41% for $\alpha = 100$. For BioGRID-Y2H and HI-union, which are smaller, sparser networks with only direct PPIs, only 300–400 nodes had scores that were statistically significant at the 0.05 level. The highest contributing source was >1 step away for as many as 10–30% of the top-ranking nodes, even for $\alpha = 10$. For the second highest contributor, this percentage increased to >50% for $\alpha = 0.01$ itself.

To further characterize the contribution of non-neighboring sources, we defined the "effective diffusion" to a protein u as the fraction of its score $s(u)$ that arose from the non-direct neighbors of u that were also SARS-CoV-2 interactors. As expected, the effective diffusion to the top-ranking proteins increased with α with values close to zero for $\alpha = 0.01$ and a median of 0.88 for $\alpha = 100$ (Fig. 3E). We concluded that the neighbors of the sources received non-trivial contributions to their RL scores from indirectly connected sources only for values of $\alpha = 1$ and higher.

We repeated these experiments for BioGRID, BioGRID-Y2H, and HI-union (see Fig. 3F and Supplementary Fig. S9). BioGRID maintained fairly similar results to STRING. On the other hand, for the other 2 networks, their effective diffusion values were quite a bit smaller (difference from STRING ∼0.2 on average). Taken together, these results suggest that in the sparser networks (BioGRID-Y2H and HI-union), a top-ranking protein has fewer sources as direct neighbors than in the denser networks (STRING and BioGRID) but a larger proportion of its score arises from these adjacent sources.

These results motivated us to test a different method for selecting an appropriate value of α for downstream analysis. As mentioned earlier, we mathematically derived a new expression for the expected value of the path length of the random walker ("Analytical Perspective on the RL and Expected Path Length" in the Supplementary Methods). To our knowledge, no such formula is known for the interpretation of the RL as a continuous-time Markov chain. This value depended on α, the topology of the network, and which proteins interacted with SARS-CoV-2. We computed the expected path length for different values of α (Supplementary Table S1). Independently, we computed the distribution of path lengths in the network from SARS-CoV-2 interactors to every other protein (Supplementary Fig. S10). The median number of edges in these paths was 3. Therefore, we set the value of $\alpha = 3.4$ for which the expected path length of the random walker was 3.04 (Supplementary Table S1). The median effective diffusion for this value of α was ∼0.3. We used this value of α to generate the results presented in this work.

Discussion

The COVID-19 pandemic and its medical and economic impact have created an urgent challenge for biomedical researchers to understand infection mechanisms used by SARS-CoV-2 and to develop therapeutics against the disease [28]. A manifestation of this community response is the first protein-protein interactome associated with the SARS-CoV-2–human interface [11]. This set of human proteins reported to interact with SARS-CoV-2 is likely to have both false-positive and false-negative results due to the properties of the proteomic screening pipeline used.

In this work, we sought to further extend the results of this study to significantly expand the resources available to the COVID-19 community by producing an extended set of putative SARS-CoV-2 interactors. Comparison of our results with independently generated SARS-CoV-2–human protein interaction networks [16–18] provides substantial experimental support for our predictions. We note that complementary efforts are based on protein structures [29], observational studies of treatments being administered to patients [30], shortest paths in protein networks [31], propagation in protein networks with predicted SARS-CoV-2 interactors [32], and exploratory analyses of virus-host-drug networks [33].

A notable new feature of our methodology is tracing the provenance of each of our predictions back to the most infor-

mative experimental sources [10]. In principle, the RL computes scores by integrating over all paths in the network. We were surprised to see that the top-contributing sources were invariably direct neighbors of the top-ranking predictions in the STRING network. A partial explanation for this trend may be the fact that as many as 5,331 proteins in the STRING network were direct neighbors of ≥1 source protein, even when we considered only interactions with weight ≥0.9 (the STRING database deems edges with such weights to be of "very high quality"). Thus, the structure of the STRING network and central location of sources within it may cause the RL both to give high ranks only to direct neighbors of sources and to channel propagation primarily along these direct connections. We stress that using only the interactions between sources and their neighbors in the network does not result in high-quality predictions, as evidenced by the relatively poor cross-validation performance of the Local algorithm. Thus, the integration of multiple paths by the RL plays a key role in prioritizing which neighbors of the sources are more likely to be potential interactors of SARS-CoV-2 proteins than others.

COVID-19 research has focused disproportionately on a small set of human proteins [34]. Our research has the potential to expand the repertoire of host proteins that are studied in the context of COVID-19 and thereby open new directions of study of the disease. The cellular processes in which our top-ranking proteins participate suggest how the virus may infect human cells. We discuss 2 illustrative examples of the type of insights provided by our approach, highlighting several proteins targeted by drugs that are already in clinical trials for COVID-19. We remind the reader that we computed functions enriched in the top-ranking proteins, performed the provenance analysis independently, and then integrated the results in the protein networks we visualized.

The role of endoplasmic reticulum stress, HSPA5, and anti-clotting drugs

Our analysis points to a connection among interactors of SARS-CoV-2, proteins involved in ER stress, and anti-clotting drugs (Fig. 3A and B). The GO biological process "protein folding in endoplasmic reticulum" was enriched in the top-ranking proteins ($P = 4.32 \times 10^{-9}$ for RL and 0.28 for interactors of SARS-CoV-2). HSPA5, also referred to as glucose regulated protein (GRP78) or immunoglobulin binding protein (BiP) in the literature, is evolutionarily conserved from prokaryotes to humans [35]. 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 [36,37]. On the cell surface, HSPA5 plays a multifunctional role in cell proliferation, cell viability, apoptosis, and regulation of innate and adaptive immunity [37,38].

HSPA5 has been proposed as a universal target for human diseases [39]. 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. SARS-CoV infection has been shown to lead to ER stress and the up-regulation of HSPA5 [40,41]. The S protein of SARS-CoV can induce transcriptional activation of HSPA5 [41]. This protein can serve as a point of attachment for both MERS-CoV and bat coronavirus (bCoV HKU9) [42]. Both Zika virus and Japanese encephalitis virus use HSPA5 to prevent apoptosis and to help in viral replication [43]. A recent molecular docking study has predicted HSPA5 as a potential receptor for the SARS-CoV S protein [44]. The observed expression *in vitro* of HSPA5 in airway epithelial cells suggests that it may serve as an additional

receptor for SARS-CoV-2 in these cells [45]. On the basis of 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 patients with COVID-19 [46]. Top-ranking proteins HSPA5 and CANX act as 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 [47]. Although Factor V, Factor VIII, and vWF are not among the top-ranking proteins and thus do not appear in Fig. 3A and B, this network is suggestive of mechanisms that SARS-CoV-2 may use to cause abnormal blood coagulation.

Anti-coagulant drugs that interact with HSPA5 or CANX 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 [48], 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 are clinical trials for tenecteplase (NCT04558125, NCT04505592) and alteplase (NCT04357730, NCT04640194) to test their effectiveness in treating COVID-19. Aspirin, also present in Fig. 3A and B, binds to and inhibits the ATPase activity of HSPA5 [49]. Aspirin is currently involved in 16 clinical trials, with one testing the effects of aspirin at various levels of COVID-19 severity (NCT04365309) and another testing whether early treatment of patients with COVID-19 with aspirin and vitamin D can inhibit the production of blood clots and decrease rates of hospitalization (NCT04363840).

Cilium assembly and tubulin-modulating drugs

GO biological processes related to cilia were significantly enriched in the top-ranking RL and SVM predictions. An example is "cilium assembly" ($P = 6.84 \times 10^{-26}$ for RL vs 0.31 in the human interactors of SARS-CoV-2). Many proteins annotated to this term belong to the tubulin family, which are components of microtubules. The SARS-CoV-2 M protein binds to 2 γ-tubulins (TUBGCP2 and TUBGCP3), which interact with several α- and β-tubulins among the top 332 predictions (Fig. 3C and D). Microtubules are polymers that provide shape and structure to eukaryotic cells and are necessary in cell transport and cell division, among other functions [50]. The α- and β-tubulins compose microtubule filaments, while γ-tubulins connect them to the microtubule organizing center.

Viruses commonly utilize microtubules for cellular entry, intra-cellular trafficking, and exit from cells [51]. For instance, the S protein of human α-coronavirus interacts with tubulin α and β chains [52], suggesting that tubulin may be involved in the transport and localization of the S protein and its assembly into virions [52]. 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 [53]. The ACE2 receptor that SARS-CoV-2 uses to enter cells appears to be expressed primarily on the cilia of respiratory tract epithelial cells [54,55], 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 [56] and the deleterious effect of COVID-19 on the renal, gastroinstestinal, and olfactory systems [57]. The drugs that target Tubulin proteins (Fig. 3C and D) are mostly anti-mitotic agents, which are being investigated as anti-cancer therapeutics. It is notable that 26 ongoing clinical trials are testing the effectiveness of colchicine against COVID-19.

Our work also sets the stage for follow-up analyses on SARS-CoV-2. Integrating new datasets of SARS-CoV-2–human protein interactions [16–18] and human proteins whose deletion inhibits viral replication [58, 59] with other omics data using our methods and with orthogonal analysis techniques promises to predict more biologically meaningful networks and processes affected by the virus. In particular, single-cell RNA-seq data offer many opportunities to examine cellular heterogeneity and context-specific interactions.

Potential Implications

The approach that we advocate here is inspired by the general framework of producing explanations for machine learning methods [60]. This area of "explanations" of predictions is receiving strong interest because of deep learning. While the idea has previously been studied in graphical models [61], most machine learning methods are not fully interpretable by the fairly strict definition of Kasif and Roberts [10]: tracing each prediction to the experimental evidence that supports it. This notion of explanation is a special but particularly important case for computational genomics and systems biology.

Causal perturbations [61] provide a general approach for producing explanations of this type for virtually any predictive model. Consider a model with experimental evidence that a gene g performs a function f. We perturb the variable associated with the gene, e.g., we change the probability $\Pr(g$ performs $f) = 1$ to $\Pr(g$ performs $f) = 0$. We then compute the change in probability of every other variable in the model due to this perturbation in order to assess the importance of this particular gene-function pair.

For network propagation, this idea yields the special case discussed in this work that is amenable to very efficient computation. Our strategy for tracing provenance extends to any algorithm that makes predictions using a linear combination of evidence such as logistic regression and GeneMania [62]. In particular, it is applicable to the large number of random-walk–based methods that have been developed for predicting disease genes or annotations to GO terms [63–66].

An important future line of research will be to develop provenance-tracing techniques for other classes of network-based methods such as Markov random fields (MRFs) [67,68] and min-cut–based methods [6,69]. For MRFs, we can apply the general perturbation-based method described above. For min-cut–based methods, it is possible to recalculate the cut for any single change in experimental data using dynamic data structures [70]. Thus, the provenance-tracing approach that we advocate here has many natural follow-ups that we expect to be studied by the community in the future.

It remains to be seen whether the trends we observed on the contributions from direct neighbors generalize to these methods and to annotations of terms in the GO or the Human Phenotype Ontology terms. In general, it is quite likely that sources that are not direct neighbors may make substantial contributions to scores. In these cases, new algorithmic developments may be required to trace the paths by which the sources spread their influence to a given node.

Our work provides significant new data and software resources to the COVID-19 community. Three properties of our results facilitate their use by experimentalists who are seeking to obtain new insights into the pathogenesis of this disease.

First, the prioritized list of predicted interactors of SARS-CoV-2 ("List of RL and SVM predictions, P-values, and top-2 contributors" [27]) contains druggable targets that may be promising to study further. Second, our provenance analysis provides the rationale underlying each prediction by directly linking to the relevant experimental input. Third, the viral-human protein interaction networks corresponding to enriched GO terms (Fig. 3 and Supplementary Fig. S6) are available for visualization and download on GraphSpace [71]. Examination of these networks provides further context for the predictions.

We conclude by noting that our methodology is general purpose and easy to generalize to a new virus. The software requires a dataset of host proteins that interact with the virus and an interaction network among the host proteins themselves. The virus-host network may be determined experimentally [11]. If such a dataset is not available, a user can predict the network computationally from the sequence of the viral genes and interaction networks for phylogenetically similar viruses [72]. Subsequently, a user can apply network propagation to predict additional human proteins and biological processes that may be targeted by the virus.

Methods

Algorithms

To facilitate the complete reproducibility of our results, we now describe the RL algorithm that we use for label propagation and prediction. We present the other methods that we use (GeneMANIA, SinkSource, RWR, Local, deepNF, the Support Vector Machine, and logistic Regression) and implementation details in "Other Algorithms" in the Supplementary Methods. We are given a weighted, undirected network $G = (V, E, w)$, where each node in V is a human protein, each edge (u, v) represents an interaction between proteins u and v, and $w: E \rightarrow (0, 1]$ is a function specifying the weight of each edge in E. Informally, the weight of an edge indicates our confidence in the experimental data supporting the corresponding protein-protein interaction. We are also given a set $P \in V$ of positive examples consisting of the human proteins that interact with SARS-CoV-2 proteins [11]. Each node in G is a human protein and each edge represents a physical or functional interaction between 2 proteins. We seek to compute a score vector $\vec{s} \in \mathbb{R}^n$, where n is the number of nodes in G. For every node v, the score $s(v)$ in this vector indicates our confidence that node v either physically interacts with or is functionally linked to a SARS-CoV-2 protein.

Regularized Laplacian [22]
Given a parameter $\alpha > 0$, we compute $\vec{s}$ using the following steps:

(1) Define a label vector $\vec{y}$ over the nodes in G where $y(u) = 1$ if node u is in P and $y(u) = 0$, otherwise.
(2) Define $W \in \mathbb{R}^{n \times n}$ as the adjacency matrix of G with edge weights, i.e., the entry in row u and column v of W equals w_{uv} if (u, v) is an edge in G and 0, otherwise.
(3) Define D as a diagonal matrix with $D_{uu} = \sum_v w_{uv}$, for every node u in G.
(4) Compute the $\mathbb{R}^{n \times n}$ matrix $\tilde{W} = D^{-1/2} W D^{-1/2}$, which denotes the normalized network.
(5) Compute the Laplacian of G as $\tilde{L} = \tilde{D} - \tilde{W}$, where we define $\tilde{D}$ to be a diagonal matrix with $\tilde{D}_{uu} = \sum_v \tilde{w}_{uv}$.
(6) Compute the vector $\vec{s} = (I + \alpha \tilde{L})^{-1} \vec{y}$.

The RL was introduced by Zhou and Schölkopf [73]. Since then, several variations of this method have been published.

The version we use is identical to the strategy used by Fouss et al. [22]. We provide the intuition behind the resulting RL matrix (i.e., $(I + \alpha \tilde{L})^{-1}$) and discuss its properties in "Analytical Perspective on the RL and Expected Path Length" in the Supplementary Methods. In particular, we derive an expression for the expected path length of the continuous-time Markov chain corresponding to the RL. As far as we know, this mathematical analysis has not previously been published.

Tracing the provenance of prediction scores

Let K denote the RL matrix $(I + \alpha \tilde{L})^{-1}$. We remind the reader that the RL algorithm ranks proteins on the basis of 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. For every protein u, we sorted the proteins in P in decreasing order of the values of K_{uv}, where v ranged over P. In this manner, we ranked the experimentally determined interactors in decreasing order of their contributions to each node's diffusion score. This analysis is important for tracing the provenance of computational predictions to their experimental sources [10].

Availability of Source Code and Requirements

- Project name: SARS-CoV-2-network-analysis
- Project home page: https://github.com/Murali-group/SARS-CoV-2-network-analysis
- Operating system(s): Platform independent (tested and applied on Linux and Mac OS)
- Programming language: Python
- Other requirements: See https://github.com/Murali-group/SARS-CoV-2-network-analysis/blob/master/requirements.txt
- License: GNU General Public License v3
- biotools id: biotools:sars-cov-2-network-analysis
- RRID:SCR_021811

Data Availability

We used publicly available datasets for our analysis. We downloaded these data from the respective publications or websites. A snapshot of the software used for this analysis and the following supplementary files are available at the *GigaScience* GigaDB database [27].

Additional Files

List of RL and SVM predictions, P-values, and top-2 contributors: 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 top-2 contributing SARS-CoV-2 interactors and corresponding SARS-CoV-2 protein. For the last piece of information, we also included the fraction of score contributed by each of the top-2 SARS-CoV-2 interactors.

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

Provenance-tracing matrix: Provenance-tracing matrix of contributions to the network propagation score from each SARS-CoV-2 interactor to every top-ranking protein.

Supplementary Methods.

Supplementary Figure S1. Cross validation results.

Abbreviations

AP-MS: affinity purification followed by mass spectrometry analysis; ATP: adenosine triphosphate; AUPRC: area under the precision-recall curve; AUROC: area under the receiver-operator characteristic curve; bCoV: bat coronavirus; COVID-19: novel coronavirus disease 2019; BioID: proximity-dependent biotinylation; BiP: immunoglobulin binding protein; ER: endoplasmic reticulum; FDR: false discovery rate; GM: GeneMania; GPL: General Public License; GO: Gene Ontology; GRP: glucose regulated protein; HIV-1: human immunodeficiency virus 1; HSV-1: herpes simplex virus type 1; iAT2: induced pluripotent stem cell–derived alveolar epithelial Type 2 cell; MERS: Middle East respiratory syndrome; NSF: National Science Foundation; RWR: random walk with restarts; RL: regularized Laplacian; SARS: severe acute respiratory syndrome; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2; SVM: Support Vector Machine; USDA-NIFA: United States Department of Agriculture National Institute of Food and Agriculture; vWF: von Willebrand factor.

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

T.M.M. and S.K. proposed the study. T.M.M., S.K., M.C., J.N.L., S.D., M.K., and J.K. contributed computational ideas. J.N.L. was the primary author of the software and led the computational analysis, with significant inputs from K.A., N.T., and C.M.D.S. All authors analyzed the results. T.M.M., M.C., P.R., and S.K. wrote the manuscript with contributions and revisions from all authors. All the authors read and approved the final manuscript.

References

1. Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol 2003;**21**(6):697–700.

2. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 2003;**19**(Suppl 1):i197–204.

3. Karaoz U, Murali TM, Letovsky S, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci U S A 2004;**101**(9):2888–93.

4. Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. J Comput Biol 2004;**11**(2-3):463–75.

5. Fraser AG, Marcotte EM. A probabilistic view of gene function. Nat Genet 2004;**36**(6):559–64.

6. Murali TM, Wu CJ, Kasif S. The art of gene function prediction. Nat Biotechnol 2006;**12**:1474–5.

7. Ideker T, Sharan R. Protein networks in disease. Genome Res 2008;**18**(4):644–52.

8. Leiserson MD, Vandin F, Wu HT, et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nat Genet 2015;**47**(2):106–14.

9. van Dijk D, Sharma R, Nainys J, et al. Recovering gene interactions from single-cell data using data diffusion. Cell 2018;**174**(3):716–29.

10. Kasif S, Roberts RJ. We need to keep a reproducible trace of facts, predictions, and hypotheses from gene to function in the era of big data. PLoS Biol 2020;**18**(11):e3000999.

11. Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020;**583**(7816):459–68.

12. Szklarczyk D, Morris JH, Cook H, et al. The STRING Database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2016;**45**(D1):D362–8.

13. Oughtred R, Rust J, Chang C, et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 2021;**30**(1):187–200.

14. Luck K, Kim DK, Lambourne L, et al. A reference map of the human binary protein interactome. Nature 2020;**580**(7803):402–8.

15. Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. Nucleic Acids Res 2018;**46**(D1):D1074–82.

16. Stukalov A, Girault V, Grass V, et al. Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. Nature 2021;**594**(7862):246–52.

17. Li J, Guo M, Tian X, et al. Virus-host interactome and proteomic survey reveal potential virulence factors influencing SARS-CoV-2 pathogenesis. Med (N Y) 2021;**2**(1):99–112.e7.

18. Samavarchi-Tehrani P, Abdouni H, Knight JDR, et al. A SARS-CoV-2 – host proximity interactome. bioRxiv 2020:doi:10.1101/2020.09.03.282103.

19. Hekman RM, Hume AJ, Goel RK, et al. Actionable cytopathogenic host responses of human alveolar type 2 cells to SARS-CoV-2. Mol Cell 2020;**80**(6):1104–22.

20. Mick E, Kamm J, Pisco AO, et al. Upper airway gene expression reveals suppressed immune responses to SARS-CoV-2 compared with other respiratory viruses. Nat Commun 2020;**11**(1):5854.

21. Cowen L, Ideker T, Raphael BJ, et al. Network propagation: A universal amplifier of genetic associations. Nat Rev Genet 2017;**18**(9):551–62.

22. Fouss F, Francoisse K, Yen L, et al. An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. Neural Netw 2012;**31**:53–72.

23. Page L, Brin S, Motwani R, et al. The PageRank citation ranking: Bringing order to the Web. Technical Report 1999-66. Stanford InfoLab; 1999. http://ilpubs.stanford.edu:8090/422/

24. Mostafavi S, Ray D, Warde-Farley D, et al. GeneMANIA: A real-time multiple association network integration algorithm for predicting gene function. Genome Biol 2008;**9**(Suppl 1):S4.

25. Murali TM, Dyer MD, Badger D, et al. Network-based prediction and analysis of HIV dependency factors. PLoS Comput Biol 2011;**7**(9):e1002164.

26. Gligorijević V, Barot M, Bonneau R. deepNF: Deep Network Fusion for Protein Function Prediction. Bioinformatics 2018;**22**:3873–81.

27. Law J, Akers K, Tasnina N, et al. Supporting data for "Interpretable network propagation with application to expanding the repertoire of human proteins that interact with SARS-CoV-2." GigaScience Database 2021. http://dx.doi.org/10.5524/100941.

28. Guy RK, DiPaola RS, Romanelli F, et al. Rapid repurposing of drugs for COVID-19. Science 2020;**368**(6493):829–30.

29. Wu C, Liu Y, Yang Y, et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm Sin B 2020;**10**(5):766–88.

30. Vaduganathan M, Vardeny O, Michel T, et al. Renin–angiotensin–aldosterone system inhibitors in patients with Covid-19. N Engl J Med 2020;**382**(17):1653–9.

31. Zhou Y, Hou Y, Shen J, et al. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov 2020;**6**:14.

32. Zhang Y, Zeng T, Chen L, et al. Identification of COVID-19 infection-related human genes based on a random walk model in a virus–human protein interaction network. Biomed Res Int 2020;**2020**:4256301.

33. Sadegh S, Matschinske J, Blumenthal DB, et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. Nat Commun 2020;**11**(1):3518.

34. Stoeger T, Nunes Amaral LA. COVID-19 research risks ignoring important host genes due to pre-established research patterns. Elife 2020;**9**:doi:10.7554/eLife.61981.

35. Lee AS. Glucose-regulated proteins in cancer: molecular mechanisms and therapeutic potential. Nat Rev Cancer 2014;**14**(4):263–76.

36. Zhang Y, Liu R, Ni M, et al. Cell surface relocalization of the endoplasmic reticulum chaperone and unfolded protein response regulator GRP78/BiP. J Biol Chem 2010;**285**(20):15065–75.

37. Tsai YL, Ha DP, Zhao H, et al. Endoplasmic reticulum stress activates SRC, relocating chaperones to the cell surface where GRP78/CD109 blocks TGF-β signaling. Proc Natl Acad Sci U S A 2018;**115**(18):E4245–54.

38. Ni M, Zhang Y, Lee AS. Beyond the endoplasmic reticulum: Atypical GRP78 in cell viability, signalling and therapeutic targeting. Biochem J 2011;**434**(2):181–8.

39. Booth L, Roberts JL, Cash DR, et al. GRP78/BiP/HSPA5/Dna K is a universal therapeutic target for human disease. J Cell Physiol 2015;**230**(7):1661–76.

40. DeDiego ML, Nieto-Torres JL, Jiménez-Guardeño JM, et al. Severe acute respiratory syndrome coronavirus envelope protein regulates cell stress response and apoptosis. PLoS Pathog 2011;**7**(10):e1002315.

41. Chan CP, Siu KL, Chin KT, et al. Modulation of the unfolded protein response by the severe acute respiratory syndrome coronavirus spike protein. J Virol 2006;**80**(18):9279–87.

42. Chu H, Chan CM, Zhang X, et al. Middle East respiratory syndrome coronavirus and bat coronavirus HKU9 both can utilize GRP78 for attachment onto host cells. J Biol Chem 2018;**293**(30):11709–26.

43. Lyoo HR, Park SY, Kim JY, et al. Constant up-regulation of BiP/GRP78 expression prevents virus-induced apoptosis in BHK-21 cells with Japanese encephalitis virus persistent infection. Virol J 2015;**12**:32.

44. Ibrahim IM, Abdelmalek DH, Elshahat ME, et al. COVID-19 spike-host cell receptor GRP78 binding site prediction. J Infect 2020;**80**(5):554–62.

45. Aguiar JA, Tremblay BJM, Mansfield MJ, et al. Gene expression and in situ protein profiling of candidate SARS-CoV-2 receptors in human airway epithelial cells and lung tissue. Eur Resp J 2020;**56**:2001123.

46. Terpos E, Ntanasis-Stathopoulos I, Elalamy I, et al. Hematological findings and complications of COVID-19. Am J Hematol 2020;**95**(7):834–47.

47. Kaufman RJ, Pipe SW, Tagliavacca L, et al. Biosynthesis, assembly and secretion of coagulation factor VIII. Blood Coagul Fibrinolysis 1997;**8**(Suppl 2):3–14.

48. Flemmig M, Melzig MF. Serine-proteases as plasminogen activators in terms of fibrinolysis. J Pharm Pharmacol 2012;**64**(8):1025–39.

49. Deng WG, Ruan KH, Du M, et al. Aspirin and salicylate bind to immunoglobulin heavy chain binding protein (BiP) and inhibit its ATPase activity in human fibroblasts. FASEB J 2001;**15**(13):2463–70.

50. Nogales E. Structural insights into microtubule function. Annu Rev Biochem 2000;**69**:277–302.

51. Greber UF, Way M. A superhighway to virus infection. Cell 2006;**124**(4):741–54.

52. Rüdiger AT, Mayrhofer P, Ma-Lauer Y, et al. Tubulins interact with porcine and human S proteins of the genus *Alphacoronavirus* and support successful assembly and release of infectious viral particles. Virology 2016;**497**:185–97.

53. Satir P, Christensen ST. Overview of structure and function of mammalian cilia. Annu Rev Physiol 2007;**69**:377–400.

54. Lee IT, Nakayama T, Wu CT, et al. ACE2 localizes to the respiratory cilia and is not increased by ACE inhibitors or ARBs. Nat Commun 2020;**11**(1):5453.

55. Sungnak W, Huang N, Bécavin C, et al. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. Nat Med 2020;**26**(5):681–7.

56. Puelles VG, Lütgehetmann M, Lindenmeyer MT, et al. Multiorgan and renal tropism of SARS-CoV-2. N Engl J Med 2020;**383**(6):590–2.

57. Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020;**395**(10223):497–506.

58. Wei J, Alfajaro MM, Hanna RE, et al. Genome-wide CRISPR screen reveals host genes that regulate SARS-CoV-2 infection. bioRxiv 2020:doi:10.1101/2020.06.16.155101.

59. Daniloski Z, Jordan TX, Wessels HH, et al. Identification of required host factors for SARS-CoV-2 infection in human cells. Cell 2021;**184**(1):92–105.e16.

60. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016:1135–44.

61. Pearl J. Causality. Cambridge University Press; 2009. http://bayes.cs.ucla.edu/BOOK-2K/. Accessed: 8 December, 2021.

62. Mostafavi S, Ray D, Warde-Farley D, et al. GeneMANIA: A real-time multiple association network integration algorithm for predicting gene function. Genome Biol 2008;**9**(Suppl 1):S4.

63. Vanunu O, Magger O, Ruppin E, et al. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol 2010;**6**(1):e1000641.

64. Komurov K, White MA, Ram PT. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. PLoS Comput Biol 2010;**6**(8):e1000889.

65. Jiang B, Kloster K, Gleich DF, et al. AptRank: An adaptive PageRank model for protein function prediction on bi-relational graphs. Bioinformatics 2017;**33**(12):1829–36.

66. Hristov BH, Chazelle B, Singh M. uKIN combines new and prior information with guided network propagation to accurately identify disease genes. Cell Syst 2020;**10**(6):470–9.e3.

67. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: A probabilistic approach. Bioinformatics 2003;**19**(Suppl 1):i197–204.

68. Deng M, Tu Z, Sun F, et al. Mapping Gene Ontology to proteins based on protein-protein interaction data. Bioinformatics 2004;**20**(6):895–902.

69. Nabieva E, Jim K, Agarwal A, et al. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 2005;**21**(Suppl 1):i302–10.

70. Goranci G, Henzinger M, Thorup M. Incremental exact min-cut in polylogarithmic amortized update time. ACM Trans Algorithms 2018;**14**(2):doi:10.1145/3174803.

71. GraphSpace SARS-Cov2 Network Anaylsis, http://graphspace.org/graphs/?query=tags:2021-sarscov2-network-analysis. Accessed: 1 November, 2021.

72. Kshirsagar M, Tasnina N, Ward MD, et al. Protein sequence models for prediction and comparative analysis of the SARS-CoV-2-human interactome. Pac Symp Biocomput 2021;**26**:154–65.

73. Zhou D. Schölkopf B A regularization framework for learning from graph data. ICML 2004 Workshop on Statistical Relational Learning and Its Connections to Other Fields 2004;132–137.