Multi-schema computational prediction of the comprehensive SARS-CoV-2 vs. human interactome

Kevin Dick [1, 2], **Anand Chopra** [3, 4], **Kyle K Biggar** [3, 4], **James R Green** [Corresp. 1, 2]

[1] Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada

[2] Institute for Data Science, Carleton University, Ottawa, Ontario, Canada

[3] Institute of Biochemistry, Carleton University, Ottawa, Ontario, Canada

[4] Department of Biology, Carleton University, Ottawa, Ontario, Canada

Corresponding Author: James R Green
Email address: jrgreen@sce.carleton.ca

Background. Understanding the disease pathogenesis of the novel coronavirus, denoted SARS-CoV-2, is critical to the development of anti-SARS-CoV-2 therapeutics. The global propagation of the viral disease, denoted COVID-19 (``coronavirus disease 2019"), has unified the scientific community in searching for possible inhibitory small molecules or polypeptides. A holistic understanding of the SARS-CoV-2 vs. human inter-species interactome promises to identify putative protein-protein interactions (PPI) that may be considered targets for the development of inhibitory therapeutics.

Methods. We leverage two state-of-the-art, sequence-based PPI predictors (PIPE4 \& SPRINT) capable of generating the comprehensive SARS-CoV-2 vs. human interactome, comprising approximately 285,000 pairwise predictions. Three prediction schemas (*all*, *proximal*, *RP-PPI*) are leveraged to obtain our highest-confidence subset of PPIs and human proteins predicted to interact with each of the 14 SARS-CoV-2 proteins considered in this study. Notably, the use of the Reciprocal Perspective (RP) framework demonstrates improved predictive performance in multiple cross-validation experiments.

Results. The *all* schema identified 279 high-confidence putative interactions involving 225 human proteins, the *proximal* schema identified 129 high-confidence putative interactions involving 126 human proteins, and the *RP-PPI* schema identified 539 high-confidence putative interactions involving 494 human proteins. The intersection of the three sets of predictions comprise the seven highest-confidence PPIs. Notably, the Spike-ACE2 interaction was the highest ranked for both the PIPE4 and SPRINT predictors with the *all* and *proximal* schemas, corroborating existing evidence for this PPI. Several other predicted PPIs are biologically relevant within the context of the original SARS-CoV virus. Furthermore, the PIPE-Sites algorithm was used to identify the putative subsequence that might mediate each interaction and thereby inform the design of inhibitory polypeptides intended to disrupt the corresponding host-pathogen interactions.

Conclusion. We publicly released the comprehensive sets of PPI predictions and their corresponding PIPE-Sites landscapes in the following DataVerse repository: https://www.doi.org/10.5683/SP2/JZ77XA. The information provided represents theoretical modeling only and caution should be exercised in its use. It is intended as a resource for the scientific community at large in furthering our understanding of SARS-CoV-2.

Multi-Schema Computational Prediction of the Comprehensive SARS-CoV-2 vs. Human Interactome

Kevin Dick[1,2], **Anand Chopra**[3,4], **Kyle K. Biggar**[3,4], **and James R. Green**[1,2]

[1]**Department of Systems & Computer Engineering, Carleton University, Ottawa, ON, Canada K1S 5B6**
[2]**Institute for Data Science, Carleton University, Ottawa, ON, Canada K1S 5B6**
[3]**Institute of Biochemistry, Carleton University, Ottawa, ON, Canada K1S 5B6**
[4]**Department of Biology, Carleton University, Ottawa, ON, Canada K1S 5B6**

Corresponding author:
James R. Green

Email address: jrgreen@sce.carleton.ca

ABSTRACT

Background. Understanding the disease pathogenesis of the novel coronavirus, denoted SARS-CoV-2, is critical to the development of anti-SARS-CoV-2 therapeutics. The global propagation of the viral disease, denoted COVID-19 ("coronavirus disease 2019"), has unified the scientific community in searching for possible inhibitory small molecules or polypeptides. A holistic understanding of the SARS-CoV-2 vs. human inter-species interactome promises to identify putative protein-protein interactions (PPI) that may be considered targets for the development of inhibitory therapeutics.

Methods. We leverage two state-of-the-art, sequence-based PPI predictors (PIPE4 & SPRINT) capable of generating the comprehensive SARS-CoV-2 vs. human interactome, comprising approximately 285,000 pairwise predictions. Three prediction schemas (*all*, *proximal*, *RP-PPI*) are leveraged to obtain our highest-confidence subset of PPIs and human proteins predicted to interact with each of the 14 SARS-CoV-2 proteins considered in this study. Notably, the use of the Reciprocal Perspective (RP) framework demonstrates improved predictive performance in multiple cross-validation experiments.

Results. The *all* schema identified 279 high-confidence putative interactions involving 225 human proteins, the *proximal* schema identified 129 high-confidence putative interactions involving 126 human proteins, and the *RP-PPI* schema identified 539 high-confidence putative interactions involving 494 human proteins. The intersection of the three sets of predictions comprise the seven highest-confidence PPIs. Notably, the Spike-ACE2 interaction was the highest ranked for both the PIPE4 and SPRINT predictors with the *all* and *proximal* schemas, corroborating existing evidence for this PPI. Several other predicted PPIs are biologically relevant within the context of the original SARS-CoV virus. Furthermore, the PIPE-Sites algorithm was used to identify the putative subsequence that might mediate each interaction and thereby inform the design of inhibitory polypeptides intended to disrupt the corresponding host-pathogen interactions.

Conclusion. We publicly released the comprehensive sets of PPI predictions and their corresponding PIPE-Sites landscapes in the following DataVerse repository: 10.5683/SP2/JZ77XA. It is intended as a resource for the scientific community at large in furthering our understanding of SARS-CoV-2.

INTRODUCTION

The novel coronavirus (CoV) pandemic has galvanized the research community into the investigation of the SARS-CoV-2 virus and the COVID-19 disease it manifests in humans (Guarner, 2020). Research has progressed with unprecedented speed in large part due to the rapid determination of the SARS-CoV-2 genome and proteome. These data enable the research community to collectively contribute to the study and understanding of SARS-CoV-2 and its disease pathogenesis. Given the emergence of three human coronaviruses (HCoVs) causative of severe disease of epidemic or pandemic proportions within the last two decades, we must expand our fundamental understanding of these viruses to rapidly identify putative

therapeutic targets, facilitate complimentary research, and inform public discussions for the present and any future outbreaks of HCoVs.

Coronaviruses share many similarities to the influenza viruses in that they are both enveloped, single-stranded, and helical RNA-viruses among the Group IV viral families (Baltimore, 1971). The four coronaviruses known to commonly infect humans are believed to have evolved such that they maximize proliferation within a population. This evolved strategy involves sickening, but not ultimately killing, their hosts. By contrast, the two prior novel coronavirus outbreaks (SARS and MERS) arose in humans after cross-species jumps from animals, as was H5N1 (the avian influenza). These latter diseases were highly fatal to humans, with relatively few mild or asymptomatic cases. A greater proportion of mild or asymptomatic cases would have resulted in wide-spread disease, however, SARS and MERS each ultimately killed fewer than 1,000 people (World Health Organization, 2020; Regional Office for the Eastern Mediterranean, 2011).

All known HCoVs arise from zoonotic origins (*i.e.* from other animal species). The wide diversity of CoVs within the animal kingdom stem from the genetic alterations to CoV genomes through acquisition of mutations and a high frequency of recombination between different CoV genomes (Makino et al., 1986; Van Der Most et al., 1992). Such genetic modifications occurring in animal CoVs may facilitate a "host jump" and are the primary reason for inter-species and animal-to-human transmission (Cui et al., 2019). The HCoVs that are endemic to the human population are causative agents of more mild disease (*e.g.* common cold) and there is less urgency to identify the animal reservoirs of these viruses.

CoVs are enveloped viruses with a mostly spherical membrane approximately 120 nm in diameter and comprised of 4-5 structural proteins. The single-stranded RNA genome is encapsulated by the Nucleocapsid (N) protein, which functions to package the viral genome into CoV particles during assembly (Chang et al., 2006). The Membrane (M) protein plays a central role in assembly of the viral particles, largely by promoting membrane curvature (Neuman et al., 2011). The Envelope (E) protein is multi-functional, playing key roles in viral assembly and maintenance, such as mediating ion-channel activity (Schoeman and Fielding, 2019). The large membrane projections are trimers of the Spike (S) glycoprotein, responsible for attachment and entry into target cells. Additional smaller 8 nm projections are inherent to lineage A βCoVs, due to the presence of hemagglutinin esterase (HE) dimers.

It is of critical importance that the cellular entry mechanism and viral replication pathways of SARS-CoV-2 and the role of accessory proteins be rapidly elucidated to develop anti-viral therapies to mitigate the spread and infectivity of the virus in the present pandemic.

Promisingly, many computational approaches have been rapidly deployed to increase our understanding of SARS-CoV-2, including protein function, three-dimensional (3D) protein structures, and possible target regions for small inhibitory molecules (Senior et al., 2020; Smith and Smith, 2020). Given that the Spike protein from the original SARS coronavirus, SARS-CoV, is known to interact with the human Angiotensin-Converting Enzyme 2 (ACE2), current efforts are focused to better characterize the SARS-CoV-2 Spike protein and its putative interaction with the ACE2 protein.

Similar efforts are being made to understand the functional and evolutionary characteristics of the SARS-CoV-2 proteome, including the determination of evolutionary conserved functional regions between related viruses to inform the use of anti-viral therapeutics (Cui et al., 2020). Given the unique infectivity characteristics of this novel coronavirus, the need for effective anti-viral therapeutics is pressing. The long viral incubation period, during which an individual is simultaneously contagious and asymptomatic, has resulted in rapid global proliferation. Leveraging what is known from the original SARS-CoV outbreak and related viral families (previously introduced) this work contributes predicted protein-protein interaction (PPI) networks to guide researchers and form the basis of testable hypotheses warranting wet-lab confirmation.

We hope to contribute to the scientific effort using the latest version of our sequence-based protein-protein interaction (PPI) predictor, PIPE4 (Dick et al., 2020b) in combination with another state-of-the-art PPI predictor, denoted Scoring PRotein INTeractions (SPRINT) (Li and Ilie, 2017). Additionally, we leverage the Reciprocal Perspective (RP) cascaded classification method to further refine predictions (Dick and Green, 2018). We leverage a *multi-schema* methodology in order to identify a high-confidence subset of putative interactors. The three predicted interactomes were leveraged in combination to produce candidate targets for experimental validation and to subsequently guide the development of inhibitory polypeptides. Finally, the PIPE-Sites algorithm was used to predict the sub-sequence regions with a high likelihood of mediating the physical interaction between two given pairs (Amos-Binks et al., 2011).

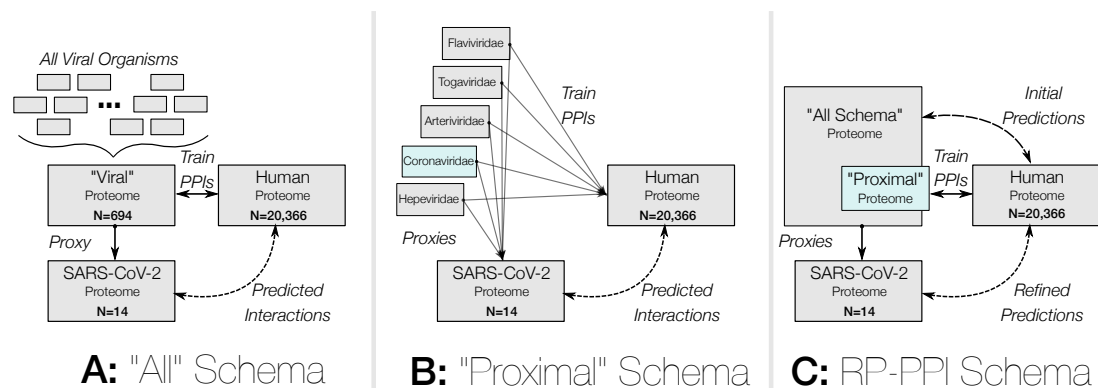



Figure 1. Overview of the Three Prediction Strategies to Generate the SARS-CoV-2 vs. Human Interactome. The three schemas depict how known PPIs are leveraged to train a prediction model to generate predictions for SARS-CoV-2.

102   Our sequence-based PPI prediction method (PIPE) was previously used during the 2015 Zika virus out-
103 break to identify putative human-Zika PPIs with the goal of informing rational drug discovery (Kazmirchuk
104 et al., 2017). In the present study, of the ∼285,000 host-virus pairs, we leverage three prediction schema
105 and two independent PPI predictors to select a highly conservative set of predicted interactions for each
106 of the 14 SARS-CoV-2 proteins considered in this study resulting in the identification of several putative
107 human protein targets. We have publicly released these predictions and related meta-data for use by the
108 broader scientific community in the following DataVerse repository: 10.5683/SP2/JZ77XA, (Dick et al.,
109 2020a).

METHODS

111 The multi-schema methodology leveraged in this work follows from and expands upon a previous study
112 of the Zika virus (Kazmirchuk et al., 2017) where we defined two initial prediction schemas from which
113 to train the PPI predictors. First, the *all* schema, contains the maximum available number of known
114 virus-human PPIs regardless of the evolutionary distance between those viruses and the target virus
115 (*i.e.* SARS-CoV-2). This schema groups all viruses into a "viral" collection to serve as a proxy for
116 SARS-CoV-2. The second schema, denoted *proximal*, is a subset of the *all* schema, where only the PPIs
117 from evolutionarily related viruses are considered. In a third schema, denoted *RP-PPI*, both the *all* and
118 *proximal* datasets are leveraged to apply the Reciprocal Perspective cascaded PPI predictor developed
119 by Dick and Green (2018). Specifically, the *proximal* PPIs are used to train the PIPE4 and SPRINT
120 method generating the comprehensive prediction matrix (CPM) representing all possible pairs between the
121 remaining *all* schema pairs and human. From this CPM, the RP features (as described in Dick and Green
122 (2018)) were extracted and used to train a downstream model to generate refined predictions between
123 SARS-CoV-2 and human protein pairs.

124   In the three schemas, as part of an independent evaluation, we remove the previously known SARS-
125 CoV Spike vs. ACE2 interaction to serve as a positive control among the set of predicted interactions. We
126 retained the other four known interactions between SARS-CoV and human within the PPI training set.

127   The dataset of experimentally elucidated human-virus PPIs was obtained from the VirusMentha
128 database (Calderone et al., 2015). These 10,693 known PPIs are used to train the PPI predictors and infer
129 new putative interactions between human proteins and the SARS-CoV-2 proteome. For the *all* schema, the
130 proteomes of the 43 viral families were collected from Uniprot and are summarized in the Supplementary
131 Materials. To generate a complimentary predicted interactome using the *proximal* schema, we tabulate
132 the 689 training PPI and the Group IV viral families over which they are distributed (Table 1). Finally, the
133 human reference proteome (UP000005640) was obtained from Uniprot, retaining only the high-quality
134 "Reviewed" Swiss-Prot proteins.

The SARS-CoV-2 Proteome

The proteome of SARS-CoV-2 was obtained from the Uniprot pre-release available at SARS-CoV-2 Pre-Release, (Swiss Institute of Bioinformatics, 2020), with the disclaimer that these data would become part of a future UniProt release and may be subject to further changes. While other SARS-CoV-2 proteins are reported among other sequence repositories, we restricted our study to these highest-confidence proteins available at the time. The 14 SARS-CoV-2 proteins and their function are tabulated in Table 2. Notably, the Spike glycoprotein (Accession: P0DTC2) is of special interest to this and related work, since its SARS-CoV homolog is known to interact with the human ACE2 protein and is presently the target of a recent mRNA-based vaccine candidate.

Computational Protein-Protein Interaction Predictors

The computational prediction of PPIs is a diverse field which encompasses multiple paradigms (*e.g.* sequence-, structure-, evolution-, and network-based methods). Sequence-based predictors rely solely upon primary sequence data, making them amenable to the investigation of proteome-wide networks. Furthermore, these methods tend to be highly efficient, where individual PPIs can be predicted in a fraction of a second.

The Protein-Protein Interaction Prediction Engine (PIPE4)

PIPE is a sequence-based method of PPI prediction that operates by examining sequence windows on each of the query proteins. If the pair of sequence windows shares significant similarity with a pair of proteins previously known to interact, then evidence for the putative PPI is increased. A similarity-weighted (SW) scoring function uses normalization to account for frequently occurring sequences, not related to PPIs. Given sufficient evidence, a PPI is predicted. PIPE has previously been validated on numerous species for both intra-species and inter-species PPI prediction tasks (Schoenrock et al., 2011; Pitre et al., 2006, 2012). Furthermore, the distribution of evidence along the length of each query protein forms a 2D landscape that can indicate the site of interaction (discussed later) (Amos-Binks et al., 2011).

The fourth version of the Protein-protein Interaction Prediction Engine (PIPE4) was recently adapted to improve predictive performance for understudied organisms (Dick et al., 2020b). That is, species for which the proteome is known, but the number of experimentally validated intra-specific PPIs is insufficient to train a model to generate the comprehensive interactome. To circumvent this, the PIPE4 algorithm leverages the known PPIs of evolutionarily similar and well-studied organisms, serving as a *proxy* training set. Using an approach denoted as *cross-species* PPI prediction, the experimentally validated PPIs from the proxy species are used to train the PPI predictor which is then applied to the proteome of the understudied target organism. Due to the limited availability of known SARS-CoV-2 PPIs, we here use the PPIs from a collection of well-studied and evolutionarily similar proxy viruses to generate these cross-species predictions as depicted in Figure 1.

The PIPE4 algorithm is particularly well-suited to cross- and inter-species PPI prediction schemas, given that the SW-scoring function appropriately normalizes the prevalence of sequence windows within each training and target species proteome (Dick et al., 2020b).

Table 1. Group IV Viral Families and their Number of PPIs used in the *Proximal* Prediction Schema.

Virus Family	Number of PPIs	Capsid Type	Capsid Symmetry	Nucleic Acid Type	Examples
Flaviviridae	569	Enveloped	Icosahedral	Single-Stranded	Hepatitis C virus, Zika virus
Togaviridae	56	Enveloped	Icosahedral	Single-Stranded	Rubella virus, Alphavirus
Arteriviridae	56	Enveloped	Icosahedral	Single-Stranded	Arterivirus
Coronaviridae	5	Enveloped	Helical	Single-Stranded	Coronavirus
Hepeviridae	3	Naked	Icosahedral	Single-Stranded	Hepatitis E virus
Astroviridae	0	Naked	Icosahedral	Single-Stranded	Astrovirus
Calciviridae	0	Naked	Icosahedral	Single-Stranded	Norwalk virus
Picornaviridae	0	Naked	Icosahedral	Single-Stranded	Enterovirus, Hepatovirus

Scoring PRotein INTeractions (SPRINT)

The SPRINT predictor is conceptually similar to PIPE; SPRINT aggregates evidence from previously known PPI interactions, depending on window similarity with the query protein pair, to inform its prediction scores (Li and Ilie, 2017). SPRINT leverages a *spaced seed* approach for determining protein window sequence similarity, where only specific positions in the two windows must be identical as defined by the bits of the spaced seeds. Furthermore, protein sequences are encoded using five bits per amino acid, enabling the use of highly efficient (SIMD) bitwise operations to rapidly compute protein window similarities and, thereby, score predictions (Li and Ilie, 2017). The present version of the SPRINT algorithm is not explicitly designed to handle inter- and cross-species prediction, nor to predict the specific subsequence site of interaction between a given pair of proteins. Nonetheless, it is among the only PPI predictors capable of predicting comprehensive interactomes in a timely manner and was demonstrated to outperform other PPI predictors, including the PIPE2 algorithm (Li and Ilie, 2017).

Determining an Appropriate Per-Protein Decision Threshold

For each of the 14 SARS-CoV-2 proteins, we predicted their interaction with each of the 20,366 human proteins resulting in 285,124 unique predictions, forming what we denote the comprehensive prediction matrix (CPM), using each of the two predictors considered. While each method, through a form of cross-validation, might determinate a highly-conservative *global* decision threshold, we know from our work in (Dick and Green, 2018) that such thresholds are sub-optimal. Furthermore, there are insufficient known PPI exemplars between human and SARS-CoV-2 from which to optimize such a threshold. Consequently, for the first time, we employ an RP-inspired method to adaptively determine *local* decision thresholds on a per-protein basis based on the distribution of prediction scores involving each protein.

From the prediction of all possible pairs, we obtain a CPM. We can then plot the rank-ordered distribution of the putative interaction scores involving each of the *individual* SARS-CoV-2 proteins separately in decreasing rank order by score, forming a *one-to-all* (O2A) score curve. This presents an opportunity to develop protein-specific local decision thresholds, where only those interactions scoring significantly above baseline are reported. These one-to-all score curves are based on the underlying assumption that we expect true SARS-CoV-2 vs. human PPIs to be rare, such that the vast majority of prediction scores should fall below the decision threshold. Furthermore, for the *RP-PPI* schema, we additionally examine the reciprocal perspective, examining one-to-all curves for each human protein and applying analogous decision logic to determine human-protein-specific decision thresholds. (Dick and Green, 2018).

Thus, for each O2A score curve, a score threshold delineating the "high-scoring" pairs from the baseline was identified and used to determine the high-confidence predicted interactions. In the absence of known PPIs between SARS-CoV-2 and human, it is difficult to determine a suitable global decision threshold. By instead examining the morphology of the O2A score curves for both perspectives, we can qualitatively identify high-scoring pairs. This process can be further automated through the identification of the baseline/knee for each view under the assumption that true PPIs are rare and high-scoring, while non-interacting pairs tend to generate scores residing below the knee in the baseline.

Table 2. The 14 Proteins in the SARS-CoV-2 Proteome Considered in this Study.

Uniprot Acc.	Gene Name	Protein Name	Protein Function
P0DTD1	R1A_WCPV	Replicase polyprotein 1a (R1a)	Viral transcription/replication
P0DTC1	R1AB_WCPV	Replicase polyprotein 1ab (R1ab)	Viral transcription/replication
P0DTC2	SPIKE_WCPV	Spike glycoprotein (S)	Attachement and entry
P0DTC3	AP3A_WCPV	Protein 3a (ORF3a)	ESCRT-independent budding
P0DTC4	VEMP_WCPV	Envelope small membrane protein (E)	ESCRT-independent budding
P0DTC5	VME1_WCPV	Membrane protein (M)	Virion morphegenesis
P0DTC6	NS6_WCPV	Non-structural protein 6 (ORF6)	Unknown; possibly host-virus modulation
P0DTC7	NS7A_WCPV	Protein 7a (ORF7a)	Unknown; possibly host-virus modulation
P0DTD8	NS7B_WCPV	Protein 7b (ORF7b)	Unknown; possibly host-virus modulation
P0DTC8	NS8_WCPV	Non-structural protein 8 (ORF8)	Unknown; possibly host-virus modulation
P0DTC9	NCAP_WCPV	Nucleoprotein (N)	Viral genome packaging
P0DTD3	Y14_WCPV	Uncharacterized protein 14 (ORF8)	Unknown; possibly host-virus modulation
P0DTD2	ORF9B_WCPV	Protein 9b (ORF9b)	Unknown; possibly host-virus modulation
A0A663DJA2	A0A663DJA2_9BETC	Hypothetical ORF10 protein	Presumably not expressed

We automated the selection of this operational decision threshold for the 14 SARS-CoV-2 proteins using the Kneedle algorithm (Satopaa et al., 2011), applied to its top-1000 predictions, using a sensitivity parameter of 2.0. An example visual illustration of the highly conservative selection of high-confidence interactions is depicted in Figure 2 and the cut-off scores for each protein are tabulated in the Supplementary Materials.

We identified the common set of predicted pairs above each locally defined knee from *both* the PIPE4 and SPRINT methods (their intersection) for each schema. For example, the *all* schema, resulted in a set of 225 putative human protein targets among 279 intersection pairs. The predicted pairs from each schema were considered to be the predicted interactome and were subsequently analyzed by PIPE-Sites; GO-term enrichment analysis was performed using the identified human proteins. The results of each schema's interactome were also combined into higher-confidence sets by taking their set intersections and were visualized as a network.

Predicting PPI Site of Interaction using PIPE-Sites & the New Similarity Weighted Landscape

The PIPE4 algorithm generates its prediction for a given pair of proteins based on a two-dimensional landscape of scores, where the score at location x, y, the number of sequence window similarity "hits", represents the weight of evidence from the xth and y^{th} subsequence of the human and SARS-CoV-2 proteins, respectively. The PIPE-Sites algorithm examines this landscape and deduces which subsequences from each protein are likely to correspond to the site of interaction (Amos-Binks et al., 2011). Such information can guide subsequent detailed investigations to determine the physical binding site which may form the target for novel interventions to disrupt the PPI.

The list of PPIs generated from both methods can be used to inform the design of anti-SARS-CoV-2 therapeutics by using peptide sequences from the predicted PPI site, which we refer to as the PPI-Site. We define the PPI-Site as the peptide sequence that is responsible for mediating a given PPI, which is here estimated using the PIPE-Sites method. A conceptual overview of the PIPE4 landscape matrix and PIPE-Site prediction is illustrated in the Supplementary Materials.

Additionally, we introduce for the first time the Similarity Weighted landscape which is derived from the original PIPE4 landscape with the following modification: the "hits" representing the weight of evidence from the xth and y^{th} subsequence of the human and SARS-CoV-2 proteins, respectively, are normalized by a cross-species variant of the SW normalization factor in (Dick et al., 2020b) which normalizes window frequency only in species for which there are available PPI training data. This suppresses the effect of highly prevalent windows that are not associated with interactions and amplifies the effect of windows that are relatively rare, yet are frequently occurring in known interactions within the proteomes of species for which training data are available. Specifically, the high-scoring "hot-spots" in the SW landscape are putative subsequences possibly mediating interactions between two proteins. For clarity, we syntactically distinguish the *interface residues* from a predicted *PPI-Site*. Since we are looking at sequence similarity across many proteins, the PPI-Site is a proxy for measuring sequence conservation. Therefore, we are identifying the subsequence that has been conserved to support the interaction site, which may include scaffolding residues distal to the actual interface.

The Reciprocal Perspective Cascaded Classifier: Combination of Multiple Experts

In previous work, we demonstrated that the use of the Reciprocal Perspective PPI cascaded classifier (RP-PPI) produced statistically significant improvement in performance (Dick and Green, 2018). Moreover, we here propose the RP-PPI method, as a cascaded machine learning algorithm, can be leveraged to combine features from multiple expert models. Here, for the first time, we jointly combine the features derived from the PIPE and SPRINT models and demonstrate the resulting improvement in performance as part of the *RP-PPI* schema. Furthermore, following from the work of Kyrollos et al. (2020), we implement the cascaded model as an eXtreme Gradient Boosting (XGBoost) regression model (Chen and Guestrin, 2016) instead of the Random Forest classifier originally proposed in (Dick and Green, 2018).

To evaluate the performance increase of the combined classifier, we perform Leave-One-Family-Out cross-validation (LOFOCV), and plot the average Receiver Operating Characteristic (ROC) curve with confidence intervals of one standard deviation. Given certain families had relatively few PPIs, we omitted those with fewer than 50 PPIs from this analysis (a negligible number of pairs were left out). The determination that the combined use of PIPE4 and SPRINT features from their respectively predicted CPMs does, in fact, result in improved performance. We then performed extensive hyper-parameter tuning, evaluated via 10-fold cross-validation, to obtain the most performant model to generate our

264 SARS-CoV-2 vs. human predictions. Varying maximum tree depth ($[3, 4, 5, ..., 18]$), number of estimators
265 ($[50, 75, 100, ..., 600]$), and the learning rate (9 values considered), we trained and evaluated 29,700 models
266 to arrive to the final model that was used to generate the comprehensive set of prediction as part of the
267 *RP-PPI* schema.

Gene Ontology (GO) Enrichment Analysis

269 To determine which human cellular pathways may be targeted by SARS-CoV-2, PANTHER Gene
270 Ontology (GO) Slim enrichment analysis was applied to each of the predicted interactomes from each
271 schema independently: the 225 human proteins predicted to interact with SARS-COV-2 proteins in the
272 *all* schema, the 123 human proteins in the *proximal* schema, and the 494 human proteins in the *RP-PPI*
273 schema. The molecular function, biological pathway, and cellular pathway *p*-values were determined
274 with the Fisher's Exact test implemented in the PANTHER GO software (Mi et al., 2019). The *p*-values
275 were corrected for multiple testing using the False Discovery Rate (FDR) method described in Mi et al.
276 (2019) and significant terms were identified at a threshold of 0.05 and ordered terms by fold enrichment
277 applying variable thresholds.

RESULTS & DISCUSSION

279 It is of critical importance that the global research community focus its efforts on the rapid understanding
280 the SARS-CoV-2 virus and the pathogenesis of COVID-19 in order to develop anti-viral therapeutics
281 and additional vaccine targets. Research into *Coronaviridae* biology sharply declined post-SARS-CoV
282 and it is the hope of this work to compliment subsequent primary research in both short-term therapeutic
283 development and long-term COVID-19 symptomology. Fortunately, the prior decades of research into
284 related viral families provide a wealth of data with which to guide current and future studies. For
285 example, the elucidation of the SARS-CoV vs. human inter-species interactome in 2011 using the
286 high-throughput (though false positive-prone) yeast-two hybrid method highlighted cyclophilins as a
287 target for pan-coronavirus inhibitors (Pfefferle et al., 2011). Previous knowledge of related coronaviruses
288 within the *Coronaviridae* family provide training samples from which we can identify a number of new
289 high-confidence PPIs that contribute to our understanding of the COVID-19 disease pathogenesis and
290 which may represent targets for novel inhibitory therapeutics.

291 It is known that the SARS-CoV Spike protein binds to the human ACE2 receptor (Glowacka et al.,
292 2010). Upon entry into the respiratory or gastrointestinal tracts, coronaviruses establish themselves
293 by entering and infecting lumenal macrophages and epithelial cells. The viral cell entry program is
294 orchestrated by the spike protein that binds to the human cellular receptors and, thereby, mediates
295 virus-cell membrane fusions.

296 While the putative interaction between the SARS-CoV-2 Spike protein and human ACE2 receptor is a
297 current focus of the research community, it is also valuable to develop a more holistic understanding of
298 the possibly numerous SARS-CoV-2 vs. human PPIs. Consequently, additional viral-human interactions
299 might be targeted and disrupted with the use of small inhibitory peptides or molecules. To this end,
300 we leveraged sequence-based predictors to score all possible interactions between the SARS-CoV-2
301 and human proteomes. To identify our highest-confidence set of predictions for the SARS-CoV-2 vs.
302 human interactome, we prepared three training, prediction, and evaluation schemas and combined their
303 predictions to produce a set of candidate interactions for wet-lab validation and the potential design of
304 inhibitory peptides.

Predictions from the *All* and *Proximal* Schemas

306 As part of the first two schemas (*all* and *proximal*), for each of the 14 viral proteins, we sorted the
307 20,366 scores (for each human protein) into a monotonically decreasing rank-order which enabled the
308 identification of the subset of high-scoring putative interactors with each viral protein. An example from
309 the *all* schema is depicted in Figure 2A,D.

310 Rather than apply a globally defined decision threshold (*i.e.* top-*k* or minimum threshold), we
311 automatically detected a highly conservative "knee" for each curve (the point of greatest rate of change
312 parameterized by a sensitivity value) to delineate those rare high-scoring pairs from the remaining baseline
313 (Figure 2B,E). For example, within the *all* schema, the union of the $n = 1,209$ predicted PIPE4 and
314 SPRINT high-confidence putative PPIs comprises only $\sim 0.42\%$ of all possible pairs, and their intersection
315 of $n = 279$ putative pairs comprises a highly conservative subset $< 0.098\%$. These data are tabulated

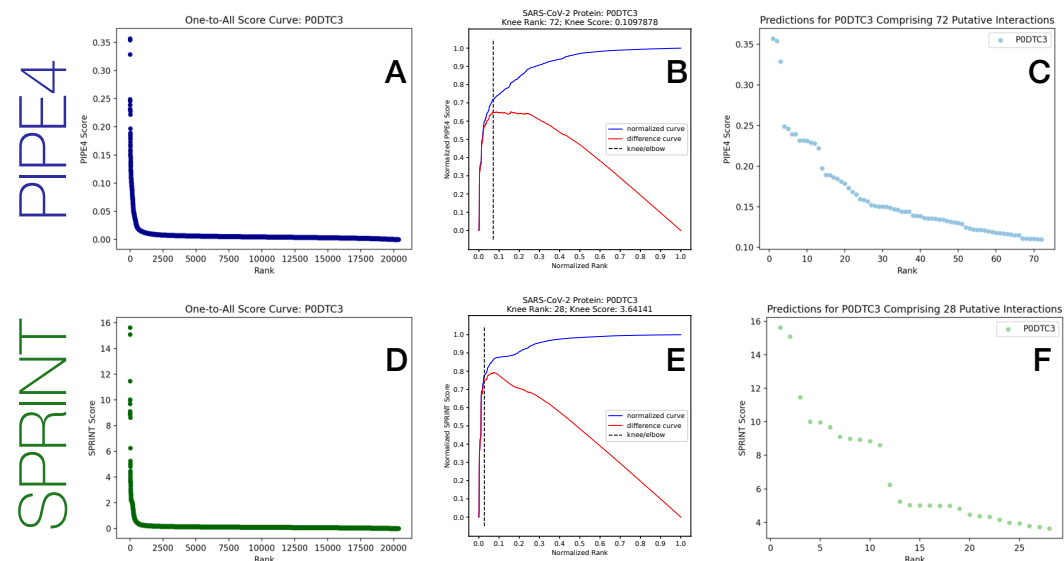



Figure 2. Example Compilation of the Spike Protein One-to-All Score Curve, Knee Detection for Local Cut-Off, and Rank Order Predictions, for each Method. Panels A & D depict the one-to-all score curves from the predicted score between the Spike protein and all proteins in the human proteome. Panels B & E depict the detected knee from the top-1000 scores in each one-to-all score curve where the dashed line represents the knee detected using the Kneedle algorithm. Panels C & F depict the predicted interactions above the knee.

316 in the Supplementary Table, plotted in Figure 2, and illustrated in Figure 3A. Taking the combinatorial
317 intersection of the high-confidence predictions from each schema resulted in the highest confidence set of
318 predictions with $n = 7$ predicted pairs (Figure 3).

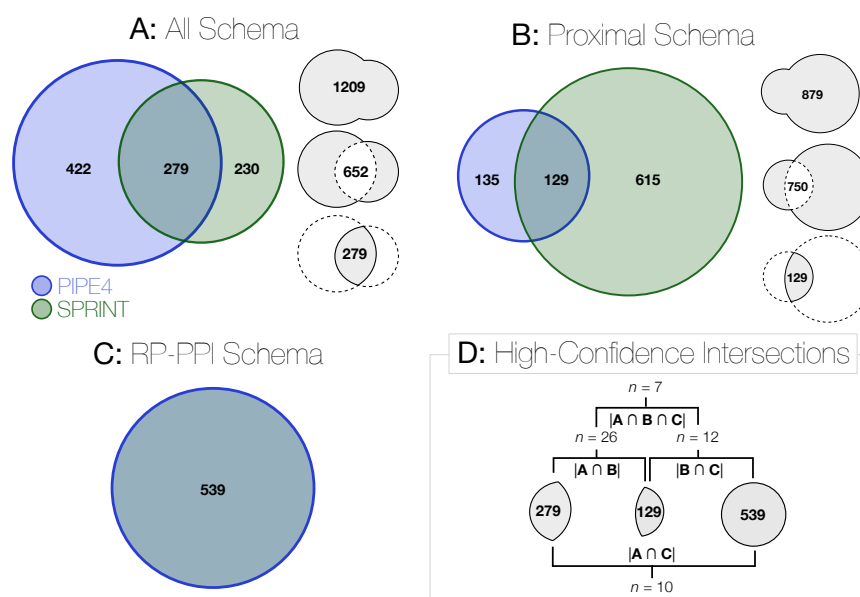



Figure 3. Venn Diagram of the Human Proteins Predicted to Interact with SARS-CoV-2 Proteins. Panels (A), (B), and (C) depict the number of predicted pairs for each of the schema's putative interactomes. In panel (D), those interactomes are combined further by taking their intersections with the highest confidence subset comprising $n = 7$ pairs.

Predictions from the *RP-PPI* Schema

Following from the experimental design of the *all* and *proximal* schemas, the independent predictions from the RP-PIPE4 model and the RP-SPRINT models would have been combined into a single intersection set. However, for the first time, we jointly combined the RP features derived from the PIPE4 O2As with those derived from the SPRINT O2As to train and evaluate a "combination of multiple experts" RP-PPI model. The joint model (using default hyperparameter settings) demonstrated an improvement over the RP-predictor model alone. Interestingly, as illustrated in Supplementary Figure 3 the improvement does not appear to be symmetric: the improvement of performance when SPRINT features are joined with the PIPE4 features (Supplementary Figure 3: A, blue & grey) is greater than when the PIPE4 features are joined with SPRINT features (Supplementary Figure 3: B, blue & grey).

Having established that the combination of experts RP-PPI approach produces improved models, we performed extensive hyperparameter tuning to determine model parameters (550 estimators, maximum tree depth of 17, learning rate of 0.1). Each experiment was evaluated via 10-fold cross-validation with performance measured using the F1 score. Following the training and evaluation of 29,700 models, we identified the best performing model parameters as having a learning rate of 0.1, a maximum tree-depth of 17, and 550 estimators (Supplementary Figure 4).

To better understand the features focused upon by the RP-PPI model, we plot the relative feature importance, measured by average information gain in Supplementary Figure 5. Many of the original features from the work of (Dick and Green, 2018) are leveraged in addition to new "statistics-type" features where a given pairs' score is measured in standard deviations away from the identified baseline of a given one-to-all score curve. Notably, baseline scores and ranks for Element A (the SARS-CoV-2 protein) of both methods are among the most distinguishing features (top-4).

With the RP-PPI model, the comprehensive set of human–SARS-CoV-2 pairs were scored to produce 14 one-to-all curves. As above, knee-detection was used to identify the highest confidence subset comprising $n = 539$ pairs, as depicted in Figure 3C.

We provide the hit and SW landscapes and predicted PIPE-Sites for each of the predicted interactions for each SARS-CoV-2 proteins of each schema. We highlight those 279 pairs within the predicted *all* interactome, the 129 pairs within the predicted *proximal* schema, and the 539 pairs within the predicted *RP-PPI* schema. All data are published in the following DataVerse repository, for broader use by the scientific community (Dick et al., 2020a).

We later discuss the biological relevance of our set of highest-confidence predictions and how these may be leveraged to develop anti-SARS-CoV-2 therapeutics. We further consider these interactions in the context of corroborating evidence from scientific literature and illustrate two particular phases of the viral life cylce that might be targeted.

Putative Interaction to Target with anti-SARS-CoV-2 Therapeutics

The genomes of SARS-CoV-2 and other coronaviruses encode for numerous proteins of diverse functions. The proteolytic cleavage products of the two polyproteins (*i.e.* non-structural proteins) play essential roles in viral replication but also participate in viral pathogenesis. Similarly, though the structural proteins (*e.g.* S, E, M, and N) are inherently involved in viral structure and virus-host interactions, such proteins further pathogenesis through interaction with numerous proteins within signaling pathways and, further, accessory proteins are not essential for viral replication; such proteins differ greatly between coronavirus species (Narayanan et al., 2008).

To guide the broader research community in expanding the basic understanding of the involvement of SARS-CoV-2 proteins in the underlying pathogenesis of COVID-19, we have visualized the predicted interactomes and incorporated relevant biological information into these networks (Figure 4). Based on biological process Gene Ontology (GO)-terms for each protein within the intersections of the three schemas, individual proteins were manually curated into single descriptors of biological roles, such as those describing viral entry modes, vesicular transport and related processes, types of immune responses, and signaling pathways related to immune responses. Associating proteins with single descriptors is not ideal as numerous proteins possess a broad range of functions. We therefore encourage investigators to assess biological functions of the predicted interactors on an individual basis. In Figure 5 we illustrate the life cycle of CoVs and highlight the mode of future peptide inhibitors potentially derived from this work.

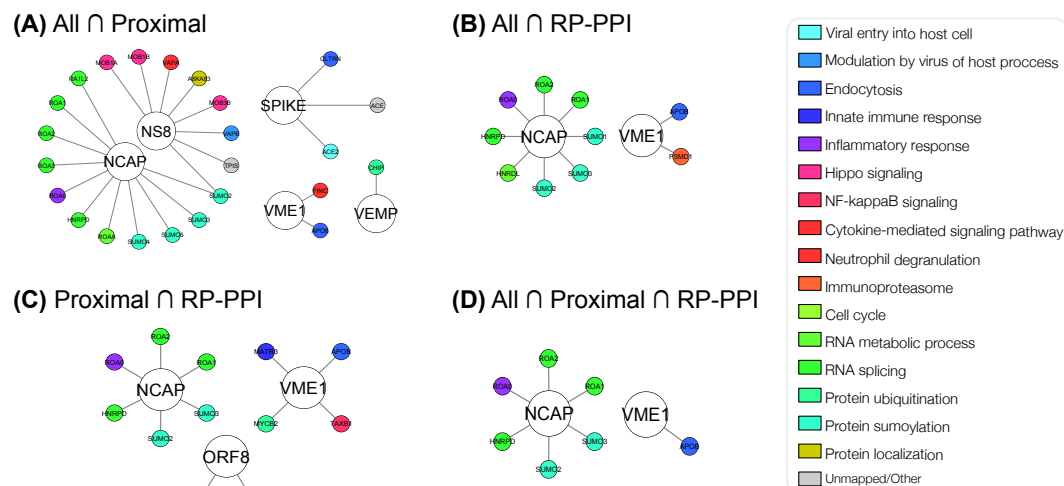



Figure 4. Network Visualization of the Highest-Confidence Predictions. The colour-function relationship is depicted in Figure 11. Created using Cytoscape (Shannon et al., 2003).

The Spike Protein vs. ACE2 Interaction

The PIPE4 and SPRINT predictors scored the SARS-CoV-2 Spike protein vs. human ACE2 protein as the top-ranking prediction in their respective one-to-all score curves (P0DTC2-Q9BYF1) (PIPE4 SW score of 2.159, SPRINT score of 29.3515) within the *all* schema and relatively high-scoring within the other two schemas. As previously noted, this was achieved despite the removal of the known SARS-CoV Spike-ACE2 PPI within the training dataset as part of an independent experiment to determine whether or not the SARS-CoV Spike-ACE2 PPI would have a large effect on scoring this prediction. We further visualize the putative subsequence region of interaction between these proteins in Figure 6.

Multiple interactions among the high-confidence predictions are biologically relevant when considering known human coronavirus biology. Most notably, predicted within the intersection of the *All* and *Proximal* schemas was the Spike-ACE2 interaction. It is now well-established that SARS-CoV-2 utilizes ACE2 as the main fusion receptor for cell entry (Hoffmann et al., 2020). Furthermore, the Spike proteins of SARS-CoV and human coronavirus NL63 interact with ACE2 to facilitate cell entry (Li et al., 2003; Hofmann et al., 2005). Although the Spike-ACE2 interaction was excluded from our training data, our computational methodology independently predicted the main PPI that permits SARS-CoV-2 cell entry; therefore this, and similar computational pipelines, hold promise for screening candidate modes of entry of future viruses. Thus, other interactions within these high-confidence intersections of the schemas may be biologically relevant and worthy of further investigation. Besides the Spike-ACE2 interaction, two other high-confidence interactors of Spike were found to be collectrin (CLTRN) and ACE. CLTRN is a homolog of ACE2, lacking the extracellular catalytic domain, sharing 47.8% identity with C-terminal regions of ACE2 (Zhang et al., 2001). Furthermore, ACE is homologous to ACE2, sharing 42% identity between catalytic domains (Donoghue et al., 2000). This finding corroborates related research reporting that SARS-CoV-2 can infect human respiratory epithelial cells through interaction with the human ACE2 receptor (Letko et al., 2020).

Certainly, if the SARS-CoV-2 and SARS-CoV Spike proteins share sufficient sequence and structural similarity, it can be expected that anti-virals designed against SARS-CoV may also be effective against SARS-CoV-2. We investigate this sequence similarity by performing a BLASTp alignment of the two sequences. Interestingly, only 76% identity was observed (Figure 7) suggesting that the SARS-CoV-2 spike protein might have evolved to be sufficiently different from its SARS-CoV variant to render existing anti-virals ineffective. Given that the Spike protein is the main point of interface with the host, we can expect that it would be rapidly evolving. The SARS-CoV-2 variant likely shares a similar mechanism of action where the recombinant SARS-CoV-2 spike protein downregulates ACE2 expression and thereby promotes lung injury (Glowacka et al., 2010).

Consequently, the elucidation of the Spike-ACE2 binding interface is needed to design novel thera-

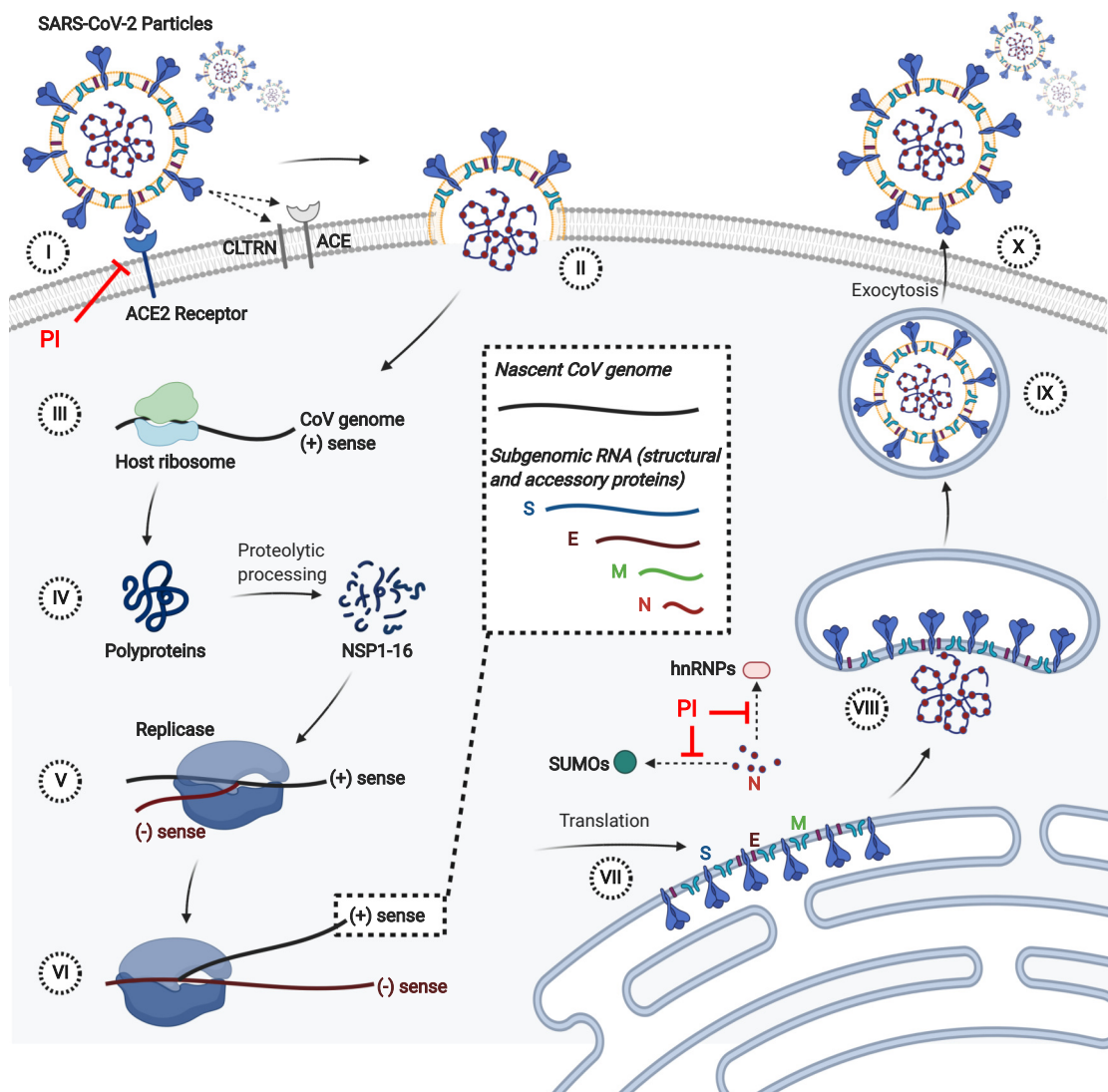



Figure 5. Life Cycle of CoVs and the Mode of Future Peptide Inhibitors (PI) Derived from this Study. (I) SARS-CoV-2 attaches to the cell surface via interaction of the spike (S) protein with the host ACE2 receptor. (II) The CoV and host membranes coalesce either at the cell surface or within endosomes, releasing the CoV genome into the cytoplasm. (III) Host ribosomes use the CoV genome as a template and translate polyproteins 1a and 1ab. (IV) The polyproteins mature into individual non-structural proteins (NSPs) 1-16 via autoproteolytic processing. (V) Multiple NSPs form a viral replicase complex, which performs negative-strand synthesis of genomeic and subgenomic RNA negative-strand templates. (VI) The viral replicase synthesizes nascent plus-strands of the full-length CoV genome and subgenomic RNAs encoding structural (S, E, M, N) and accessory proteins (not shown). (VII) S, Membrane (M), and Envelope (E) proteins are translated at the endoplasmic reticulum (ER) and inserted into the ER membrane. The Nucleocapsid (N) protein is translated within the cytoplasm. (VIII) The N protein encapsulates the nascent CoV genome and interacts with the other structural proteins within the ER-Golgi intermediate compartment (ERGIC). (IX) Mature CoV particles are formed within vesicles upon budding into the lumen of the ERGIC. (X) CoV particles are released upon exocytosis. Besides the validated S-ACE2 interaction, other notable predicted protein-protein interactions are indicated by dashed arrows. This figure was made in ©BioRender - biorender.com.

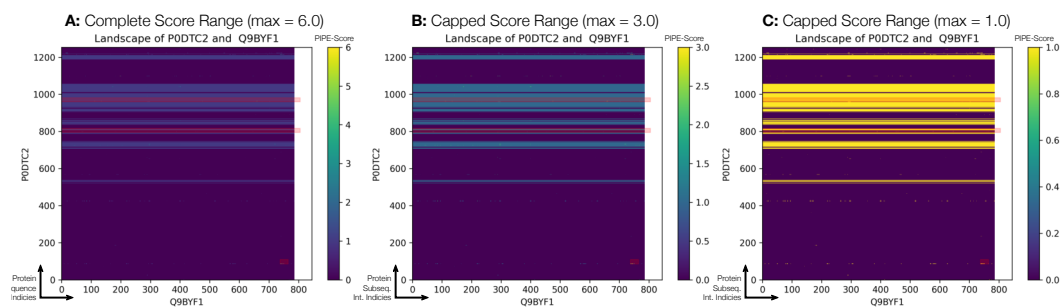



Figure 6. The PIPE-Sites Landscape between the SARS-CoV-2 Spike protein and human ACE2 protein. Within each panel, the three red rectangles represent the predicted PIPE-Sites regions. Panel A depicts the completely predicted landscape with the complete numerical range of scores depicted. To more easily visualize the high-scoring subsequence regions, panels B & C apply a numerical "capped threshold" where any value greater than or equal to the maximum threshold is set to that value. This has the effect of emphasizing the regions of potential interest. A threshold of 3.0 is applied in panel B and a threshold of 1.0 in panel C. See the supplementary material for guidance on the interpretation of these landscapes.

405 peutics. To that end, we used the PIPE-Sites algorithm to predict the three most likely putative interaction
406 interfaces between the Spike (P0DTC2) and ACE2 (Q9BYF1) proteins (Figure 6). Note that all predicted
407 subsequence offsets are 0-indexed. With a maximum landscape peak of 6, the PIPE-Sites algorithm
408 identified three putative interaction interfaces:

409 1. **P0DTC2**: [86–109]; **Q9BYF1**: [738–816]
410 2. **P0DTC2**: [795–816]; **Q9BYF1**: Entire sequence
411 3. **P0DTC2**: [960–981]; **Q9BYF1**: Entire sequence

412 Interestingly, the PIPE-Sites score landscape in Figure 6 exhibits a number of horizontal bands
413 indicative of subsequence regions along the Spike protein that correspond to a relatively high likelihood
414 of interaction. While the PIPE-Sites algorithm only identifies three putative regions, these bands suggest
415 additional regions of interest.

416 The highest-scoring predicted PIPE-Site interface corresponds to the Spike [86–109] subsequence
417 and the ACE2 [738–816] subsequence, which resides within the *intracellular* cytoplasmic domain of
418 ACE2. However, upon closer inspection of other "hot spot" regions within the landscape, we note several
419 that reside within the *extracellular* N-terminal region of ACE2 (*i.e.* residues ∼[30-84] & [353-357]). In
420 particular, we note the three following regions of interest:

421 *Visually high-scoring region:*   **P0DTC2**: near residue 1224; **Q9BYF1**: [15–23]
422 *Within ACE2 residues [30-84]:*   **P0DTC2**: near residue 420; **Q9BYF1**: [80-84]
423 *Within ACE2 residues [353-537]:*   **P0DTC2**: near residue 420; **Q9BYF1**: [355-357]

424 Most interestingly, certain of these region along the Spike protein appears to coincide with mismatched
425 or gap regions along the dot plot comparing the SARS-CoV vs. SARS-CoV-2 alignment depicted in
426 Figure 7). For example, upon closer investigation of the alignment around residue 420, we note six
427 mismatches. Their proximity to a candidate region of interaction certainly warrant additional experimental
428 investigation (Figure 7).

429 While numerous inhibitory strategies exist, including the use of small molecules or small interfering
430 RNAs, this research is most directly amenable to the design of small inhibitory peptides that inhibit virus
431 infection by preventing Spike protein-mediated receptor binding and blocking viral fusion and entry
432 (Figure 5). Unfortunately, much like small peptides and interfering RNAs, peptide-based solutions are
433 disadvantaged by their *low antiviral potency*.

434 ### HLA Class I/II Histocompatibility Antigen
435 Among the 225 human proteins identified in the *all* schema, six Human Leukocyte Antigen (HLA) class
436 I/II histocompatibility antigens were predicted to interact with P0DTC3, the SARS-CoV-2 Protein 3a
437 (ORF3a):

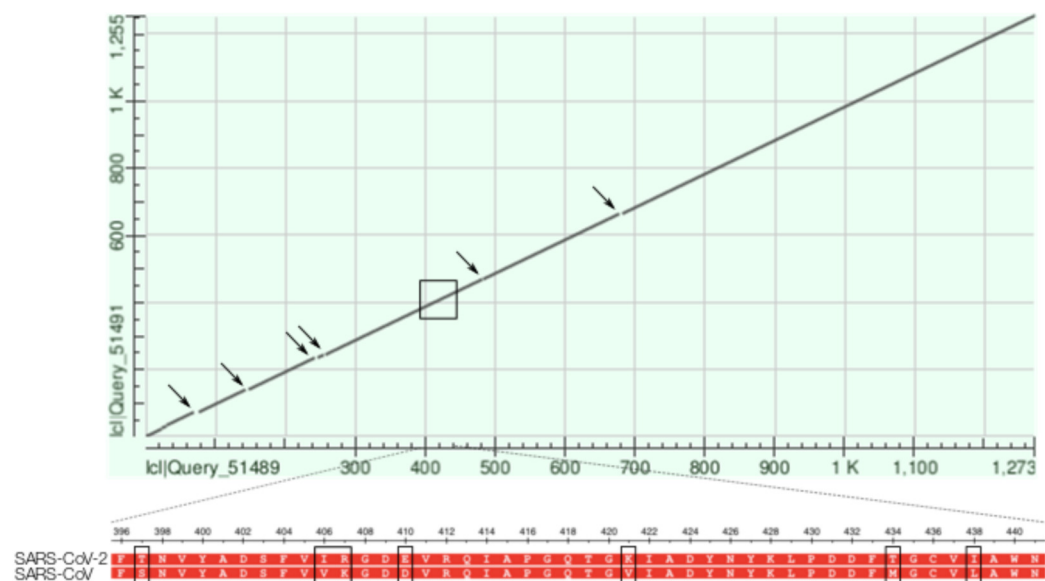



Figure 7. Dot Plot of the BLASTp Alignment of the SARS-CoV and SARS-CoV-2 Spike Protein.
The alignment of the two proteins results in a *max score* of 2039, a *total score* of 2039, 100% coverage, an *E-value* of 0.0, and 76.04% identity. Specifically: 971/1277 (76%) identities, 1109/1277 (86%) positives, and 26/1277 (2%) gaps. Arrows indicate gaps within the alignment and the zoomed-in region highlights the six mismatches around residue 420.

438 • **P13747**: HLA-E HLA-6.2 HLAE
439 • **P01911**: HLA-DRB1
440 • **P17693**: HLA-G HLA-6.0 HLAG
441 • **P04439**: HLA-A HLAA
442 • **P10321**: HLA-C HLAC
443 • **P30511**: HLA-F HLA-5.4 HLAF

444   The visualization of the predicted site of interaction for the six HLA interactions highlight a consistent
445 subsequence region of the SARS-CoV-2 protein 3a between amino acids [202–222] (Figure 8). Literature
446 review reveals that one of the open reading frames (ORFs) of the SARS-CoV virus, the ORF3a, encodes
447 the variant 274 AA-long Protein 3a. A previous study used sequence analysis that suggested that the
448 ORF3a aligned to a calcium pump present in *Plasmodium falciparum* and glutamine synthetase found in
449 *Leptospira interrogans*. This sequence similarity between the three organisms was found to be limited
450 only to amino acid residues [209–264], which form the cytoplasmic domain of ORF3a. This subsequence
451 region was predicted to be involved in calcium binding and then confirmed *in vitro* (Minakshi et al., 2014).
452   Given the important role that calcium plays as part of virion structure formation, virus entry, viral gene
453 expression, virion maturation, and release, these regions of Protein 3a are of possible interest for disruption
454 of SARS-CoV-2. Specifically, the design of a small inhibitory peptide targeting this subsequence region
455 of Protein 3a might disrupt the viral life cycle.

Heterogeneous Nuclear Ribonuclear Proteins (hnRNPs)

456
457 The Nucleocapsid (abbreviated N or NCAP) protein was predicted to interact with four heterogeneous
458 nuclear ribonuclear proteins (hnRNPs) within the intersection of the three schemas. PPIs between the
459 NCAP protein and those involved in RNA related processes are not surprising, especially considering that
460 NCAP protein of coronaviruses plays a role in viral RNA genome packaging as it is capable of binding
461 single-stranded RNA (Huang et al., 2004). Notably, the N protein of SARS-CoV-2 was predicted to
462 interact with hnRNP A1 (*i.e.* ROA1). This interaction has previously been validated in the context of
463 SARS-CoV NCAP protein and was found to be a high affinity interaction (Luo et al., 2005). Additionally,
464 this physical interaction is also inherent to a mouse coronavirus species (i.e., mouse hepatitis virus, MHV)
465 (Wang and Zhang, 1999). The role of hnRNP A1 as a host cell factor in MHV coronavirus biology is

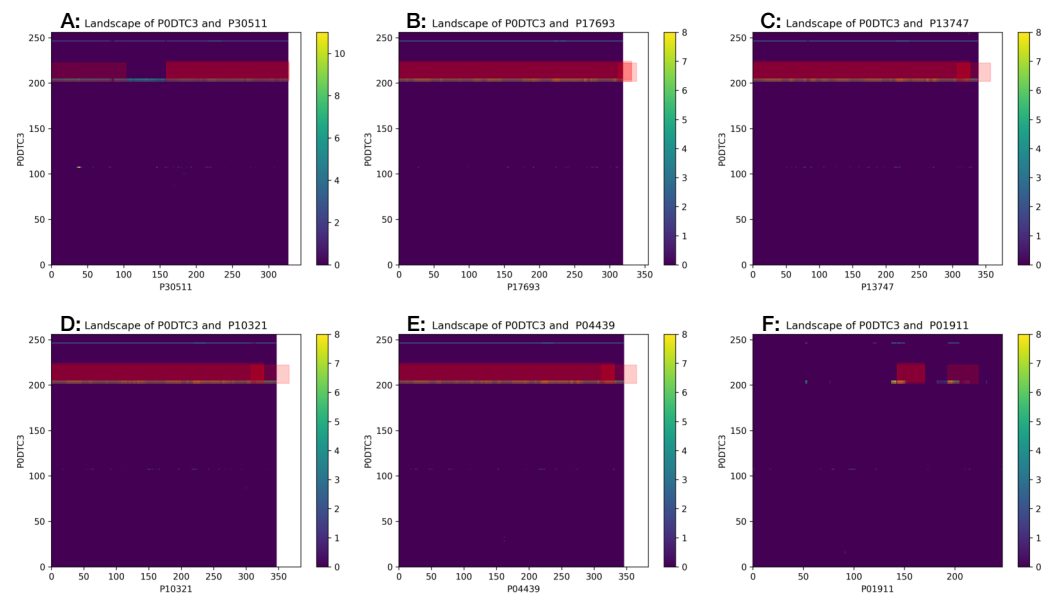



Figure 8. Landscapes of the Six Predicted HLA interactors with SARS-CoV-2 Protein 3a. The three red rectangles represent the predicted PIPE-Sites regions. They're "shifted" relative to the highlighted cells due to the algorithm's use of a window of 20 amino acids in length that extends both to the left (along the x-axis) and upwards (along the y-axis). This implementation may also result in the predicted site extending past the coloured matrix, either to the right or above. The PIPE-Sites may overlap when numerous hits appear within close proximity, as is the case when a "band" of hits appears in the matrix. See the supplementary material for guidance on the interpretation of these landscapes.

466   not clear, as initially it was shown that this protein functions in MHV RNA synthesis in the cytoplasm,
467   however involvement in these roles (e.g., RNA genome replication and discontinuous transcription)
468   were later contradicted (Shi et al., 2000; Shen and Masters, 2001). Furthermore, multiple hnRNPs were
469   shown to be upregulated in SARS-CoV infected cells (Jiang et al., 2005). The function of hnRNPs in
470   SARS-CoV-2 pathogenesis may relate to previous findings that suggest a role for such proteins in viral
471   RNA synthesis, and therefore these NCAP-hnRNP interactions may present as druggable targets. To that
472   end, Figure 9 illustrates both the hit and SW landscapes which may serve to identify putative subsequences
473   that may mediate the virus-host interactions.

Small Ubiquitin-related Modifier (SUMO) Proteins

475   Lastly, several small ubiquitin-related modifier (SUMO) proteins were predicted interactors within the
476   high confidence intersections. Notably, SUMO proteins were primarily predicted to interact with the
477   NCAP protein. Although interactions are contextually different than post-translational modifications, the
478   SARS-CoV NCAP was previously shown to both interact with an E2 enzyme involved in SUMOylation
479   as well as undergo SUMOylation at the lysine-62 residue (Fan et al., 2006; Li et al., 2005). SUMOylation
480   of NCAP at lysine-62 promotes homo-oligomerization and this residue may be involved in the disruption
481   of host cell division (Li et al., 2005). Whether this SUMOylation occurs within the context of SARS-
482   CoV-2 and the significance of this remains to be explored; however, based on the previous findings, this
483   interaction should be further investigated.

GO-Term Analysis of Human Proteins Among the Predicted Interactomes

485   For each of the schemas, the human proteins within the respective intersections of the PIPE4 and SPRINT
486   predicted interactions were used to run a number of GO-term analyses to better understand the functional
487   role of the human proteins involved. To this end, the GO-Slim Panther Classification System was used
488   to run over/under-representation analysis of the predicted sets of human proteins as compared to the
489   reference human proteome. A Fisher's Exact test with correction for False Discovery Rate was used to
490   extract a list of the most enriched GO-terms among the human proteins for which GO-term data were

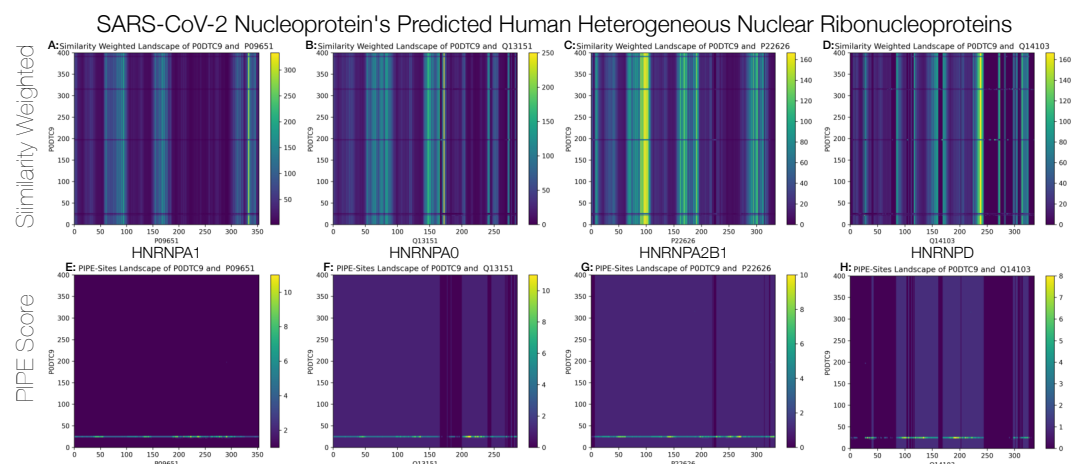



Figure 9. Hit and SW Landscapes for the Four Predicted hnRNPs to Guide the Design of Peptide Inhibitors. Both "hotspots" and "bands" identify subsequence regions of interest to target with peptide inhibitors. See the supplementary material for guidance on the interpretation of these landscapes.

⁴⁹¹ available. To limit the number of functions, variable thresholds for fold enrichment were applied. For ⁴⁹² example, among the tables for the *all* schema, the Molecular Functions exhibiting a fold enrichment ⁴⁹³ greater than 3 are reported; the Biological Processes exhibiting a fold enrichment greater than 50 are ⁴⁹⁴ reported; and the Cellular Components exhibiting a fold enrichment greater than 15 are reported. The fold ⁴⁹⁵ enrichment cut-offs were selected to limit the size of the tables; the complete tables are available in the ⁴⁹⁶ appendix of the Supplementary Materials and at public repository, Dick et al. (2020a).

⁴⁹⁷ While this current analysis combines all predicted human interactors together, a more revealing ⁴⁹⁸ analysis might investigate the resultant GO-terms on a per-viral-protein basis to identify those human ⁴⁹⁹ pathways and biological processes most sensibly targeted by SARS-CoV-2. This analysis is left to future ⁵⁰⁰ work.

⁵⁰¹ We encourage the scientific community to delve into the findings of this study. For example, of ⁵⁰² the GO-terms observed from the *all* schema alone, the highly over-represented biological processes ⁵⁰³ in Supplementary Materials Table 4 are interesting. Notably, the top-9 GO-terms have a 96.98 fold ⁵⁰⁴ enrichment given that the predicted set of human interactors contain all of the proteins from the *H. sapiens* ⁵⁰⁵ reference (*i.e.* the number of proteins present in the reference are also in the sample: 2/2, 8/8, and 3/3 ⁵⁰⁶ among the top-3, respectively). We specifically highlight the "antigen processing and presentation of ⁵⁰⁷ exogenous peptide antigen via MHC class Ib" (GO:0002477) and the "calcium ion transport from cytosol ⁵⁰⁸ to endoplasmic reticulum" (GO:1903515). Moreover, the top-ranking cellular component GO-terms ⁵⁰⁹ (Supplementary Materials Table 5) show notable over-representation of "MHC class Ib protein complex" ⁵¹⁰ (GO:0032398), "MHC class I protein complex" (GO:0042612), and numerous proteasome complex ⁵¹¹ terms. While only a shallow analysis is presented here, a more involved investigation into these predicted ⁵¹² interactions promises to reveal putative targets for novel inhibitory peptides.

A Literature Curated Subset of Candidate Human Protein Targets

⁵¹⁴ In the work of Gordon et al. (2020), $n = 332$ pairs between the $m = 26$ SARS-CoV-2 proteins and $m = 332$ ⁵¹⁵ human proteins (*i.e.* each human protein was involved in exactly one pair) were uncovered. While our ⁵¹⁶ highest-confidence interactomes do not predict any of these pairs (*i.e.* there is, unfortunately, no overlap ⁵¹⁷ between the $n = 322$ pairs in the work of Gordon et al. (2020) and the $n = 907$ union-of-three-interactome ⁵¹⁸ pairs, $A \cup B \cup C$), we emphasize that the work of Gordon et al. (2020) considered $m = 26$ proteins of the ⁵¹⁹ SARS-CoV-2 proteome while this work comprises only an $m = 14$ subset of those proteins.

⁵²⁰ As the $m = 332$ human proteins identified in the work of Gordon et al. (2020) are of putative interest ⁵²¹ to the scientific community in an effort to counter the COVID19 pandemic, for convenience, from the ⁵²² union of the three high-confidence interactomes (*i.e.* $A \cup B \cup C$, notation from Figure 3D) we extracted any ⁵²³ predicted interactions involving one of the $m = 332$ human proteins resulting in the $m = 14$ putative pairs. ⁵²⁴ This set of protein pairs may represent a focused subset of candidate pairs for subsequent investigation.

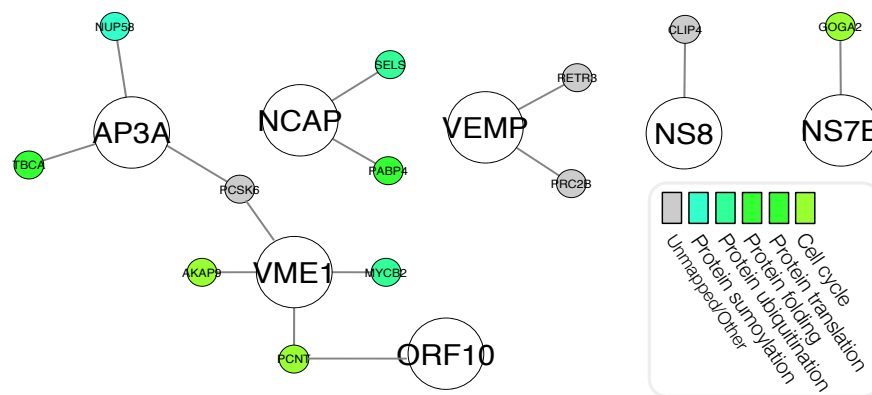



Figure 10. **Network Visualization of the 14 Predicted Pairs involving one of the 332 Human Proteins from Gordon et al. (2020).** Created using Cytoscape (Shannon et al., 2003).

This small network is depicted in Figure 10.

Finally, even if no overlap between Gordon *et al.* and our predicted interactomes exists, there is, indeed, an overlap of the GO-terms represented by the proteins of both sets. Notably, we identify 220 shared GO-terms between our sets indicative of a large functional overlap including such terms as viral process (GO:0016032), vesicle-mediated transport (GO:0016192), and response to virus (GO:0009615).

Complete Predicted Interactomes

To better visualize the predicted interactome the complete network-based representation is depicted in Figure 11. Much like the HLA proteins highlighted above, we note a number of highly represented GO-terms around several of the proteins of interest including those related to the immune response, various types of signalling, and the viral life cycle. We hope that this work will guide the broader research community in their search for putative inhibitory molecules.

CONCLUSIONS

The purpose of this work is to help guide the broader research community in the collective pursuit to understand the SARS-CoV-2 viral pathogenesis. To that end, we assessed 285,124 protein pairs using two state-of-the-art sequence-based PPI predictors within three prediction schemas, thereby creating the comprehensive SARS-CoV-2 vs. human interactome. For each of the 14 SARS-CoV-2 proteins considered in this study, a highly conservative locally defined decision threshold was determined to obtain a predicted interactome comprising putative PPIs within the predicted intersection of the PIPE4 and SPRINT methods. Furthermore, the PIPE-Sites algorithm was used to predict the putative interaction interfaces to identify the subsequence regions of interest that might mediate these interactions.

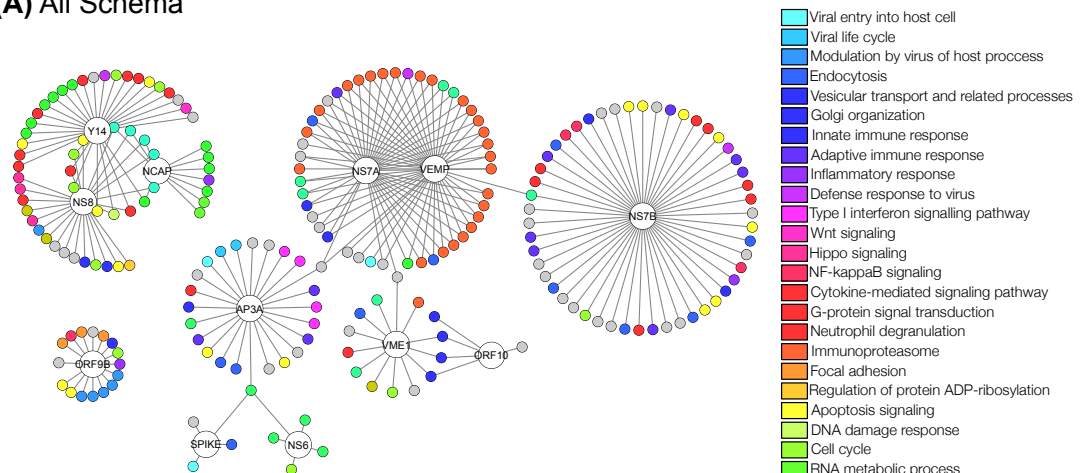
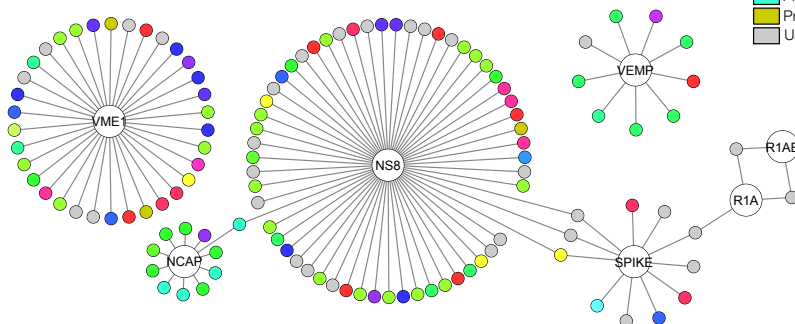
These predictions have been deposited in a public DataVerse repository for use by the broader scientific community in the collective effort to combat the COVID-19 pandemic (Dick et al., 2020a). We re-emphasize that the information provided is theoretical modelling only and caution should be exercised in its use. It is intended only as a resource for the scientific community at large in furthering our understanding of SARS-CoV-2.

ACKNOWLEDGMENTS

The authors would like to thank François Charih for his valuable and constructive suggestions during the planning and development of this research work. This study was funded by a grant from the Canadian Natural Sciences and Engineering Research Council (NSERC) to JRG (RGPIN/327498-2011).

REFERENCES

Amos-Binks, A., Patulea, C., Pitre, S., Schoenrock, A., Gui, Y., Green, J. R., Golshani, A., and Dehne, F. (2011). Binding site prediction for protein-protein interactions and novel motif discovery using re-occurring polypeptide sequences. *BMC bioinformatics*, 12(1):225.

(A) All Schema

(B) Proximal Schema

(C) RP-PPI Schema

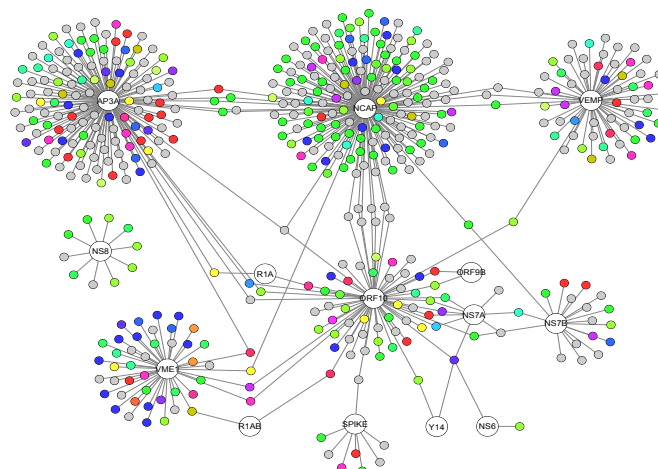



Figure 11. Network Visualization of the Complete Predicted Interactomes for each Schema.
Nodes represent proteins with color representing biological function and edges represent a predicted
interaction. A complete analysis of GO-terms for each network is available in the supplementary material.
GO-terms are approximately ordered by function, where related functions are closer in colour. Created
using Cytoscape (Shannon et al., 2003).

558 Baltimore, D. (1971). Expression of animal virus genomes. *Bacteriological reviews*, 35(3):235.

559 Calderone, A., Licata, L., and Cesareni, G. (2015). Virusmentha: a new resource for virus-host protein
560 interactions. *Nucleic acids research*, 43(D1):D588–D592.

561 Chang, C.-k., Sue, S.-C., Yu, T.-h, Hsieh, C.-M., Tsai, C.-K., Chiang, Y.-C., Lee, S.-j., Hsiao, H.-h., Wu,
562 W.-J., Chang, W.-L., Lin, C.-H., and Huang, T.-h. (2006). Modular organization of sars coronavirus
563 nucleocapsid protein. *Journal of biomedical science*, 13(1):59–72.

564 Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd*
565 *acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

566 Cui, H., Gao, Z., Liu, M., Lu, S., Mkandawire, W., Narykov, O., Srinivasan, S., Sun, M., and Korkin, D.
567 (2020). Structural genomics and interactomics of 2019 wuhan novel coronavirus, sars-cov-2, indicate
568 evolutionary conserved functional regions of viral proteins. 10.20944/preprints202002.0372.v1.

569 Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. *Nature Reviews*
570 *Microbiology*, 17(3):181–192.

571 Dick, K., Biggar, K. K., and Green, J. R. (2020a). Comprehensive Prediction of the SARS-CoV-2 vs.
572 Human Interactome using PIPE4, SPRINT, and PIPE-Sites.

573 Dick, K. and Green, J. R. (2018). Reciprocal perspective for improved protein-protein interaction
574 prediction. *Scientific reports*, 8(1):1–12.

575 Dick, K., Samanfar, B., Barnes, B., Cober, E. R., Mimee, B., Molnar, S. J., Biggar, K. K., Golshani, A.,
576 Dehne, F., and Green, J. R. (2020b). Pipe4: Fast ppi predictor for comprehensive inter-and cross-species
577 interactomes. *Scientific Reports*, 10(1):1–15.

578 Donoghue, M., Hsieh, F., Baronas, E., Godbout, K., Gosselin, M., Stagliano, N., Donovan, M., Woolf, B.,
579 Robison, K., Jeyaseelan, R., Breitbart, R. E., and Acton, S. (2000). A novel angiotensin-converting
580 enzyme–related carboxypeptidase (ace2) converts angiotensin i to angiotensin 1-9. *Circulation research*,
581 87(5):e1–e9.

582 Fan, Z., Zhuo, Y., Tan, X., Zhou, Z., Yuan, J., Qiang, B., Yan, J., Peng, X., and Gao, G. F. (2006).
583 Sars-cov nucleocapsid protein binds to hubc9, a ubiquitin conjugating enzyme of the sumoylation
584 system. *Journal of medical virology*, 78(11):1365–1373.

585 Glowacka, I., Bertram, S., Herzog, P., Pfefferle, S., Steffen, I., Muench, M. O., Simmons, G., Hofmann,
586 H., Kuri, T., Weber, F., Eichler, J., Drosten, C., and Pöhlmann, S. (2010). Differential downregulation
587 of ace2 by the spike proteins of severe acute respiratory syndrome coronavirus and human coronavirus
588 nl63. *Journal of Virology*, 84(2):1198–1205.

589 Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj,
590 V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Hüttenhain, R., Kaake, R. M., Richards, A. L.,
591 Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., Kim, M., Haas, P., Polacco, B. J.,
592 Braberg, H., Fabius, J. M., Eckhardt, M., Soucheray, M., Bennett, M. J., Cakir, M., McGregor, M. J.,
593 Li, Q., Meyer, B., Roesch, F., Vallet, T., Mac Kain, A., Miorin, L., Moreno, E., Naing, Z. Z. C., Zhou,
594 Y., Peng, S., Shi, Y., Zhang, Z., Shen, W., Kirby, I. T., Melnyk, J. E., Chorba, J. S., Lou, K., Dai,
595 S. A., Barrio-Hernandez, I., Memon, D., Hernandez-Armenta, C., Lyu, J., Mathy, C. J. P., Perica,
596 T., Pilla, K. B., Ganesan, S. J., Saltzberg, D. J., Rakesh, R., Liu, X., Rosenthal, S. B., Calviello, L.,
597 Venkataramanan, S., Liboy-Lugo, J., Lin, Y., Huang, X.-P., Liu, Y., Wankowicz, S. A., Bohn, M., Safari,
598 M., Ugur, F. S., Koh, C., Savar, N. S., Tran, Q. D., Shengjuler, D., Fletcher, S. J., O'Neal, M. C., Cai,
599 Y., Chang, J. C. J., Broadhurst, D. J., Klippsten, S., Sharp, P. P., Wenzell, N. A., Kuzuoglu-Ozturk, D.,
600 Wang, H.-Y., Trenker, R., Young, J. M., Cavero, D. A., Hiatt, J., Roth, T. L., Rathore, U., Subramanian,
601 A., Noack, J., Hubert, M., Stroud, R. M., Frankel, A. D., Rosenberg, O. S., Verba, K. A., Agard, D. A.,
602 Ott, M., Emerman, M., Jura, N., von Zastrow, M., Verdin, E., Ashworth, A., Schwartz, O., d'Enfert, C.,
603 Mukherjee, S., Jacobson, M., Malik, H. S., Fujimori, D. G., Ideker, T., Craik, C. S., Floor, S. N., Fraser,
604 J. S., Gross, J. D., Sali, A., Roth, B. L., Ruggero, D., Taunton, J., Kortemme, T., Beltrao, P., Vignuzzi,
605 M., García-Sastre, A., Shokat, K. M., Shoichet, B. K., and Krogan, N. J. (2020). A sars-cov-2 protein
606 interaction map reveals targets for drug repurposing. *Nature*, pages 1–13.

607 Guarner, J. (2020). Three Emerging Coronaviruses in Two Decades: The Story of SARS, MERS, and
608 Now COVID-19. *American Journal of Clinical Pathology*, 153(4):420–421.

609 Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T. S.,
610 Herrler, G., Wu, N.-H., Nitsche, A., Müller, M. A., Drosten, C., and Pöhlmann, S. (2020). Sars-cov-2
611 cell entry depends on ace2 and tmprss2 and is blocked by a clinically proven protease inhibitor. *Cell*,
612 181(2):271–280.e8.

Hofmann, H., Pyrc, K., Van Der Hoek, L., Geier, M., Berkhout, B., and Pöhlmann, S. (2005). Human coronavirus nl63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. *Proceedings of the National Academy of Sciences*, 102(22):7988–7993.

Huang, Q., Yu, L., Petros, A. M., Gunasekera, A., Liu, Z., Xu, N., Hajduk, P., Mack, J., Fesik, S. W., and Olejniczak, E. T. (2004). Structure of the n-terminal rna-binding domain of the sars cov nucleocapsid protein. *Biochemistry*, 43(20):6059–6063.

Jiang, X.-S., Tang, L.-Y., Dai, J., Zhou, H., Li, S.-J., Xia, Q.-C., Wu, J.-R., and Zeng, R. (2005). Quantitative analysis of severe acute respiratory syndrome (sars)-associated coronavirus-infected cells using proteomic approaches: implications for cellular responses to virus infection. *Molecular & Cellular Proteomics*, 4(7):902–913.

Kazmirchuk, T., Dick, K., Burnside, D. J., Barnes, B., Moteshareie, H., Hajikarimlou, M., Omidi, K., Ahmed, D., Low, A., Lettl, C., Hooshyar, M., Schoenrock, A., Pitre, S., Babu, M., Cassol, E., Samanfar, B., Wong, A., Dehne, F., Green, J. R., and Golshani, A. (2017). Designing anti-zika virus peptides derived from predicted human-zika virus protein-protein interactions. *Computational Biology and Chemistry*, 71:180–187.

Kyrollos, D. G., Reid, B., Dick, K., and Green, J. R. (2020). Rpmirdip: Reciprocal perspective improves mirna targeting prediction. *Scientific reports*, 10(1):1–13.

Letko, M., Marzi, A., and Munster, V. (2020). Functional assessment of cell entry and receptor usage for sars-cov-2 and other lineage b betacoronaviruses. *Nature microbiology*, pages 1–8.

Li, F. Q., Xiao, H., Tam, J. P., and Liu, D. (2005). Sumoylation of the nucleocapsid protein of severe acute respiratory syndrome coronavirus. *FEBS letters*, 579(11):2387–2396.

Li, W., Moore, M. J., Vasilieva, N., Sui, J., Wong, S. K., Berne, M. A., Somasundaran, M., Sullivan, J. L., Luzuriaga, K., Greenough, T. C., Choe, H., and Farzan, M. (2003). Angiotensin-converting enzyme 2 is a functional receptor for the sars coronavirus. *Nature*, 426(6965):450–454.

Li, Y. and Ilie, L. (2017). Sprint: ultrafast protein-protein interaction prediction of the entire human interactome. *BMC bioinformatics*, 18(1):485.

Luo, H., Chen, Q., Chen, J., Chen, K., Shen, X., and Jiang, H. (2005). The nucleocapsid protein of sars coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein a1. *FEBS letters*, 579(12):2623–2628.

Makino, S., Keck, J. G., Stohlman, S. A., and Lai, M. (1986). High-frequency rna recombination of murine coronaviruses. *Journal of Virology*, 57(3):729–737.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic acids research*, 47(D1):D419–D426.

Minakshi, R., Padhan, K., Rehman, S., Hassan, M. I., and Ahmad, F. (2014). The sars coronavirus 3a protein binds calcium in its cytoplasmic domain. *Virus research*, 191:180–183.

Narayanan, K., Huang, C., and Makino, S. (2008). Sars coronavirus accessory proteins. *Virus research*, 133(1):113–121.

Neuman, B. W., Kiss, G., Kunding, A. H., Bhella, D., Baksh, M. F., Connelly, S., Droese, B., Klaus, J. P., Makino, S., Sawicki, S. G., Siddell, S. G., Stamou, D. G., Wilson, I. A., Kuhn, P., and Buchmeier, M. J. (2011). A structural analysis of m protein in coronavirus assembly and morphology. *Journal of Structural Biology*, 174(1):11–22.

Pfefferle, S., Schöpf, J., Kögl, M., Friedel, C. C., Müller, M. A., Carbajo-Lozoya, J., Stellberger, T., von Dall'Armi, E., Herzog, P., Kallies, S., Niemeyer, D., Ditt, V., Kuri, T., Züst, R., Pumpor, K., Hilgenfeld, R., Schwarz, F., Zimmer, R., Steffen, I., Weber, F., Thiel, V., Herrler, G., Thiel, H.-J., Schwegmann-Weßels, C., Pöhlmann, S., Haas, J., Drosten, C., and von Brunn, A. (2011). The sars-coronavirus-host interactome: Identification of cyclophilins as target for pan-coronavirus inhibitors. *PLOS Pathogens*, 7(10):1–15.

Pitre, S., Dehne, F., Chan, A., Cheetham, J., Duong, A., Emili, A., Gebbia, M., Greenblatt, J., Jessulat, M., Krogan, N., Luo, X., and Golshani, A. (2006). Pipe: a protein-protein interaction prediction engine based on the re-occurring short polypeptide sequences between known interacting protein pairs. *BMC bioinformatics*, 7(1):365.

Pitre, S., Hooshyar, M., Schoenrock, A., Samanfar, B., Jessulat, M., Green, J. R., Dehne, F., and Golshani, A. (2012). Short co-occurring polypeptide regions can predict global protein interaction maps. *Scientific reports*, 2:239.

668 Regional Office for the Eastern Mediterranean (2011). Mers situation update.

669 Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a" kneedle" in a haystack: Detecting
670 knee points in system behavior. In *2011 31st international conference on distributed computing systems*
671 *workshops*, pages 166–171. IEEE.

672 Schoeman, D. and Fielding, B. C. (2019). Coronavirus envelope protein: current knowledge. *Virology*
673 *journal*, 16(1):1–22.

674 Schoenrock, A., Dehne, F., Green, J. R., Golshani, A., and Pitre, S. (2011). Mp-pipe: a massively
675 parallel protein-protein interaction prediction engine. In *Proceedings of the international conference*
676 *on Supercomputing*, pages 327–337.

677 Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W.,
678 Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver,
679 D., Kavukcuoglu, K., and Hassabis, D. (2020). Improved protein structure prediction using potentials
680 from deep learning. *Nature*, pages 1–5.

681 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski,
682 B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular
683 interaction networks. *Genome research*, 13(11):2498–2504.

684 Shen, X. and Masters, P. S. (2001). Evaluation of the role of heterogeneous nuclear ribonucleoprotein a1
685 as a host factor in murine coronavirus discontinuous transcription and genome replication. *Proceedings*
686 *of the National Academy of Sciences*, 98(5):2717–2722.

687 Shi, S. T., Huang, P., Li, H.-P., and Lai, M. M. (2000). Heterogeneous nuclear ribonucleoprotein a1
688 regulates rna synthesis of a cytoplasmic virus. *The EMBO journal*, 19(17):4701–4711.

689 Smith, M. and Smith, J. C. (2020). Repurposing therapeutics for covid-19: Supercomputer-based docking
690 to the sars-cov-2 viral spike protein and viral spike protein-human ace2 interface. *ChemRxiv*.

691 Swiss Institute of Bioinformatics (2020). SARS Coronavirus 2 Proteome, ViralZone.

692 Van Der Most, R. G., Heijnen, L., Spaan, W. J., and De Groot, R. J. (1992). Homologous rna recombination
693 allows efficient introduction of site-specific mutations into the genome of coronavirus mhv-a59 via
694 synthetic co-replicating rnas. *Nucleic acids research*, 20(13):3375–3381.

695 Wang, Y. and Zhang, X. (1999). The nucleocapsid protein of coronavirus mouse hepatitis virus interacts
696 with the cellular heterogeneous nuclear ribonucleoprotein a1 in vitro and in vivo. *Virology*, 265(1):96–
697 109.

698 World Health Organization (2020). *Laboratory Biosafety Manual, 3rd edition.*

699 Zhang, H., Wada, J., Hida, K., Tsuchiyama, Y., Hiragushi, K., Shikata, K., Wang, H., Lin, S., Kanwar,
700 Y. S., and Makino, H. (2001). Collectrin, a collecting duct-specific transmembrane glycoprotein, is a
701 novel homolog of ace2 and is developmentally regulated in embryonic kidneys. *Journal of Biological*
702 *Chemistry*, 276(20):17132–17139.