Protein sequence models for prediction and comparative analysis of the SARS-CoV-2 −human interactome

Meghana Kshirsagar[†], Nure Tasnina[*], Michael D. Ward[‡], Jeffrey N. Law[*], T. M. Murali[*], Juan M. Lavista Ferres[†], Gregory R. Bowman[‡], Judith Klein-Seetharaman[a]

[†] *Microsoft, AI for Good Research Lab, Redmond, WA, USA*
‡*Dept. of Biochemistry & Molecular Biophysics, Washington Univ., St. Louis, MO, USA*
[a] *Colorado School of Mines, Initiative for AI in Bio and Health, Golden, CO, USA*
[*] *Dept. of Computer Science, Virginia Tech, Blacksburg, VA, USA*

Viruses such as the novel coronavirus, SARS-CoV-2, that is wreaking havoc on the world, depend on interactions of its own proteins with those of the human host cells. Relatively small changes in sequence such as between SARS-CoV and SARS-CoV-2 can dramatically change clinical phenotypes of the virus, including transmission rates and severity of the disease. On the other hand, highly dissimilar virus families such as *Coronaviridae*, *Ebola*, and *HIV* have overlap in functions. In this work we aim to analyze the role of protein sequence in the binding of SARS-CoV-2 virus proteins towards human proteins and compare it to that of the above other viruses. We build supervised machine learning models, using Generalized Additive Models to predict interactions based on sequence features and find that our models perform well with an AUC-PR of 0.65 in a class-skew of 1:10. Analysis of the novel predictions using an independent dataset showed statistically significant enrichment. We further map the importance of specific amino-acid sequence features in predicting binding and summarize what combinations of sequences from the virus and the host is correlated with an interaction. By analyzing the sequence-based embeddings of the interactomes from different viruses and clustering them together we find some functionally similar proteins from different viruses. For example, `vif` protein from *HIV-1*, `vp24` from *Ebola* and `orf3b` from SARS-CoV all function as interferon antagonists. Furthermore, we can differentiate the functions of similar viruses, for example `orf3a`'s interactions are more diverged than `orf7b` interactions when comparing SARS-CoV and SARS-CoV-2.

Keywords: protein interaction prediction; SARS-CoV-2; SARS-CoV; generalized additive models ; protein sequence

1. Introduction

Disease-causing pathogens such as viruses introduce their proteins into the host cells where they interact with the host's proteins enabling the virus to replicate inside the host. These interactions between pathogen and host proteins are key to understanding infectious diseases. The experimental discovery of protein-protein interactions (PPI) in general, and including those between host and pathogen, involves biochemical and biophysical methods, most frequently on a large scale using yeast two-hybrid (Y2H) assays and co-immunoprecipitation (co-IP) usually combined with mass spectrometry, but also many others usually applied at smaller scales such as co-crystallization or surface plasmon resonance. Computational techniques complement laboratory-based methods by predicting highly probable PPIs. Supervised machine learning based methods use the known interactions as training data and formulate the interaction prediction problem in a classification setting.[1–3]

For a newly emerged virus such as SARS-CoV-2, the type of information that is most easily obtained is genome sequence information. Within the first few weeks of its discovery, thousands of DNA sequences had been deposited. The much more complex task of discovering the interactome took a few months of the pandemic and the first global interactome study was published in Gordon et al.[4] A sequence based PPI prediction approach, which can use protein sequences derived from the viral DNA sequence, can thus be very informative in the initial stages of understanding a new virus. The rationale behind a sequence-based approach is that the amino-acid sequences of proteins determine its structure and consequently its function in the organism. By using amino-acid sequences of the two proteins of interest as inputs to a model, we can capture the dependence between their individual structural properties, their functions and their binding affinities. Towards this, we make the following contributions:

- We present an *interpretable* model for SARS-CoV-2 – human PPI prediction using only sequence-based features and evaluate these models on various metrics. We show that the performance of our interpretable model on SARS-CoV-2 PPI prediction, is better than that of Random Forests (which have been popular in prior work) and a deep learning approach that uses a Transformer based architecture for modeling protein sequences
- We analyze the interactomes from a sequence perspective, within SARS-CoV-2 and in comparison to other viruses and find interesting observations
- We validate predictions from our model using an additional recently published dataset from Stukalov et al.[5]

2. Methods

Given a virus-human protein interaction represented as the tuple: (p_v, p_h), we model the joint dependence of both the virus protein p_v and human protein p_h's sequences on the output variable, explicitly in the form of sequence feature level interactions. Towards this, we use a non-linear model GA^{2}M (Lou et al.[6]), which extends traditional Generalized Additive Models (GAMs) by incorporating higher-order feature interactions.

The standard GAM model is a generalized linear model in which the predictor depends linearly on unknown smooth functions f_i of some input covariates x_i. It has the following form: $g(E[y]) = \sum_{i \in [1,\ldots,d]} f_i(x_i)$, where d is the number of features or covariates, y is the output variable for an input, g is the link function (for instance: *log*). Here f_i is a linear function over the i^{th} feature of example x.

2.1. *Generalized Additive Models with interactions (GA^{2}M)*

While GAMs usually model the dependent variable as a sum of univariate terms, GA^{2}M permits interactions and consists of univariate and a small number of pairwise interaction terms between pairs of features:

$$g(E[y]) = \sum_i f_i(x_i) + \sum_{i,j\in[1,\ldots,d],i\neq j} f_{ij}(x_i, x_j)$$

Here i, j are indices over the set of all features. In Section 3.2 we describe our feature set in detail. To represent each virus-human PPI example (p_v, p_h), we concatenate the protein sequence features of both p_v and p_h to get a single feature vector of dimension d.

Since GA^{2}M only include one- and two-dimensional components, these components can be visualized and interpreted which has been difficult with neural networks. Lou et al.[6] propose an algorithm to learn GA^{2}M models that learn non-linear functions (trees) for every univariate and bivariate term, with pairs of features for the latter being selected by efficiently ranking all possible pairs of features as candidates and choosing the top k, where k is a hyper-parameter.

3. Gold Standard Interaction Datasets

We consider the following datasets (details in Table 1) in various experimental settings.

(1) a set of human proteins that physically interact with SARS-CoV-2 in human embryonic kidney cells (HEK293) based on affinity-purification mass spectrometry[4]
(2) a multi-level proteomics study[5] of SARS-CoV and SARS-CoV-2 proteins that also involves an affinity-purification mass spectrometry-based binding study but carried out in a human lung epithelial cell line (A549)
(3) Virus-human interactions data for other viruses was downloaded from VirHostNet[7a]

Unlike the interactions reported in the first mass spectrometry study,[4] the data from the second study[5] has homologous PPI within each dataset as well as several interologs between SARS-CoV and SARS-CoV-2. We downloaded the sequences for *Ebola* and *HIV-1* proteins from UniprotKB and those for SARS-CoV and SARS-CoV-2 from the respective publications' supplementary materials.

Table 1. Dataset characteristics

Virus and source	Interactions	Human proteins	Virus proteins
SARS-CoV-2 (Gordon et al.[4])	332	332	28
SARS-CoV (Stukalov et al.[5])	711	624	24
SARS-CoV-2 (Stukalov et al.[5])	1089	882	22
SARS-CoV (VirHostNet)[7]	141	122	23
Ebola (VirHostNet)[7]	221	221	7
HIV-1 (VirHostNet)[7]	618	583	8

[a] `http://virhostnet.prabi.fr/`

3.1. *Dealing with the lack of negative examples*

Due to the way protein interaction studies are designed, it is not possible to identify non-binding proteins: we cannot rule out interactions between baits and preys that are not co-purified in an affinity purification experiment, for instance. In order to build supervised machine learning models from PPI data, negative datasets comprising pairs of proteins that are unlikely to interact are constructed using heuristics such as (a) randomly selecting pairs of proteins from the set of all possible protein pairs,[8] which has ≈600,000 pairs (b) considering two proteins that do not co-locate within a cell. An approach using (b) is infeasible when considering cross-species protein interactions and also has a bias towards functionally dissimilar proteins. Other negative sets are manually curated in databases like Negatome [b], which are based on known protein domain properties such as hydrophobicity and derived from observational studies that note specific protein domains' lack of affinity towards certain other domains. While this data adjusts the bias mentioned above, it does not contain protein domains or families of many viruses, in particular none from *Coronaviridae.*

We found that using the set of 6,532 non-interacting pairs from Negatome resulted in models that were discriminating virus proteins from other proteins (AUC-PR of 0.98) due to the lack of virus proteins in the negative class. The negatives generated by approach (a) do not have this issue or the functional bias discussed above. Hence we randomly sample the requisite number of negatives from a combination of Negatome and the heuristic in (a).

Choice of class skew: We sample negatives at various positive to negative class-skews: balanced, 1:5 meaning we sample five times as many negatives as the number of positives, 1:10, 1:20 and 1:50. Using a balanced set of positives and negatives results in a biased model that has many false positives whereas using a large class-skew (1:50) that represents our prior that most pairs of proteins are unlikely to interact results in a model that captures the properties of the random protein pairs rather than the positive class (which is out-numbered). We analyzed the ranking of positives from the validation data (using the metric Precision @ 10% Recall) to decide the class-skew, which we treat as a global hyper-parameter. We found 1:10 to be the optimal setting that lead to the best Precision @ 10% Recall.

3.2. *Features*

We derived amino acid sequence k-mer features: consisting of the normalized frequency of 1-mers, 2-mers and 3-mers in the protein sequence. In addition to the above, we also derive conjoint triad features.[9] This approach first partitions the twenty amino acids into seven classes based on their dipoles and the volumes of the side chains. Trimers are represented using the classes of amino acids; hence trimers with amino acids belonging to the same classes, such as ART and VKS, are treated identically. There are 7^3 such tri-mers owing to the 7 classes. The `protr`[c] package was used for generating the conjoint triad features and the `fasta2matrix`[d]

[b] `http://mips.helmholtz-muenchen.de/proj/ppi/negatome/`

[c] `https://cran.r-project.org/web/packages/protr/vignettes/protr.html#46_conjoint_triad_descriptors`

[d] `https://noble.gs.washington.edu/proj/nucsvm/fasta2matrix.py`

utility was used to generate other k-mer features. For each virus-host protein pair, we concatenated the feature vectors of the individual proteins. Therefore, each virus-host protein pair had a feature vector of length 17,526 ($20 + 20^2 + 20^3 + 7^3$ from each protein).

Feature selection: The implementation of GA^{2}M that we use[e] does not scale well beyond a few thousand features because the number of pairs of features to consider is very large and the computational complexity of the feature-pair ranking algorithm.[6] To reduce the number of feature-pairs to consider, we select the top 2500 tri-mers in a feature selection step that builds a linear model on other virus-human interactions. This reduces the number of features in our model to ≈7000 features ($20 + 20^2 + 7^3 + 2500$ features per protein to be precise).

4. Experiments

We train various supervised machine learning models on these datasets to explore the strengths and weaknesses of each approach and illustrate that our method of choice, namely GA^{2}M perform well while giving us interpretability. We compare GA^{2}M with Random Forests, which have been popular in prior work on protein-sequence based prediction and a deep learning embeddings based approach, TAPE.[10]

4.1. *TAPE: Transformer based model for protein sequences*

We use the Unirep model[f] from the TAPE repository[10] which was pretrained on masked language modeling of 31 million protein sequences using a Transformer architecture derived from BERT. This model takes as input, a protein, in the form of its amino acid sequence $x = (x_1, \dots x_n)$, where n is the length of the protein sequence and outputs a sequence of continuous embeddings $y = (y_1 \dots y_n)$. The architecture comprises 12 encoder layers, each of which includes multiple attention heads. Intuitively, attention weights define the influence of every token on the next layer's representation for the current token.

To derive TAPE-based embeddings, we apply a `UniRep babbler-1900` model on all protein sequences in our dataset, which gives us 1900 dimensional embeddings for each protein in two modes: `pooled` and `avg` where the former incorporates the temporal aspect of the input sequence and the latter averages over the per-position embeddings. We concatenate the embeddings from the virus and human proteins to get a 3800 dimensional embedding. We trained two types of supervised models using these as features: Logistic Regression and Random Forests. We found no significant difference in the performance from either and show results from the Logistic Regression based models due to computational efficiency. For the embeddings, we found the setting `avg` worked better probably because it captures homology better. The hyper-parameters of all algorithms were trained using nested cross-validation and grid-search over various values. For GA^{2}M, the main hyper-parameter is the number of interaction terms k for which we tried the following values: $0, 10, 50, 100, 200, 500$. We observed that the performance improved until $k = 100$ and then got worse with higher k. We choose $k = 50$ to trade-off computational speed against a small drop in accuracy.

[e] `https://github.com/interpretml/interpret`

[f] `https://github.com/songlab-cal/tape`

5. Results

5.1. *Prediction performance and validation of predicted interactions*

In Fig 1 we show the predictive performance of all approaches in a 5-fold cross validation setup, for a class-skew of 1:10, where each experiment was repeated 20 times, each time with a different set of negative examples. The reported numbers show the mean (horizontal line in the bar) and standard deviation of the metrics. The GA^{2}M model has an AUC-PR of 0.67 on predicting SARS-CoV-2-human PPI and 0.59 on predicting SARS-CoV-human PPI. The results from the TAPE embeddings are similar to that of Random Forests on SARS-CoV-2-human PPI possibly due to the small scale of PPI data.

To evaluate our models further, we score the set of all possible SARS-CoV-2-human protein-pairs: let us call this set U comprising of 29 x $\approx 21,000 = 609,000$ for ≈21,000 `reviewed` proteins from UniprotKB, and validate these scores using the more recently published PPI from Stukalov et al.[5] Towards this, we first train 100 different models on the gold standard dataset from Gordon et al.[4] by using the 332 positives and sampling a random set of negatives from the unlabeled protein pairs for each of the 100 runs. Since the predictions from a single model are likely to have a bias dependent on the exact set of negatives used, we train 100 different models and apply each of them on the set U. The score for each example from U is averaged over the scores from the 100 different models.

After excluding the gold-standard PPI from the set U, we found that 10,211 examples crossed the classifier score threshold of 0.5. Suppl. Table S1 shows the 28 predictions from this list of 10,211 which appear as experimentally determined interactions in.[5]We performed Fisher's exact test to evaluate the statistical significance of this observation (i.e the probability of seeing 28 of 1089 interactions if 10,211 pairs of proteins are sampled from 609,840 pairs) and obtained a p-value of 0.014. Since pull-down/mass spectrometry based methods are prone to false negatives because of technical limitations in the technology, it is likely that additional pairs within the 10,211 highly ranked predictions are also interacting.

5.2. *Enrichment analysis of predicted human binding partners*

Our models predicted 113 unique human proteins to have at least one interaction with a SARS-CoV-2 protein having a score larger than 0.9. We used Fisher's exact test to determine the enrichment of Gene Ontology (GO) biological processes and cellular components in this set of proteins. We considered terms with Benjamini-Hochberg corrected p-value ≤ 0.01. To remove the redundancy resulting from the parent-child relationships in the GO, we used REVIGO[11] to simplify the sets of enriched terms. REVIGO forms groups of highly similar GO terms using a clustering algorithm (which is similar to hierarchical clustering) and then chooses one representative for each cluster while ensuring that no two representatives are more similar than a user-provided cutoff. We used SimRel[12] as the semantic similarity measure and 0.7 as the cutoff. We now discuss some key enriched GO cellular components. The full set of enriched cellular components and biological processes is available in the supplementary materials.

The GO cellular component "actin cytoskeleton" was significantly enriched (p-value 1.74×10^{-14}) in the predicted human binding proteins. Many viruses use and modify the

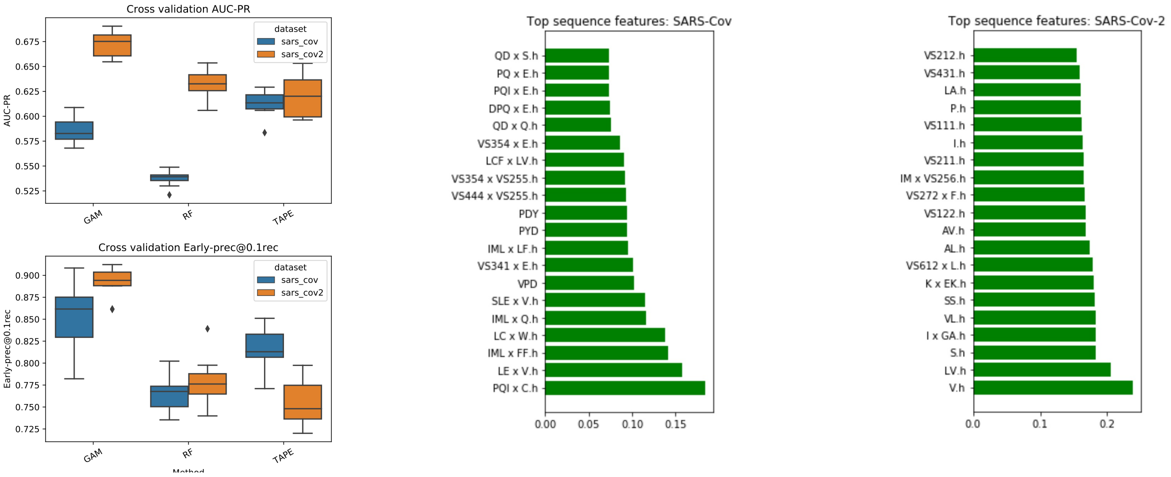


Fig. 1. **(left)** AUC-PR and Precision at 10% Recall averaged over 20 runs for a class skew of 1:10. **(center)** Sequence features relevant to predicting interactions between SARS-CoV and human proteins and **(right)** SARS-CoV-2 and human proteins. Statistics obtained by averaging feature weight from 20 models. Feature names with no suffix are from the virus protein and the suffix '`.h`' refers to that feature from the human protein. Pairwise interaction features are shown as: f_1 x f_2, for instance: `I x GA.h` refers to an interaction between feature `I` from the virus protein and `GA` from the human protein. Features with a prefix of `VS` are conjoint triad features. `VS612` represents a trimer that contains amino-acids from classes 6, 1, and 2. See Fig 3 for the mapping of these classes.

host cell's actin cytoskeleton at different stages of their life cycle including entry, replication, egress, and infection of new cells.[13] In uninfected host cells, viral particles bind to cellular receptors associated with actin filaments in order to travel along filopodia and reach entry sites where endocytosis occurs.[13] Filopodial extensions also act as bridges between infected to uninfected cells to transport virus particles.[13] A global phosphoproteomic analysis[14] of SARS-CoV-2 infection in Caco-2 cells found that the virus induced substantial increase in filopodial protrusions. The authors hypothesized that induction of filopodia might be crucial for egress of SARS-CoV-2 and/or its spread from one cell to another within epithelial monolayers.

The GO cellular component "kinesin complex" was significantly enriched (p-value 1.28×10^{-8}). Kinesins are a family of motor proteins that play an important role in the replication and spread of different viruses by mediating their long distance movement in the microtubule transport system.[15] Our predictions suggest that SARS-CoV-2 may also use kinesins for transport within infected host cells.

6. Discussion

6.1. *Visualizing the virus-human interactions*

Fig. 2(left) shows the embedding of the PPI datasets from Stukalov et al[5] in comparison to HIV, Ebola and SARS. PCA was used for dimensionality reduction from 17,526 features to 100 dimensions, followed by t-SNE to visualize the embedding. The interactive versions of these figures are available in our repository [g], where the user can hover over each entry and

[g] `https://github.com/meghana-kshirsagar/sars_ppi/blob/master/allviruses_plot.html`

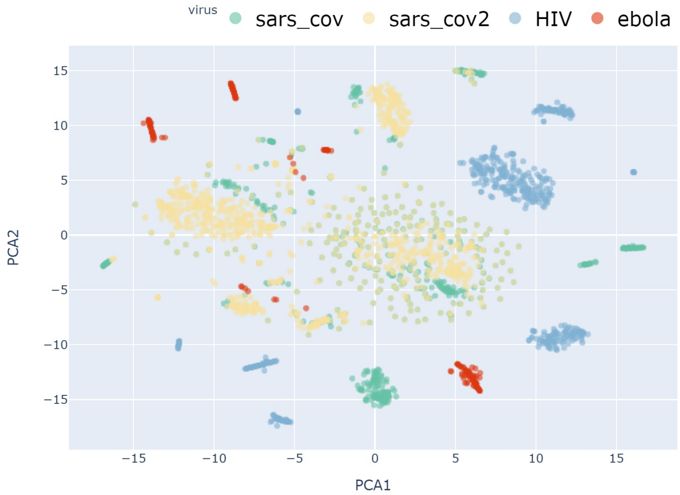
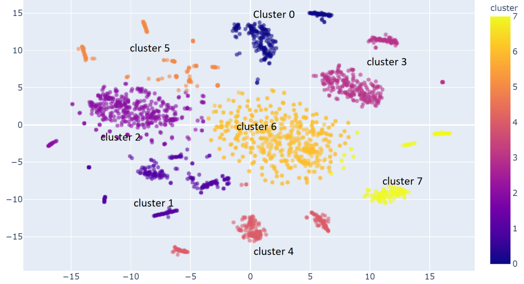


Fig. 2. Embedding of the SARS-CoV and SARS-CoV-2 PPI from Stukalov et al[5] jointly with the *Ebola* and *HIV-1* PPI described in Table 1. Each dot represents a virus-human PPI, colored by the virus species (details in Section 6.1). The large cluster of overlapping yellow and green points at the center shows the interologs between SARS-CoV and SARS-CoV-2.



Class	1	2	3	4	5	6	7
Amino acids	Ala, Gly, Val	Ile, Leu, Phe, Pro	Tyr, Met, Thr, Ser	His, Asn, Gln, Trp	Arg, Lys	Asp, Glu	Cys

Fig. 3. **(top)** K-means clustering of the dimensionality reduced data from the left panel. Each dot is a virus-human PPI coloured by the cluster it was assigned to by the k-means algorithm. **(bottom)** The seven amino-acid classes used in the conjoint triad features; details of the properties used in their classification can be found in Shen et al.[9]

find the protein pair's identity. One can see that in both graphs, there are obvious clusters of interactions, some of which involve only proteins from a single type of virus. In contrast, others show overlap with several viruses.

For further analysis of the PPI clusters, we apply k-means clustering on the 100-dimensional data obtained from PCA and colour the PPI based on which cluster they were assigned to. The result of k-means clustering is shown in Fig. 3 (right). There are 8 clusters, some of which we discuss here. Cluster 0 contains several visually distinct sub-clusters. On the right, there are mostly SARS `N` protein interactions, overlapping with SARS-CoV-2 `N` protein interactions, while those on the left are mostly `M` protein interactions. Cluster 1 includes sub-clusters for *HIV-1* `rev`, SARS `nsp6` (a protein with 4 transmembrane helices and a protease domain), as well as SARS and SARS-CoV-2 `E` and `orf7a` proteins. In the vicinity of the E protein interactions there are several *Ebola* `vp40` interactions as well as a subcluster of *HIV-1* `vpu` interactions. The close proximity of all four viruses implies that there may be commonality in the functions of these interactions. Indeed, in SARS the `M`, `E`, and `N` proteins are required for efficient assembly, trafficking, and release of virus-like particles, as evidenced by the need for co-expression of both `E` and `N` proteins with `M` protein.[16] This is remarkably similar to what has been observed in *Ebola*, where expression of `vp40` alone in mammalian cells induces the production of virus particles with a density similar to that of virions but proper particles require co-expression of `vp40` and `GP`.[17] How do the `nsp6` and `orf7a` proteins fit into this process? While it is known that `nsp6` is involved in autophagy (it limits autophagosome diameter), the proximity to the SARS/SARS-CoV-2 E protein interactions and the *Ebola* `vp40` interactions suggest that there is a connection to virion formation. Unclear is also the role of the *HIV-1* accessory protein vpu, and this proximity may shed light on its function.

Cluster 2 contains a small subcluster on the left, composed mostly with orf9b SARS, and a few orf9b SARS-CoV-2 interactions, but the majority of this cluster are orf3 interactions from SARS-CoV-2, lined with some on the top and the bottom of orf3a from SARS. Cluster 4 contains three subclusters, left: *HIV-1* vif interactions, middle: SARS orf3b interactions and right: *Ebola* vp24 interactions. The functions of vif are not well understood, but for vp24 and orf3b it is clear that they act as IFN antagonists,[18] although the two proteins don't share any detectable sequence similarity. Furthermore, the vif protein in another virus, the caprine arthritis encephalitis virus, appears to be an interferon antagonist as well.[19] This cluster is a particularly strong validation for the concept that the PPI network that a virus protein engages in defines its functions and provides a novel way to identify functional similarity where sequence and structure similarity is not detectable.

Cluster 6 is a large cluster that contains only orf7b from SARS and SARS-CoV-2. Clusters 0, 2 and 6 are the ones most unique to the coronaviruses but with different levels of similarity within. It has been speculated that the differences between orf9b in SARS and SARS-CoV-2 may contribute to the enhanced transmissibility of SARS-CoV-2, possibly due to increased ability to suppress the interferon response.[20] Finally, cluster 7 involves three subclusters, HIV tat, SARS orf8, and orf8a. tat activates RNA Polymerase II,[21] while the functions of orf8/a are not known.[22] Thus, it is tempting to speculate that there may be overlap in these functions with those of tat in HIV.

6.2. *Highly ranked sequence features*

Fig. 1 (center) shows the top-ranked features from SARS-CoV-human interactions and (right) SARS-CoV-2-human interactions. Single letters refer to amino acids in the k-mer, while those with a prefix VS refer to the conjoint triad feature with amino acid groups shown in Fig. 3 (bottom). An extension .h indicates that the feature refers to the human binding partner. One can clearly see that the top-ranked features for the two viruses are different in their detail (which supports that experimental observation that the sequence variations between the two viruses affect their PPIs[5]) but follow similar trends. For example, many of the features refer to hydrophilic amino acid combinations such as QD, PQ, DPQ, QD reflecting the fact that it is the water-exposed surfaces of proteins that engage in PPI interfaces. Furthermore, it is well established that the bulky aromatic, yet hydrophilic side-chain Y is often found as anchor residues in PPI interfaces. Thus, it is encouraging to find PDY, PYD and triad features involving class 3 amongst the top ranked features.

6.3. *Structural analysis*

Highly ranked sequence features from the model correspond to amino acid residues that form cryptic pockets. Cryptic pockets are cavities that form in protein structures due to thermal fluctuations in vivo, but are not observed in experimentally derived protein structures.[23] These pockets can expose functionally important residues to the surface of a protein and can also be used as targets for drug development.[24] A recent study performed molecular dynamics simulations on the majority of proteins in the SARS-CoV-2 proteome to sample the ensemble of structural poses that each protein adopts,[25] using a specialized algorithm to focus on sampling cryptic pockets.[26] The group curated a dataset indicating which residues are part of a cryptic pocket based on analysis using LIGSITE,[27] which performs a grid-based search for pockets, and exposons,[28] which identifies residues that have cooperative changes in their solvent exposure. Overlaying sequence features from the PPI model onto one of the SARS-CoV-2 proteins, Nonstructural protein 16 (`nsp16`), we find that the positions found significant by the model coincide with the location of 3 out of the 5 pockets. This protein is of particular interest since it has more pockets than any other protein in the dataset, and is an interesting drug target since it is known to be involved in evading the host immune response.[29]

7. Prior Work

Network analysis of SARS-CoV-2 has been carried out since the first SARS-CoV-2 related PPI dataset was deposited in BioRxiv on March 22, 2020.[4] The majority of analyses have focused on identifying targets for repurposing drugs,[30–32] and/or to better understand the molecular details underlying viral pathogenesis.[33,34] These network analysis papers use known human-human PPI to follow the paths from original human-virus pair into the human interactome. This network propagation approach has also been extended to include predicted human-human PPI.[35] A few groups have also looked at the prediction of new interactions between virus and human host proteins: PIPE[36] uses sequence-based PPI predictors PIPE4 and SPRINT to predict interactions for only 14 of the 29 SARS-CoV-2 proteins based on known PPI obtained from the VirusMentha database[37] which currently contains 5 SARS (not SARS-CoV-2) PPIs.

8. Conclusion

We developed a sequence-only based feature prediction model for interactions between SARS-CoV-2 and human proteins. Validation by an independent dataset showed significant enrichment of experimentally validated interactions in the highly-ranked predictions, strongly supporting the approach. The interpretability of our model also allows designing hypotheses toward disrupting these interactions, a crucial step in exploiting PPI prediction for antiviral drug discovery.

Supplementary material: Additional plots, tables, predicted PPI and enrichment analysis are available at: `https://github.com/meghana-kshirsagar/sars_ppi`

9. Acknowledgements

We would like to thank Mark Crovella and Simon Kasif for their help in discussions. This work was supported by National Science Foundation CISE grants 2031614 and 1940169 (to J.K-S.) and NSF grants DBI-1759858 and MCB-1817736 (to T.M.M.) and the Computational Tissue Engineering Graduate Education Program at Virginia Tech.

References

1. M. D. Dyer, T. Murali and B. W. Sobral, Supervised learning and prediction of physical interactions between human and HIV proteins, *Infection, Genetics and Evolution* **11**, 917 (2011).

2. M. Kshirsagar, K. Murugesan, J. G. Carbonell and J. Klein-Seetharaman, Multitask matrix completion for learning protein interactions across diseases, *Journal of Computational Biology* **24**, 501 (2017).
3. M. Chen, C. J.-T. Ju, G. Zhou, X. Chen, T. Zhang, K.-W. Chang, C. Zaniolo and W. Wang, Multifaceted protein–protein interaction prediction based on siamese residual rcnn, *Bioinformatics* **35**, i305 (2019).
4. D. E. Gordon, G. M. Jang, M. Bouhaddou, J. Xu, K. Obernier, K. M. White, M. J. O'Meara, V. V. Rezelj, J. Z. Guo, D. L. Swaney *et al.*, A SARS-CoV-2 protein interaction map reveals targets for drug repurposing, *Nature* , 1 (2020).
5. A. Stukalov, V. Girault, V. Grass, V. Bergant, O. Karayel, C. Urban, D. A. Haas, Y. Huang, L. Oubraham, A. Wang *et al.*, Multi-level proteomics reveals host-perturbation strategies of SARS-CoV-2 and SARS-CoV, *bioRxiv* (2020).
6. Y. Lou, R. Caruana, J. Gehrke and G. Hooker, Accurate intelligible models with pairwise interactions, *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* , 623 (2013).
7. T. Guirimand, S. Delmotte and V. Navratil, VirHostNet 2.0: surfing on the web of virus/host molecular interactions data, *Nucleic acids research* **43**, D583 (2015).
8. M. Kshirsagar, J. Carbonell and J. Klein-Seetharaman, Multitask learning for host–pathogen protein interactions, *Bioinformatics* **29**, i217 (2013).
9. J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li and H. Jiang, Predicting protein–protein interactions based only on sequences information, *Proceedings of the National Academy of Sciences* **104**, 4337 (2007).
10. R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel and Y. S. Song, Evaluating protein transfer learning with tape, *Advances in Neural Information Processing Systems* (2019).
11. F. Supek, M. Bošnjak, N. Škunca and T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms, *PloS one* **6**, p. e21800 (2011).
12. A. Schlicker, F. S. Domingues, J. Rahnenführer and T. Lengauer, A new measure for functional similarity of gene products based on Gene Ontology, *BMC bioinformatics* **7**, p. 302 (2006).
13. M. P. Taylor, O. O. Koyuncu and L. W. Enquist, Subversion of the actin cytoskeleton during viral infection, *Nature Reviews Microbiology* **9**, 427 (2011).
14. M. Bouhaddou, D. Memon, B. Meyer, K. M. White, V. V. Rezelj, M. C. Marrero, B. J. Polacco, J. E. Melnyk, S. Ulferts, R. M. Kaake *et al.*, The global phosphorylation landscape of sars-cov-2 infection, *Cell* **182**, 685 (2020).
15. M. P. Dodding and M. Way, Coupling viruses to dynein and kinesin-1, *The EMBO journal* **30**, 3527 (2011).
16. Y. Siu, K. Teoh, J. Lo, C. Chan, F. Kien, N. Escriou, S. Tsao, J. Nicholls, R. Altmeyer, J. Peiris *et al.*, The M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles, *Journal of virology* **82**, 11318 (2008).
17. T. Noda, H. Sagara, E. Suzuki, A. Takada, H. Kida and Y. Kawaoka, Ebola virus VP40 drives the formation of virus-like filamentous particles along with GP, *Journal of virology* **76**, 4855 (2002).
18. A. P. Zhang, D. M. Abelson, Z. A. Bornholdt, T. Liu, V. L. Woods, Jr and E. O. Saphire, The ebolavirus VP24 interferon antagonist: know your enemy, *Virulence* **3**, 440 (2012).
19. Y. Fu, D. Lu, Y. Su, H. Chi, J. Wang and J. Huang, The vif protein of caprine arthritis encephalitis virus inhibits interferon production, *Archives of virology* (2020).
20. H. Jiang, H. Zhang, Q. Meng, J. Xie, Y. Li, H. Chen, Y. Zheng, X. Wang, H. Qi, J. Zhang *et al.*, SARS-CoV-2 orf9b suppresses type i interferon responses by targeting TOM70, *Cellular*

& Molecular Immunology , 1 (2020).

21. A. P. Rice, The HIV-1 tat protein: mechanism of action and target for hiv-1 cure strategies, *Current pharmaceutical design* **23**, 4098 (2017).
22. C.-T. Keng and Y.-J. Tan, Molecular and biochemical characterization of the sars-cov accessory proteins orf8a, orf8b and orf8ab, in *Molecular Biology of the SARS-Coronavirus*, (Springer, 2010) pp. 177–191.
23. C. R. Knoverek, G. K. Amarasinghe and G. R. Bowman, Advanced methods for accessing protein shape-shifting present new therapeutic opportunities, *Trends in Biochemical Sciences* (2019).
24. D. Beglov, D. R. Hall, A. E. Wakefield, L. Luo, K. N. Allen, D. Kozakov, A. Whitty and S. Vajda, Exploring the structural origins of cryptic sites on proteins, *Proceedings of the National Academy of Sciences* (2018).
25. M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, M. F. Hurley, A. M. Harbison, C. A. Fogarty, J. E. Coffland, E. Fadda, V. A. Voelz, J. D. Chodera and G. R. Bowman, Citizen scientists create an exascale computer to combat COVID-19, *bioRxiv* (2020).
26. M. I. Zimmerman and G. R. Bowman, FAST conformational searches by balancing exploration/exploitation trade-offs, *Journal of Chemical Theory and Computation* (2015).
27. M. Hendlich, F. Rippmann and G. Barnickel, LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins, *Journal of Molecular Graphics and Modelling* (1997).
28. J. R. Porter, K. E. Moeder, C. A. Sibbald, M. I. Zimmerman, K. M. Hart, M. J. Greenberg and G. R. Bowman, Cooperative changes in solvent exposure identify cryptic pockets, switches, and allosteric coupling, *Biophysical Journal* (2018).
29. T. Viswanathan, S. Arya, S.-H. Chan, S. Qi, N. Dai, A. Misra, J.-G. Park, F. Oladunni, D. Kovalskyy, R. A. Hromas, L. Martinez-Sobrido and Y. K. Gupta, Structural basis of rna cap modification by SARS-CoV-2, *Nature Communications* (2020).
30. J. Bullock, A. S. Luccioni, K. H. Pham, C. S. N. L. Lam and M. Luengo-Oroz, Mapping the landscape of artificial intelligence applications against COVID-19, *arXiv* (2020).
31. J. N. Law, N. Tasnina, M. Kshirsagar, J. Klein-Seetharaman, M. Crovella, P. Rajagopalan, S. Kasif and T. Murali, Identifying human interactors of SARS-CoV-2 proteins and drug targets for COVID-19 using network-based label propagation, *bioRxiv* (2020).
32. Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin and F. Cheng, Network-based drug repurposing for novel coronavirus 2019-ncov/SARS-CoV-2, *Cell Discovery* **6** (2020).
33. N. Kumar, B. Mishra, A. Mehmood, M. Athar and S. Mukhtar, Integrative network biology framework elucidates molecular mechanisms of SARS-CoV-2 pathogenesis, *iScience* (2020).
34. D. Domingo-Fernandez, S. Baksi, B. Schultz, Y. Gadiya, R. Karki, T. Raschka, C. Ebeling, M. Hofmann-Apitius *et al.*, COVID-19 knowledge graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology, *BioRxiv* (2020).
35. K. B. Karunakaran, N. Balakrishnan and M. K. Ganapatiraju, Interactome of SARS-CoV-2/ncov19 modulated host proteins presents clinically actionable targets for COVID-19, *Research Square* (2020).
36. K. Dick, K. K. Biggar and J. R. Green, Computational prediction of the comprehensive SARS-CoV-2 vs. human interactome to guide the esign of therapeutics, *bioRxiv* (2020).
37. A. Calderone, L. Licata and G. Cesareni, VirusMentha: a new resource for virus-host protein interactions, *Nucleic acids research* **43**, D588 (2015).