MAPPING AND QUERYING PROTEIN-PROTEIN INTERACTIONS FROM SARS-COV-2, HPV, AND HIV DATASETS

SHANZAY FARZAN

DEC. 25, 2020

CHM-240 RESEARCH IN CHEM

DR. PHALGUNI GHOSH

Contents

TITLE

Mapping & Querying Protein-Protein Interactions from SARS-CoV-2, HPV, and HIV Datasets

AUHOR

Shanzay Farzan

ABSTRACT

The novel coronavirus SARS-CoV-2 has infected nearly 13.2 million individuals in the United States and caused the deaths of 266,000 Americans. Using research on viruses that accompanies protein-protein interaction (PPI) networks, we used a data-driven approach to compare the interactions of viral and human proteins. We used PPI visual networks experimentally observed in 3 viruses: SARS-CoV-2, HPV, and HIV. We merged and queried the data for these viruses in one database and created new networks around shared interactions. With these methods, we found human-viral protein interactions common to all three virus types. We used the provided Mass Spectrometry Interaction Statistics (MIST) scores to isolate high-confidence PPI in all three virus types. Using data from the original SARS-CoV-2 network, we also created a composite network of common druggable interaction sites. This gives us a starting point for all sites in the human body where human genes which will potentially react to viral infection, opening up a comparative framework for viral-human protein analysis.

KEYWORDS

COVID-19, protein-protein interactions, PPI, SARS-CoV-2, human papillomavirus, HIV, MIST, Cytoscape

INTRODUCTION

This paper examines protein-protein interactions experimentally observed in 3 viruses: SARS-CoV-2, HPV, and HIV. We used a bioinformatics software called Cytoscape to learn more about the SARS-CoV-2 virus strain and how it functions in the human body in comparison to other viruses.

There are four objectives in this project. The first is to understand and create visual networks in Cytoscape, showing interactions between viral and human proteins. The second objective is to build a database of all of the data among the three viruses. The third is to create a composite SARS-CoV-2 network that shows drug compounds which can inhibit viral/human protein interactions. The fourth objective is to run SQL queries to get some discrete answers about what the data is saying.

We mainly used two programs. One is Cytoscape, which is an open-source software platform for visualizing complex networks. The other program is Microsoft Access to run queries in structured query language, or SQL, a language which is used to communicate with a database.

METHODS AND PROCEDURES

2.1 Data from Virus-Human Protein Interaction

A quick review of the viral life cycle tells us that a virus typically enters the human body, hijacks its protein-building machinery and makes copies of itself to attack other cells. Viruses contain three main elements: an outer viral envelope that will bind to the receptor on the human cell, and an inner capsid and viral RNA contained inside that will be released into the cell's cytoplasm. Once there, this RNA then functions as a template for synthesis of complementary strands, which create new copies of the genome, and translation of new capsid and envelope proteins that will form the newly exiting virus in a process called exocytosis.[1][2]

Three research articles were used alongside their accompanying networks on NDEx, which is collaborative software infrastructure for storing, sharing and publishing biological network info.

The first paper, called "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing" was published in June 2020 by a group of experts attempting to find drugs that could be repurposed for treatment of the infection from SARS-CoV-2. The paper identified human proteins that interact with SARS-CoV-2 protein, and 66 potential human proteins targeted by 69 compounds. The SARS-CoV-2 structure includes four structural, and sixteen non-structural proteins. The structural proteins include the Spike Protein, or S, which allows the virus to penetrate the cell, shown here as pink spikes on the outer layer; the Envelope and Membrane proteins, or E and M help with viral assembly; the Nucleocapsid, or N, allows the virus to camouflage its genetic material to the immune system. The process of infection involves the spike protein binding the to the ACE2 protein receptor site, where it then undergoes the simplified process of commandeering the ribosome to make copies. [3] [4]

In order to plot data into our database, we used the 332 high-confidence interactions in the HEK293 cell line from Supplementary Table 2 with a MIST score greater than or equal to 0.60. [Appendix 1]

We also added data from Supplementary Table 4 and 5, "Literature-derived drugs and reagents that modulate SARS-CoV-2 interactors" and "Expert-identified drugs and reagents that modulate SARS-CoV-2 interactors" as separate rows in the excel spreadsheet in order to create nodes in Cytoscape to identify all high-confidence interactions alongside all of their drug-target associations in one combined network. [Appendix 2]

The second paper, “Global landscape of HIV-human protein complexes,” looks at a network of HIV protein-protein interactions. The HIV genome consists of two identical single-stranded RNA molecules that are enclosed within the core. The genome contains nine genes that encode fifteen viral proteins. There are 3 structural proteins: gag, pol, and env. The *gag* gene provides basic physical make-up of the virus, and *pol* provides the basic mechanism by which retroviruses reproduce, while the others help HIV to enter the host cell and enhance its reproduction. The envelope consists of gp120 and gp41, which are the “spikes” that bind to the CD4 receptor site of white blood cells in the immune system. Although initially dormant, copies of HIV mutate over time until the new copies begin attacking T-cells in the human body, which are responsible for immune system defense. Once T-cell count begins to drop, the body’s ability to fight disease and infection drastically decreases – leading to the development of AIDS. [5]

For our database, we utilized Supplementary Data 3 from the paper, which shows a total of 522 high-confidence interactions in two cell lines, HEK293 and JurKat with a MIST score greater than or equal to 0.75. Due to 25 entries not having a valid gene names, 497 interactions were used. [6]

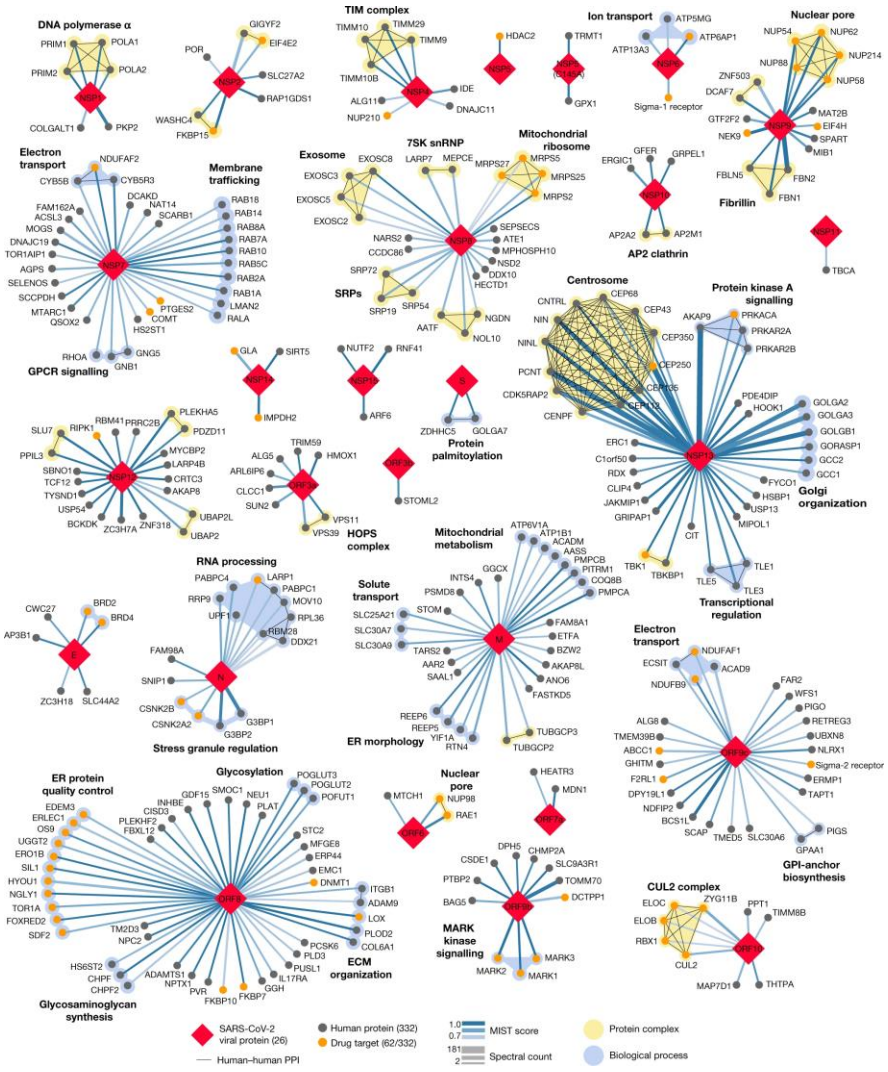
Finally, the third paper used was “Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network” in order to map HPV-31 protein-protein interactions for cross examination of common proteins across all networks. The structure of HPV contains two structural proteins, L1 and L2. The proteins E1, E2, E1^E4, and E8^E2C are responsible for viral replication and regulation, and E5, E6 and E7 are oncogenic, meaning they have the potential to cause cancer. HPV is responsible for cervical cancer, which occurs after the E6 and E7 genes overexpress and stimulate unnatural growth of cells in the cervical lining. [7]

For our database, we utilized Supplementary Table S2 in the C33A cell line from the paper, with a total of 4056 viral-human protein interactions. In order to use relevant data, we used only 567 high-confidence interactions with MIST scores greater than or equal to 0.60. [7]

2.2 NDEx

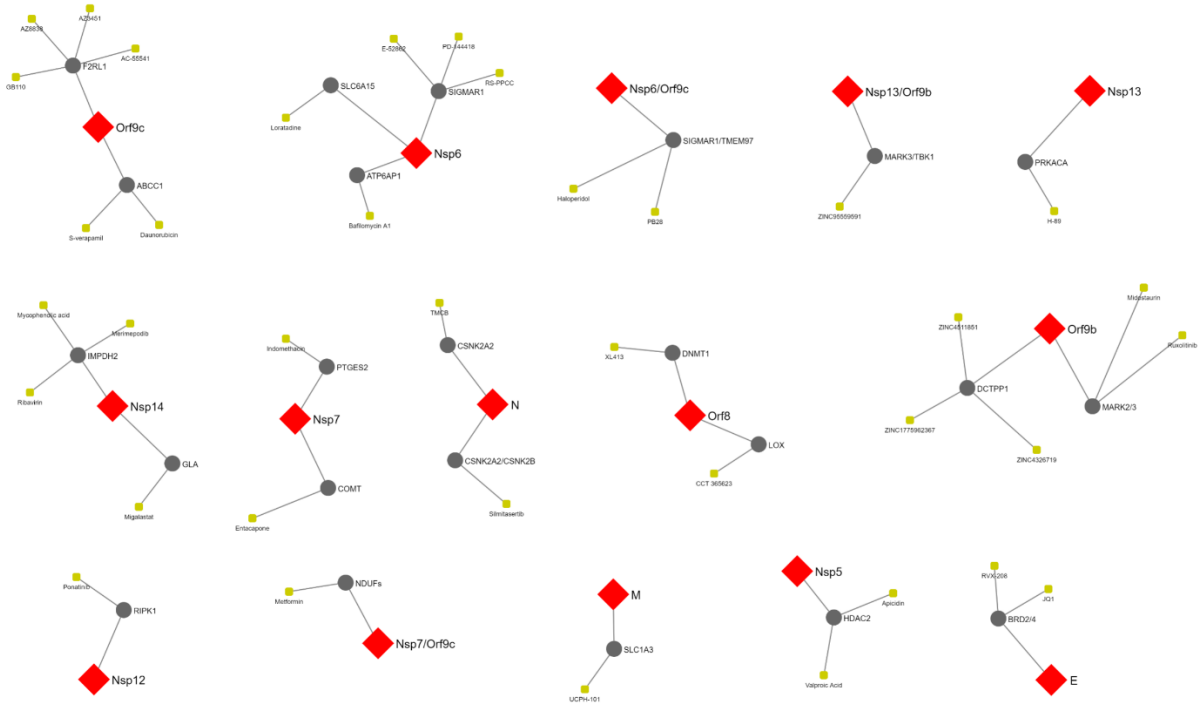
From the three articles mentioned, accompanying networks of human-viral protein interactions were found from NDEx and imported into Cytoscape.

Figure 3 from "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," displays the SARS-CoV-2 Host-Pathogen Interaction Map. The red diamonds represent SARS-CoV-2 viral proteins, and the small gray circles represent human proteins interacting with a viral protein. The clustering mechanism used to display the yellow-highlighted human proteins were not used for this experiment. Finally, the spectrum of blue edges represent the strength of MIST scores, with the light blue edges representing the weakest relative interactions and dark blue edges representing the strongest relative interactions. [8]



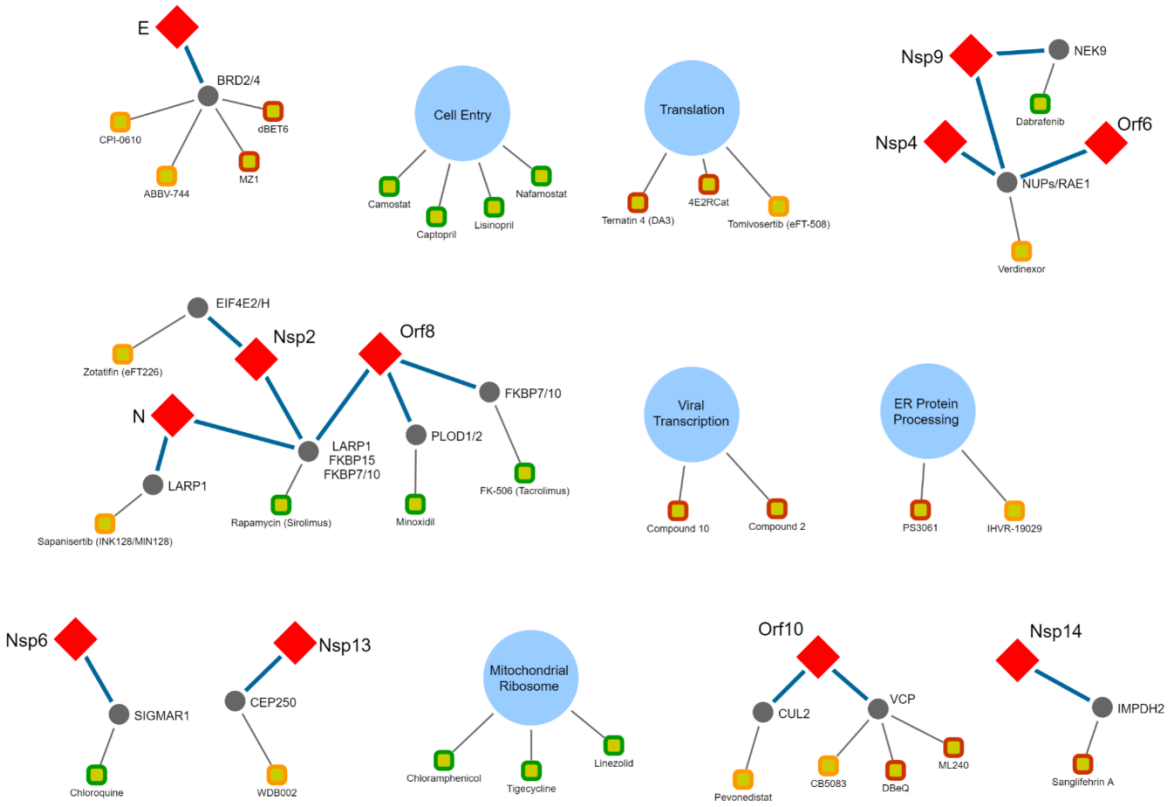
SARS-CoV-2 Host-Pathogen Interaction Map [Fig. 3]

Supplementary Table 3 from the paper shows Literature-derived drugs and reagents that modulate SARS-CoV-2 interactors. The red diamonds represent SARS-CoV-2 viral proteins, and the small gray circles represent human proteins interacting with a viral protein. The green squares represent drugs and reagents that inhibit viral-human protein interactions. Note that this network does not visually identify MIST score strength. [9]



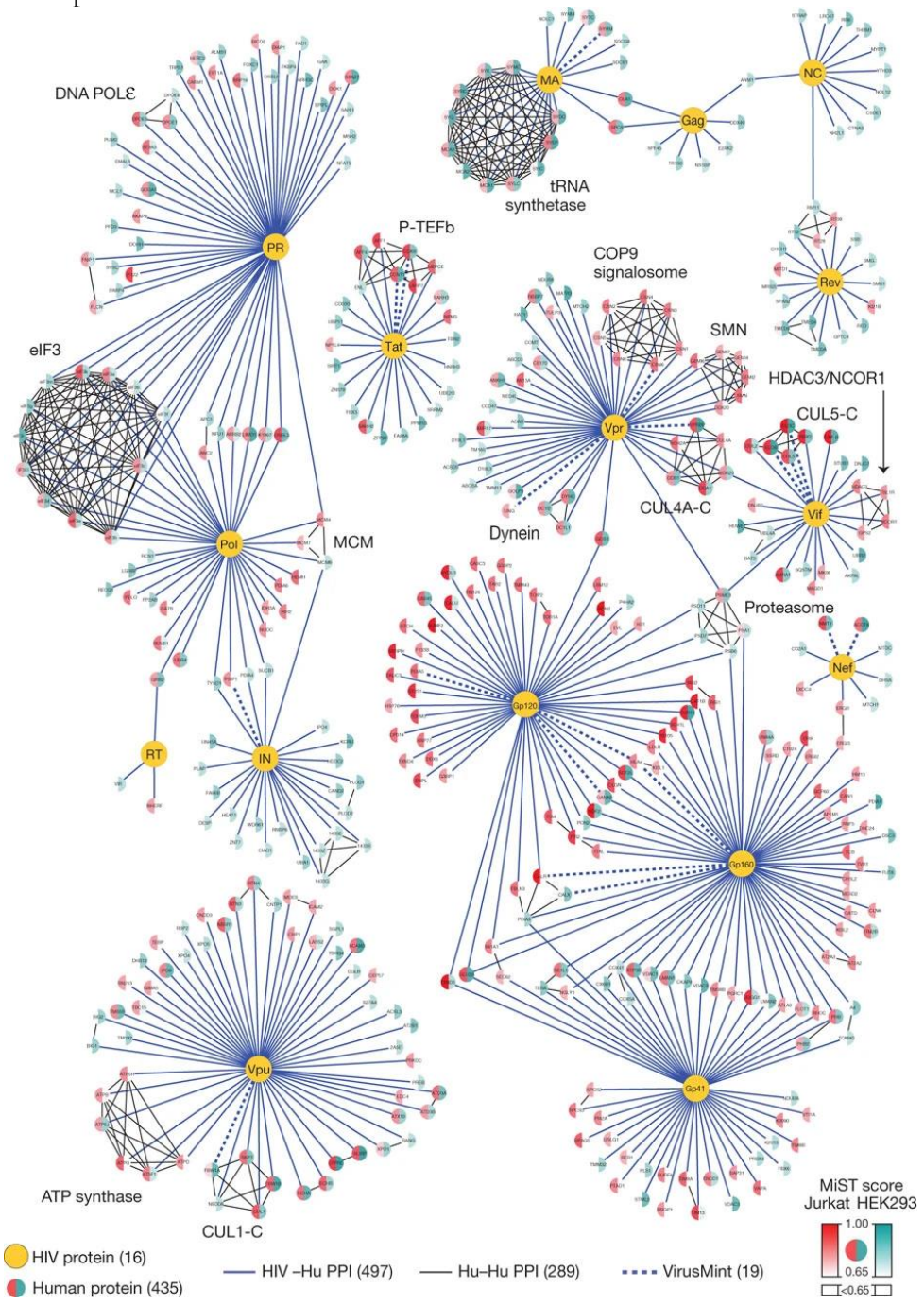
Literature-derived drugs and reagents that modulate SARS-CoV-2 interactors (ST3)

Supplementary Table 4 from the paper shows Expert-identified drugs and reagents that modulate SARS-CoV-2 interactors. The red diamonds represent SARS-CoV-2 viral proteins, and the small gray circles represent human proteins interacting with a viral protein. The green squares represent drugs and reagents that inhibit viral-human protein interactions. Red, green and orange borders show development status of the compound. A green border represents a compound that has been approved, an orange border indicates that the compound is in clinical trials, and a red border represents pre-clinical status. Thicker dark blue edges represent host-pathogen protein-protein interactions. [10]



Expert-identified drugs and reagents that modulate SARS-CoV-2 interactors (ST4)

In the paper, "Global landscape of HIV-human protein complexes," Figure 3 displays a Network representation of the 497 HIV–human PPIs. The yellow circles represent HIV viral proteins, and the smaller red/green/white circles represent human protein interactions. Note that the MIST score is not visually represented in the edge strength, but as darker or lighter red/green gradients in the human protein nodes. This is because results are measured across two cell lines (HEK293 and JurKat) instead of one. Human-Human PPIs were not used in this experiment. [11]



Network representation of HIV–human PPIs

Finally, in "Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network," Figure 2 displays the HPV-Human Protein Network Map. The green diamonds represent HPV viral proteins, and the gray/purple/orange circles represent human protein interactions. Note that the MIST score is not visually represented in the edge strength. The results of human protein interactions are measured across two cell lines (HEK293 shown in purple and gray and Het-1A shown in orange and gray) but only data from HEK293 is used in this experiment. Human-Human PPIs were not used in this experiment. [12]

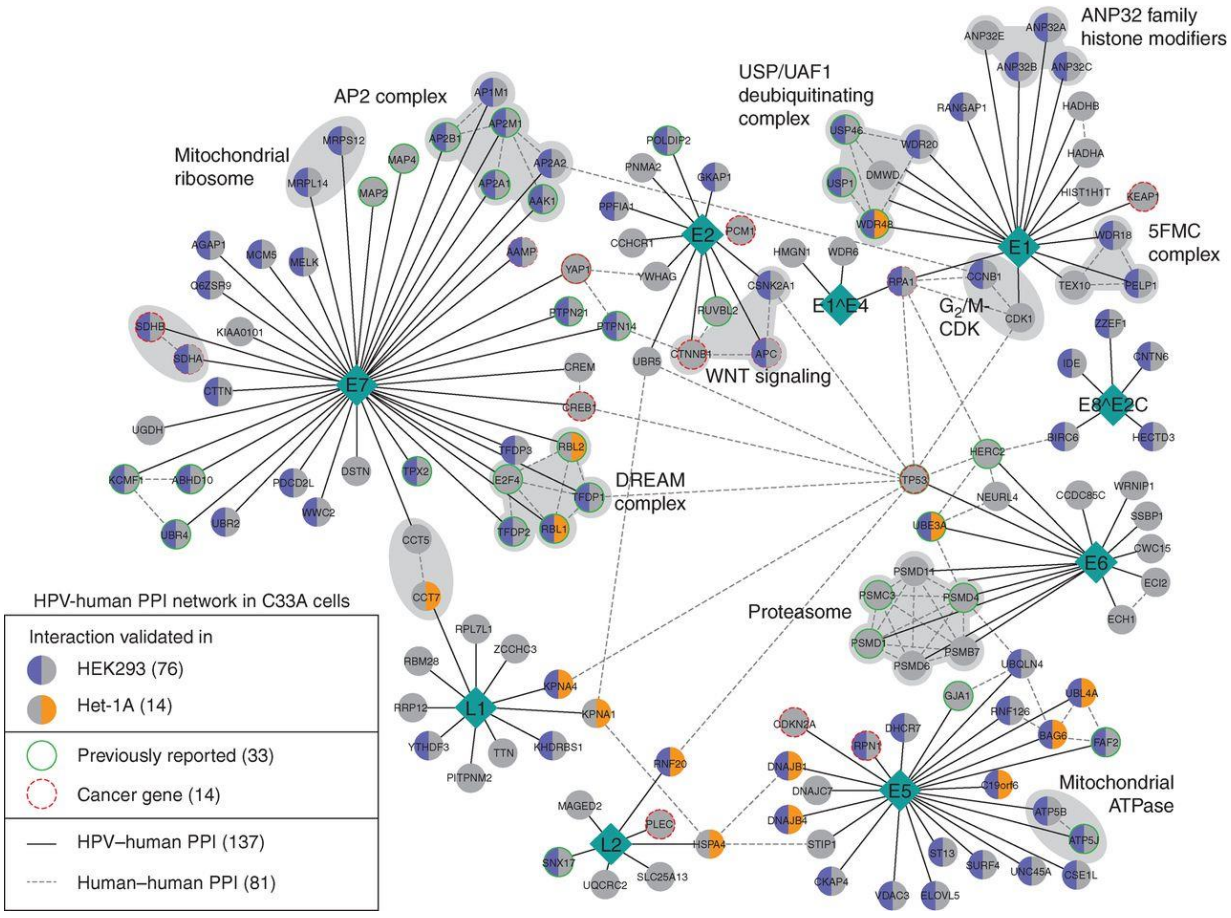


Fig. 2 HPV-human Protein Network Map

2.3 Classification of Virus-Human Protein Interaction & Building database

The first objective is to create a visual network in Cytoscape. When working with data in Excel [Appendix 1], there were 3 columns that were essential to creating networks: the viral protein, the human gene, and MIST.

In Cytoscape, the elements of a visual network are *nodes*, and *edges* that connect the nodes. The three columns mentioned above were assigned as either nodes or edges in Cytoscape. The human protein represents the source node, from which the interaction to the viral protein is being mapped. The viral protein is the target node. There is a connection between them determined by the "edge" – which links the two together explicitly.

The edge strength is made prominent by making the edge line thicker or a different color. For this project, we used an experimentally determined quantity called MIST.

MIST is an experimentally determined quantity that stands for Mass Spectrometry Interaction Statistics. It's a weighted sum of three measures: **protein abundance,** invariability of that abundance over replicated experiments, or **reproducibility, a**nd **specificity**, or uniqueness of an observed host-pathogen interaction across all viral purifications. [6] For our purposes: it is a quantitative score to establish how strong protein-protein interactions are between the human gene and the viral protein. It's a number from 0-1 and is always less than 1, where 0 represents weak interactions and numbers close to 1 represent high confidence interactions.

In order to build a consistent database of all three virus-human interactions, all columns were labeled consistently across the three tables. The columns include: Primary ID, Viral Protein, Virus Name, Viral Protein Type, MIST, Human Gene, UNP ID (uniport ID), and Protein Name. For the SARS-CoV-2 table, additional columns: Compound, Drug Status, and Activity were added. Consistency in labeling made relational querying in Microsoft Access easier across multiple tables. The three essential columns were used to recreate and manipulate networks in Cytoscape. Further, Cytoscape only allows for one node and one edge. Therefore, for the SARS-CoV-2 network, some columns for compounds that inhibit viral-human protein interactions were merged in Microsoft Excel and imported into Cytoscape.

Microsoft Access was used to create the database of all three virus types. A primary key was assigned to all viral proteins in three separate tables labeled: HIV, HPV, and Main. The "Main" table represents the SARS-CoV-2 PPI.

2.4 Building networks in Cytoscape

From "Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network," the HPV network shown in figure 2 above does not show MIST strength in the edges of interactions. [12]

We can manipulate and recreate this network to include a spectrum for MIST score edge strength. By extracting and reorganizing data out of the original network, we re-enter data from the joint database into Cytoscape and exclude the human-human protein interactions.

From "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing," figure 3 displays only viral-human protein interactions. From the same paper, supplementary tables 3 and 4 also display potential drug compounds that can be repurposed for treatment of SARS-CoV-2. These compounds are labeled in the joint database under the columns: Compound, Drug Status, and Activity. Because Cytoscape only allows assigning one target node to the viral protein and one source node to the human protein, the columns for compounds can be added as new rows merged with columns Viral Protein, Viral Type, and Activity respectively, where the compounds and viral proteins are target nodes for the human protein source nodes.

The second objective of this experiment is to create a composite SARS-CoV-2 network showing all of the drug compounds linked to human proteins in the same network as all of the human and viral proteins. After merging the columns of viral proteins with the drug compounds, we stylistically change the node shapes, colors and sizes as consistent with the paper in Cytoscape.

2.5 Building SQL query

The basic structure of a query starts with a SELECT, UPDATE, or DELETE statement. Then, the relevant columns are shown in brackets. Indicate the table from which the columns are used. Finally, attach a condition that specifies the query answer you're looking for.

```
SELECT column_1, column_2, etc.
FROM table
WHERE condition;
```

For example, the viral protein "E" in SARS-CoV-2 network is listed in the database under the table "Main." In order to list all human proteins that interact with the viral protein "E," SELECT the columns "Viral Protein" and "Human Gene" from Main where the viral protein is "E" and the MIST score is greater than 0. This is because there is additional data for drug compounds using KD values, which are above 1. Specifying this condition eliminates redundant instances or blank fields. The result shows six proteins that interact with "E."

```
SELECT Main.[Viral Protein],
   Main.[Human Gene]
FROM Main
WHERE ((Main.[Viral Protein])="E")
   And ((Main.[MIST]>0));
```

Viral Protein	Human Gene
E	AP3B1
E	BRD2
E	BRD4
E	CWC27
E	SLC44A2
E	ZC3H18

Table i

Other queries are discussed below and included in Appendix 4. [Appendix 4]

RESULTS AND DISCUSSION

3.1 Cytoscape findings on SARS-CoV-2 & HPV

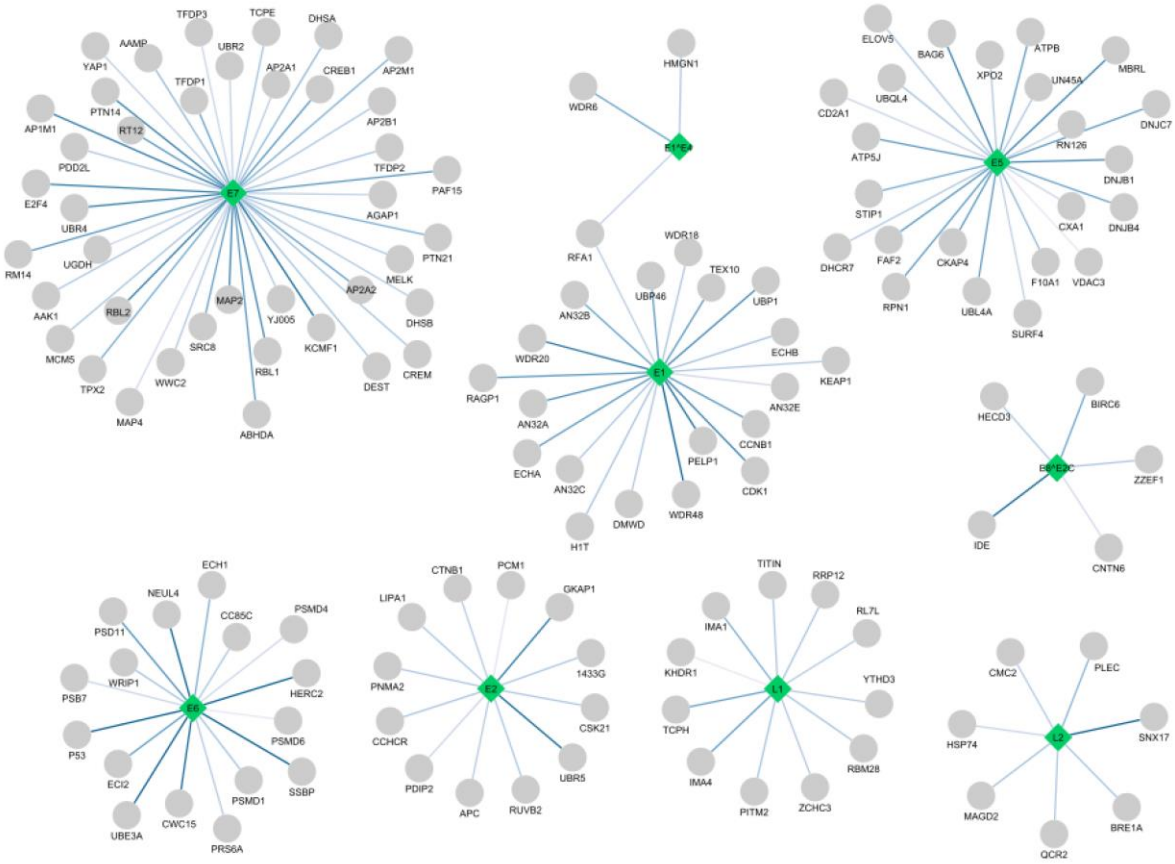


Figure i

The new HPV network is shown above. Visualizing the edge color makes it easily apparent which human proteins strongly interact with HPV viral proteins. The darker the edge, the stronger the recorded interaction between the target and source nodes, or the human and

viral proteins. This lays out a visual representation for any potential drug targets that could have the capacity to inhibit strong HPV PPI.

For the SARS-CoV-2 network, the indicated green dotted edges emphasize drug-human protein interactions in the full recreated network with integrated drug compound nodes. [Appendix 3] From this composite network, an example drug like Selexinor is of interest, because it clearly has linked interactions to several human proteins, and this network directly allows us to see how many. Another example shows the popular drug target chloroquine, from which we can determine it interacts with the SIGMAR1 human protein.

The edges between drug compounds and human proteins are gray and thin, compared to the thicker blue edges between the human and viral proteins. That's because there is no MIST score for drug compound interactions. Instead, another column that measures something called KD values is included, which is not included in the visual network. The development status of the drugs is, again, indicated by the borders around the green squares.

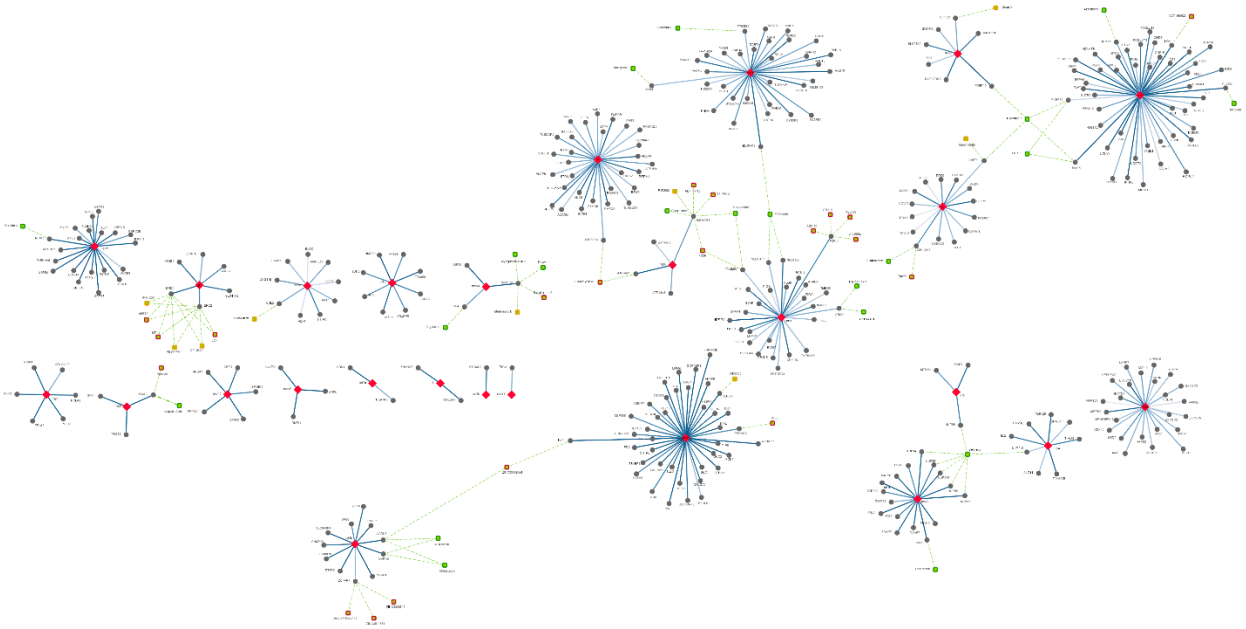


Figure ii

The two drugs in clinical trials are shown in orange, and the drugs in the pre-clinical stage are indicated by a red border.

While this network is extensive, it can still be clearly *seen* in Cytoscape, and that is from the simple fact that the data used in this network contains only 332 high-confidence interactions in one paper. For larger sets of data, such as the combined data between SARS-CoV-2, HIV, and HPV, the visual network will appear clustered and difficult to visualize. For that reason, the joint database is necessary – in order to run SQL queries to get more exact answers to questions.

3.2 Querying the database: SQL commands & findings

In order to see how many interactions between human proteins and viral proteins are really strong, and how many are really weak, the condition needs to be specified. We assign discrete values and categorize the ranges within these discrete values. Since the lowest score doesn't go below 0.6, everything with a MIST score from 0 to 0.7, but not including 0.7 can be assigned "0.6". We create a new column within the query using "Switch" called "MISTRange" which creates discrete values.

```
SELECT DISTINCT Main.[MIST],

Switch(
    Main.[MIST] < 0.70, 0.60,
    Main.[MIST] < 0.80, 0.70,
    Main.[MIST] < 0.90, 0.80,
    Main.[MIST] < 0.99, 0.90,
    Main.[MIST] < 1, 0.99
    ) AS MISTRange

FROM Main;
```

MIST	MISTRange
0.618948613	0.6
0.628043125	0.6
0.631167513	0.6
0.63637328	0.6
0.63921796	0.6
0.647089409	0.6
0.655233005	0.6
0.660192254	0.6
0.669574426	0.6
0.672587659	0.6
0.701764404	0.7
0.703620586	0.7
0.704291901	0.7
0.705393478	0.7
0.706328526	0.7
0.706466834	0.7
0.710174697	0.7
0.710961769	0.7
0.71407894	0.7
0.717059992	0.7
0.717265558	0.7
0.718398295	0.7
0.720285746	0.7
0.720456009	0.7
0.721790867	0.7

Table ii

Then, we used two operators: the inclusive if, and sum. Inside of the sum, we created a function that either evaluates to true or false, i.e. "does this discrete value = 0.6?" if yes, evaluate as true. If no, evaluate as false. Then, outside of that function is another command, which is to "Sum." In this case, it is not literally summing up 0.6+0.6+0.6+... What it is summing is the TRUE values of 0.6. This results in a chartable set of values, where interactions can be specified to fall within the 0.99 range, 0.9 range, 0.8, etc.

```
SELECT

Sum(IIf((MISTRange.MISTRange)=0.6,1,0)) AS Pt6,
Sum(IIf((MISTRange.MISTRange)=0.7,1,0)) AS Pt7,
Sum(IIf((MISTRange.MISTRange)=0.8,1,0)) AS Pt8,
Sum(IIf((MISTRange.MISTRange)=0.9,1,0)) AS Pt9,
Sum(IIf((MISTRange.MISTRange)=0.99,1,0)) AS Pt99

FROM MISTRange

WHERE (((MISTRange.MISTRange) In (0.7,0.8,0.9,0.6,0.99)));
```

Pt6	Pt7	Pt8	Pt9	Pt99
12	84	73	124	39

Table iii

We can then plot this information in a bar graph, as shown below.

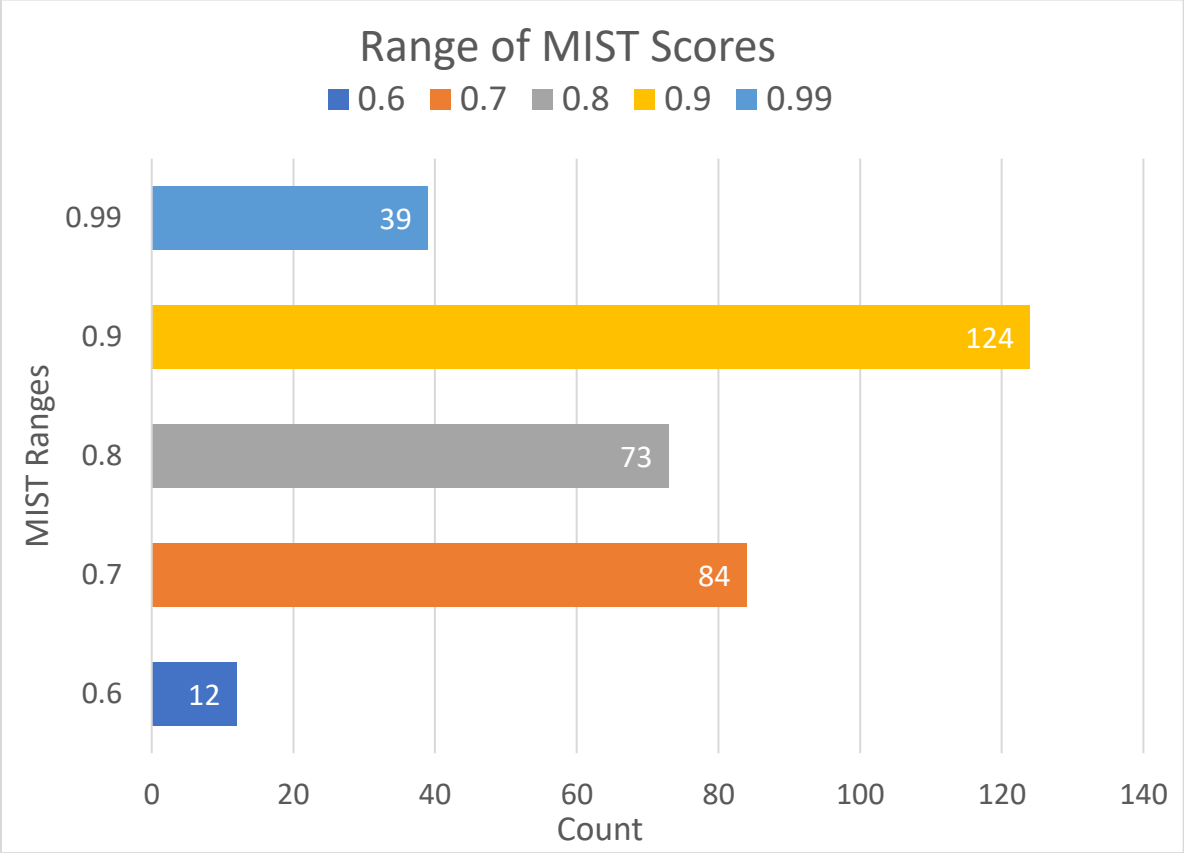


Figure iii

From this, we can see that the MIST score can be used reliably from this paper, since most of the interactions fall within the 0.9 range (124 to be exact) as well as an additional 39 in the 0.99 range, which indicates very strong interactions between the viral and human proteins.

3.3 Commonality & differences in SARS-CoV-2, HPV, & HIV

In order to find human proteins common across the tables "Main" which is SARS-CoV-2, and "HPV," we used an inner join in SQL. The red column indicates the human proteins, the purple columns indicate SARS-CoV-2 viral proteins and MIST scores. Finally, the HPV viral proteins and MIST score columns are blue. We added another condition that both tables have to display results that are greater than 0.6, so only the higher confidence interactions are shown.

```
SELECT DISTINCT

Main.[Human Gene], Main.[Viral Protein], Main.[MIST], HPV.[Viral Protein], HPV.[MIST]

FROM Main

INNER JOIN HPV ON Main.[Human Gene] = HPV.[Human Gene]

WHERE Main.MIST>=0.6 And HPV.MIST>=0.6;
```

	A	B	C	D	E
1	**Human Gene**	**SARS-CoV-2 Viral Protein**	**SARS-CoV-2 MIST (HEK293)**	**HPV Viral Protein**	**HPV MIST (C33A)**
2	**HDAC2**	nsp5	0.993708403	L2	0.685510627
3	**AP2A2**	nsp10	0.99112813	E7	0.882203347
4	**AP2M1**	nsp10	0.982905884	E7	0.853236047
5	**WFS1**	orf9c	0.955124813	E5	0.685510241
6	**IDE**	nsp4	0.918031442	E8^E2C	0.974307196
7	**RTN4**	M	0.873826097	E5	0.685510371
8	**AAR2**	M	0.801486724	E6	0.685511079
9	**ACADM**	M	0.724348569	E7	0.685511333
10	**EMC1**	orf8	0.723777507	E5	0.685510216
11	**AKAP8**	nsp12	0.717059992	L1	0.685510281
12	**RBM28**	N	0.628043125	L1	0.816328465

Table iv

From this we can see that HDAC2 interacts with nsp5 in SARS-CoV-2, but with L2 in HPV.

In order to find common human proteins between HIV and SARS-CoV-2, we perform another inner join. HIV has two columns for MIST scores because the experimental data is from two separate cell lines: HEK293 and JurKat. Regardless, after running the query, there are 25 results of higher confidence interactions.

```
SELECT DISTINCT

Main.[Human Gene], Main.[Viral Protein], Main.[MIST],
        HIV.[Viral Protein], HIV.[MIST HEK293], HIV.[MIST JurKat]

FROM Main

INNER JOIN HIV ON Main.[Human Gene] = HIV.[Human Gene]
WHERE Main.MIST>=0.6 And
      (HIV.[MIST HEK293]>=0.6 OR HIV.[MIST JurKat]>=0.6);
```

	A	B	C	D	E	F
1	**Human Gene**	**SARS-CoV-2 Viral Protein**	**SARS-CoV-2 MIST HEK293**	**HIV Viral Protein**	**HIV MIST HEK293**	**HIV MIST JurKat**
2	**NGLY1**	orf8	0.993220351	GP160	0	0.834
3	**NGLY1**	orf8	0.993220351	GP41	0	0.756
4	**PLOD2**	orf8	0.992483418	IN	0.754	0
5	**FBN2**	nsp9	0.991012329	TAT	0.838	0
6	**AKAP9**	nsp13	0.990813809	PR	0	0.787
7	**CSDE1**	orf9b	0.988959751	NC	0.812	0
8	**G3BP2**	N	0.958133672	GP120	0	0.838
9	**G3BP1**	N	0.95331626	GP120	0	0.784
10	**OS9**	orf8	0.931530938	GP160	0	0.915
11	**ACSL3**	nsp7	0.897068932	VPU	0.821	0.058
12	**TOR1A**	orf8	0.879789245	GP120	0	0.846
13	**RTN4**	M	0.873826097	VPU	0.809	0.819
14	**MTCH1**	orf6	0.851173737	NEF	0.752	0
15	**SDF2**	orf8	0.826107348	GP120	0.856	0.981
16	**SDF2**	orf8	0.826107348	GP160	0.829	0.931
17	**EDEM3**	orf8	0.81834398	GP120	0	0.898
18	**CUL2**	orf10	0.818290141	VIF	0.776	0.897
19	**LARP7**	nsp8	0.812479682	TAT	0.54	0.915
20	**MEPCE**	nsp8	0.790978117	TAT	0.692	0.857
21	**HYOU1**	orf8	0.77235306	GP120	0.775	0.96
22	**COMT**	nsp7	0.745231765	VPR	0.775	0.244
23	**LMAN2**	nsp7	0.725773983	GP160	0	0.839
24	**LMAN2**	nsp7	0.725773983	GP41	0.86	0
25	**ELOB**	orf10	0.655233005	VIF	0.96	0.928
26	**ELOC**	orf10	0.618948613	VIF	0.956	0.972

Table v

Finally, in order to combine all three tables and run the same query, we must place the first inner join between parenthesis so it can apply an inner join to another inner join. SELECT DISTINCT is used to only display unique instances of results.

```
SELECT DISTINCT

Main.[Human Gene],
    Main.[Viral Protein], Main.[MIST],
    HIV.[Viral Protein], HIV.[MIST HEK293], HIV.[MIST JurKat]
    HPV.[Viral Protein], HPV.[MIST],

FROM (Main
    INNER JOIN HPV ON Main.[Human Gene] = HPV.[Human Gene])

INNER JOIN HIV ON HPV.[Human Gene] = HIV.[Human Gene];
```

	A	B	C	D	E	F	G	H
1	**Human Gene**	**SARS-CoV-2 Viral Protein**	**SARS-CoV-2 MIST HEK293**	**HIV Viral Protein**	**HIV MIST HEK293**	**HIV MIST JurKat**	**HPV Viral Protein**	**HPV MIST (C33A)**
2	**CSDE1**	orf9b	0.988959751	NC	0.812	0	E2	0.176316087
3	CSDE1	orf9b	0.988959751	NC	0.812	0	E6	0.0698267
4	CSDE1	orf9b	0.988959751	NC	0.812	0	E8^E2C	0.186285804
5	CSDE1	orf9b	0.988959751	NC	0.812	0	L1	0.356805069
6	CSDE1	orf9b	0.988959751	NC	0.812	0	L2	0.103342845
7	**G3BP2**	N	0.958133672	GP120	0	0.838	E1	0.322921149
8	G3BP2	N	0.958133672	GP120	0	0.838	E1^E4	0.363596661
9	G3BP2	N	0.958133672	GP120	0	0.838	E2	0.383268897
10	G3BP2	N	0.958133672	GP120	0	0.838	E5	0.290177749
11	G3BP2	N	0.958133672	GP120	0	0.838	E6	0.317960495
12	G3BP2	N	0.958133672	GP120	0	0.838	E7	0.289612609
13	G3BP2	N	0.958133672	GP120	0	0.838	E8^E2C	0.385217699
14	G3BP2	N	0.958133672	GP120	0	0.838	L1	0.365959601
15	G3BP2	N	0.958133672	GP120	0	0.838	L2	0.32994147
16	**G3BP1**	N	0.95331626	GP120	0	0.784	E1	0.196739252
17	G3BP1	N	0.95331626	GP120	0	0.784	E1^E4	0.200569845
18	G3BP1	N	0.95331626	GP120	0	0.784	E2	0.263651559
19	G3BP1	N	0.95331626	GP120	0	0.784	E5	0.161791186
20	G3BP1	N	0.95331626	GP120	0	0.784	E6	0.20952895
21	G3BP1	N	0.95331626	GP120	0	0.784	E7	0.26312747
22	G3BP1	N	0.95331626	GP120	0	0.784	E8^E2C	0.204416776
23	G3BP1	N	0.95331626	GP120	0	0.784	L1	0.335680758
24	G3BP1	N	0.95331626	GP120	0	0.784	L2	0.258371514
25	**RTN4**	M	0.873826097	VPU	0.809	0.819	E5	0.685510371
26	**ELOB**	orf10	0.655233005	VIF	0.96	0.928	E1	0.120568043
27	ELOB	orf10	0.655233005	VIF	0.96	0.928	E1^E4	0.016731554
28	ELOB	orf10	0.655233005	VIF	0.96	0.928	E5	0.016963384
29	ELOB	orf10	0.655233005	VIF	0.96	0.928	E7	0.189816337

Table vi

The redundant results have been grayed out to show only the relevant proteins. For example, the CSDE1 human gene interacts with orf9b in SARS-CoV-2, NC in HIV, and The HPV viral proteins E2, E6, E8^E2C, L1, and L2.

This data can then be entered into Cytoscape to create a new network from a set of queries run from the joint database. Note the five highlighted human genes, the five proteins from SARS-CoV-2 and HIV and the nine viral proteins from HPV. Again, edge strength is

represented by the blue spectrum of MIST scores. In Figure iv, there are lines for HPV that are very light, indicating very weak interactions.

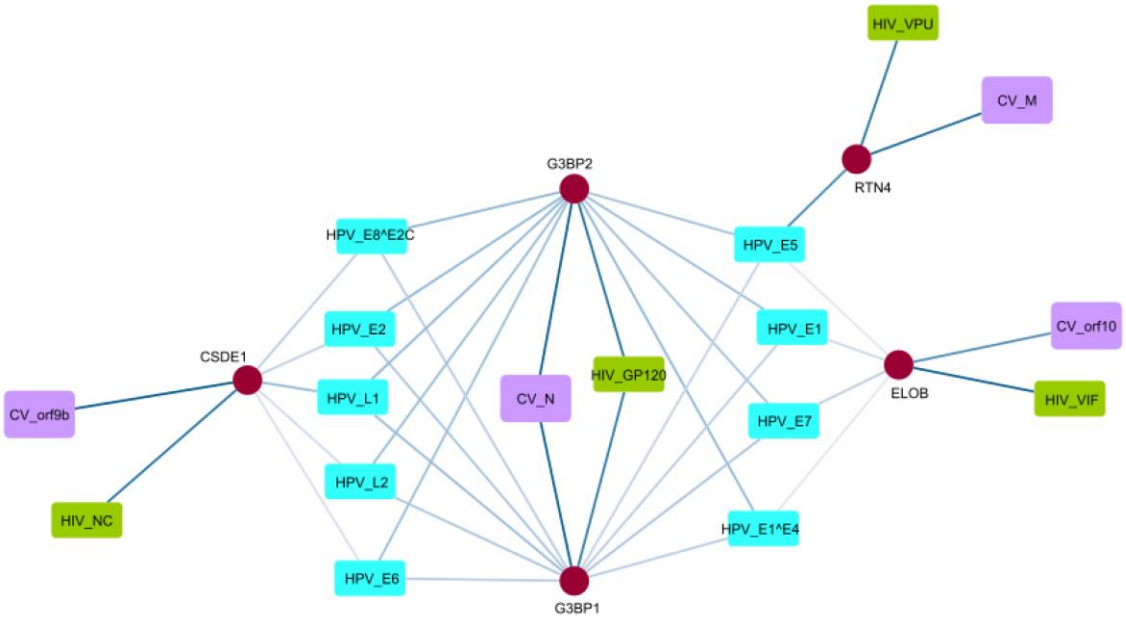


Figure iv

Looking back at the MIST scores, HPV does not include high-confidence interactions, so we specify another condition to only display results where all three tables have MIST scores above 0.6. There is one human gene "RTN4," or Reiculon-4, that has strong interactions with viral proteins from all three viruses. This gene is involved with the formation and stabilization of the endoplasmic reticulum. [13]

```
SELECT DISTINCT Main.[Human Gene], Main.[UNP ID],

    Main.[Viral Protein], Main.[MIST],
    HIV.[Viral Protein], HIV.[MIST HEK293], HIV.[MIST JurKat],
    HPV.[Viral Protein], HPV.[MIST]

FROM (Main INNER JOIN HIV ON Main.[Human Gene] = HIV.[Human Gene])
INNER JOIN HPV ON HIV.[Human Gene] = HPV.[Human Gene]

WHERE Main.[MIST]>=0.6 AND HPV.[MIST]>=0.6
    AND (HIV.[MIST HEK293]>=0.6 OR HIV.[MIST JurKat]>=0.6);
```

	A	B	C	D	E	F	G	H	I
1	UNP ID	Human Gene	SARS-CoV-2 Viral Protein	SARS-CoV-2 MIST HEK293	HIV Viral Protein	HIV MIST HEK293	HIV MIST JurKat	HPV Viral Protein	HPV MIST (C33A)
2	Q9NQC3	RTN4	M	0.873826097	VPU	0.809	0.819	E5	0.685510371

Table vii

Finally, Figure v shows the combined Cytoscape network from all three viral types: SARS-CoV-2, HIV and HPV tables in our joint database. [Appendix 5]

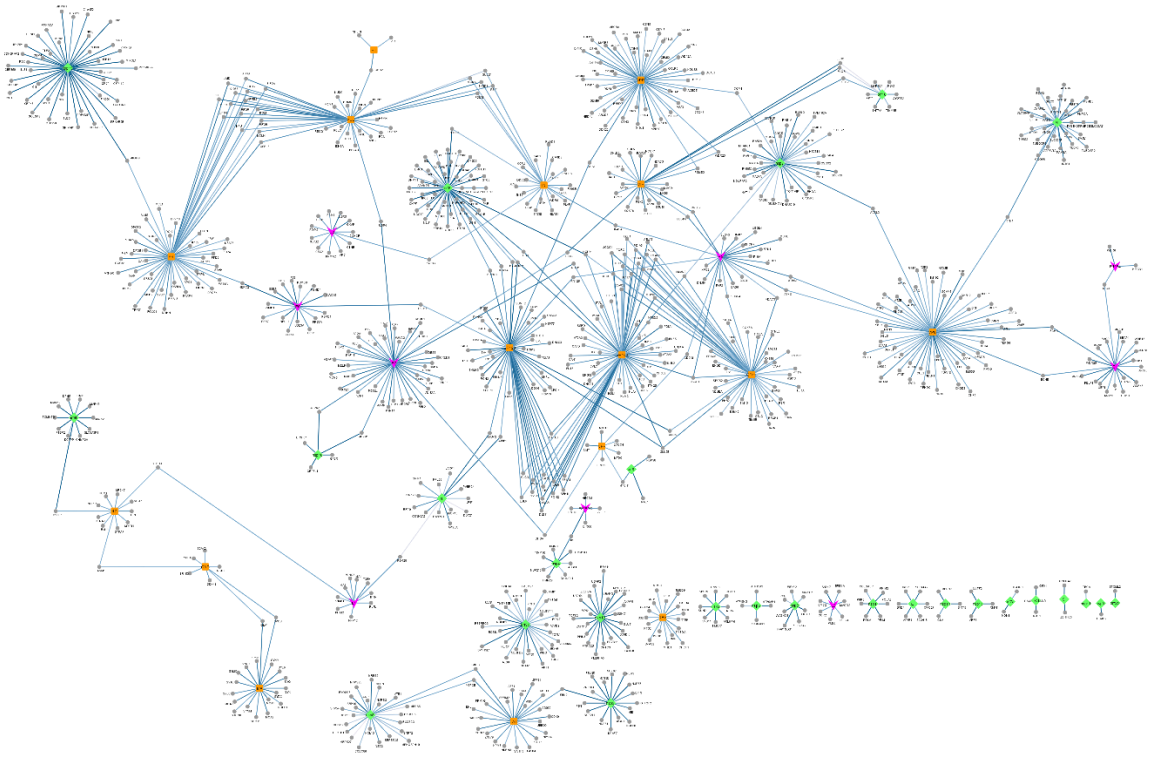


Figure v

CONCLUSION

There are many applications for examining data between the three datasets for SARS-CoV-2, HIV, and HPV in both Microsoft Access and Cytoscape. By merging and querying data for these viruses, we were able to examine novel networks in shared interactions. We were able to find a number of human-viral protein interactions across all three virus types, extending the ability to find common human proteins across other viruses. MIST scores were used to isolate high-confidence PPI in all three virus types. Finally, for the SARS-CoV-2 network, a composite network showcased the most common and most connected druggable interaction sites. Querying and mapping these interactions lays out a clear path for efficiently locating drugs and compounds that have the most potential to interact with human and viral proteins, and slowing the spread of viral infection.

REFERENCES

[1] Urry, L. A., Cain, M. L. 1., Wasserman, S. A., Minorsky, P. V., Reece, J. B., & Campbell, N. A. (2017). *Essential biology*. Eleventh edition. Figure 19.7. New York, NY: Pearson Education, Inc.

[2] "The Protein Data Bank" *RCSB*, pdb101.rcsb.org/.

[3] "A SARS-CoV-2 protein interaction map reveals targets for drug repurposing" Gordon, D.E., Jang, G.M., Bouhaddou, M. *et al. Nature* 583, p. 459-468 (2020).

[4] ScrippsResearch. "Coronavirus Anatomy Explained: Science, Simplified." *YouTube*, YouTube, 13 May 2020, www.youtube.com/watch?v=8hgc2iZflTI.

[5] Votteler J, Schubert U (2008). "Human Immunodeficiency Viruses: Molecular Biology". *Encyclopedia of Virology* (3rd ed.). pp. 517–525.

[6] "Global landscape of HIV-human protein complexes" Jäger, S., Cimermancic, P., Gulbahce, N. *et al. Nature* 481, p. 365-370 (2012).

[7] "Multiple Routes to Oncogenesis Are Promoted by the Human Papillomavirus-Host Protein Network" Eckhardt, M., Zhang, W., Gross, A.M. *et al. Cancer Discovery*, American Association for Cancer Research. Vol. 8, No. 11, p. 1474-1489 (2018).

[8] Krogan, Nevan. "SARS-CoV-2 Host-Pathogen Interaction Map (Fig. 3)." *NDEx*, ndexbio.org/viewer/networks/5d97a04a-6fab-11ea-bfdc-0ac135e8bacf.

[9] Krogan, Nevan. "Literature-derived drugs and reagents that modulate SARS-Cov-2 interactors (ST3)." *NDEx*, http://ndexbio.org/viewer/networks/a49d7cc7-6e1e-11ea-bfdc-0ac135e8bacf.

[10] Krogan, Nevan. "Expert-identified drugs and reagents that modulate SARS-CoV-2 interactors (ST4)." *NDEx*, http://ndexbio.org/viewer/networks/9ee12f90-6fd0-11ea-bfdc-0ac135e8bacf.

[11] Krogan, Nevan. "HIV-Human." *NDEx*, http://ndexbio.org/viewer/networks/1c568feb-d660-11e9-bb65-0ac135e8bacf.

[12] Krogan, Nevan. "Fig. 2 HPV-human Protein Network Map." NDEx, http://ndexbio.org/viewer/networks/c5dd2e82-b888-11e8-9520-0ac135e8bacf.

[13] "Reticulon-4." UNIPROT, https://www.uniprot.org/uniprot/Q9NQC3.

[Appendix 1] Raw Excel data for SARS-CoV-2, HIV, and HPV viral-human proteins.

[Appendix 2] Drugs/compounds merged into SARS-CoV-2 columns for Cytoscape.
[Appendix 3] Full recreated Cytoscape network with integrated drug compound nodes.
[Appendix 4] All SQL queries in Microsoft Access.
[Appendix 5] Combined SARS-CoV-2, HPV, and HIV Cytoscape network.