SARS-CoV-2 Protein Protein Interactions

Led by Snow Naing and Isha Karim

Valerie, Joyce, Ami, Arhana, Yomn, Esha

Our Goal

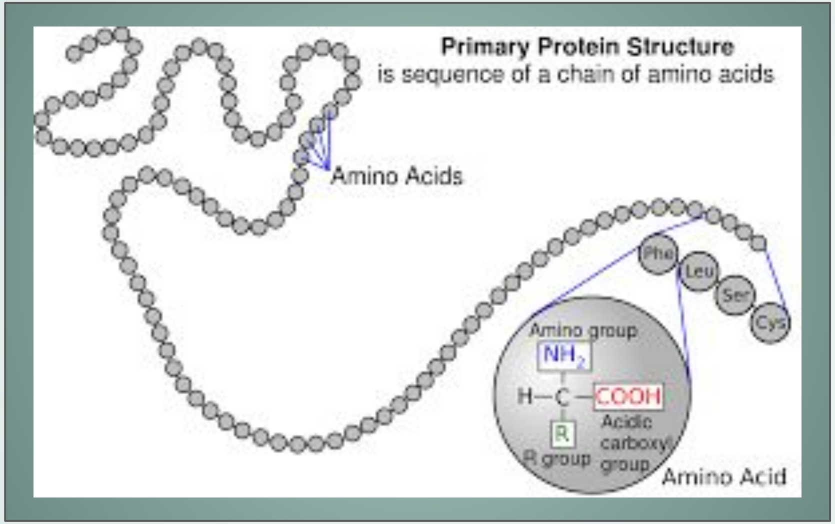
- Covid-19 took everyone by surprise

- Scientists knew little about about how it would interact with the human body

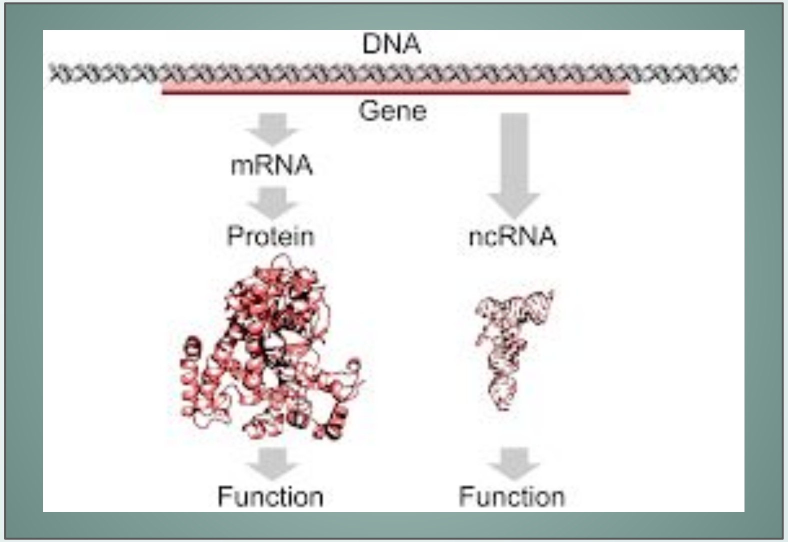
- Create an algorithm that predicts how a new pathogen (Sars-COV-2) will interact with human proteins

THE SCIENCE

PROTEINS

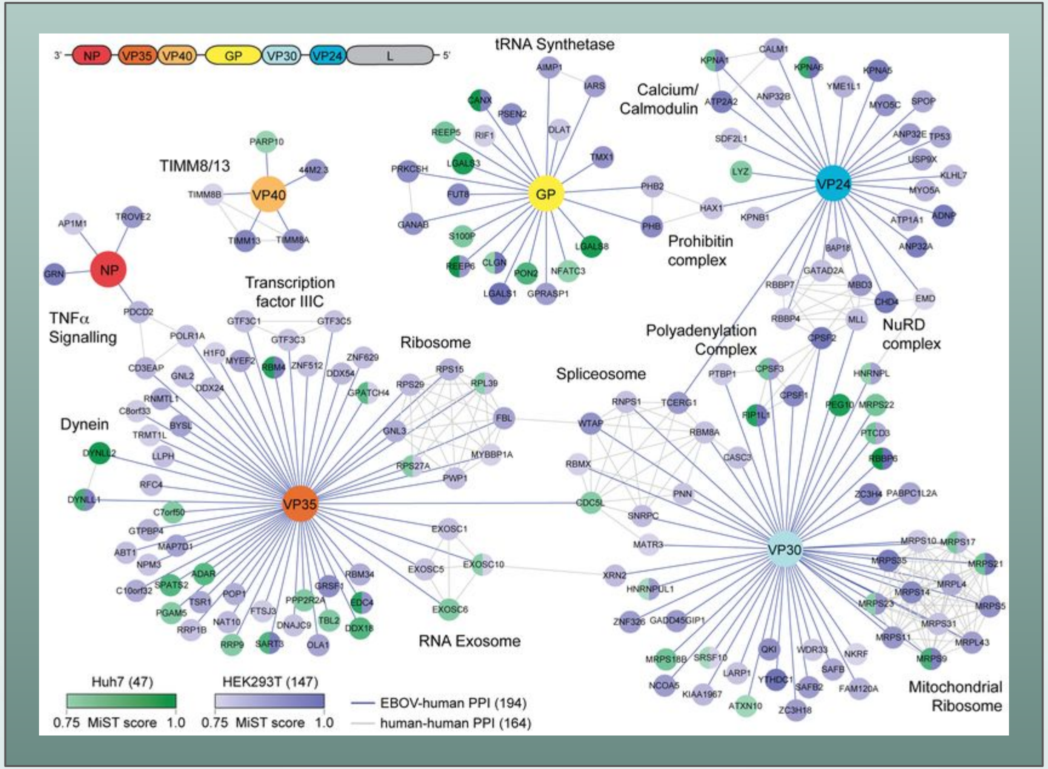


Primary Protein Structure
is sequence of a chain of amino acids
Amino Acids
Phe
Leu
Ser
Cys
Amino group
NH$_2$
H—C—COOH
R
R group
Acidic carboxyl group
Amino Acid

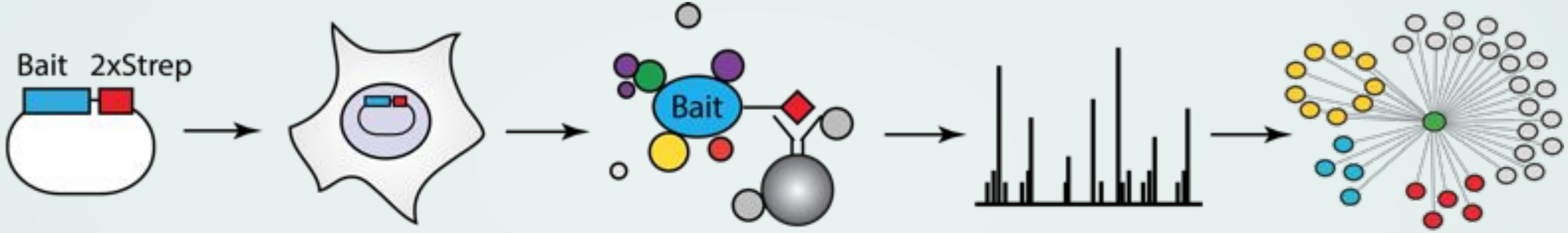


DNA
Gene
mRNA
Protein
ncRNA
Function
Function

PROTEIN-PROTEIN INTERACTION



Affinity Purification-Mass Spectrometry



Plasmid encoding tagged **bait** protein

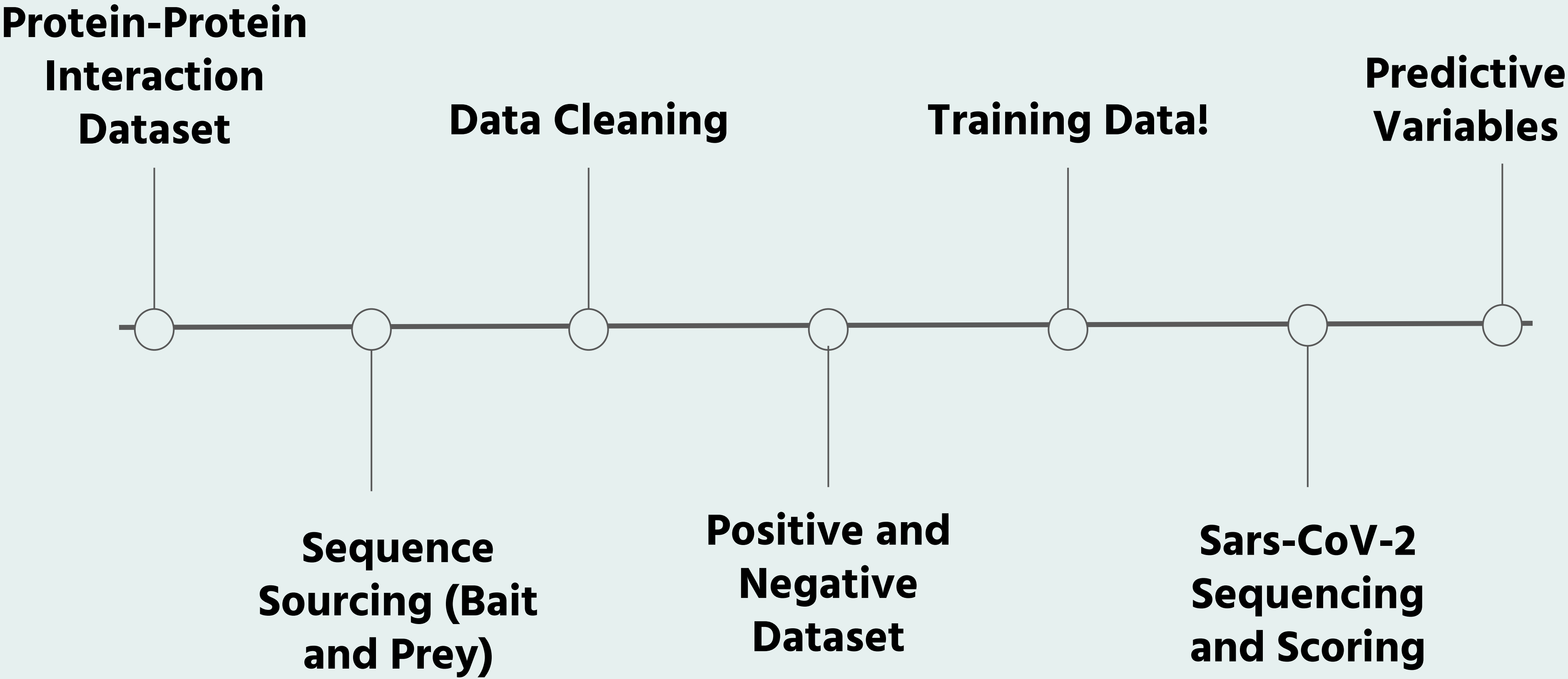
Expression in HEK293T cells

Affinity purification of **bait** and **prey** proteins

Mass spectrometry analysis

Data analysis

DATA



Protein-Protein Interaction Dataset

Sequence Sourcing (Bait and Prey)

Data Cleaning

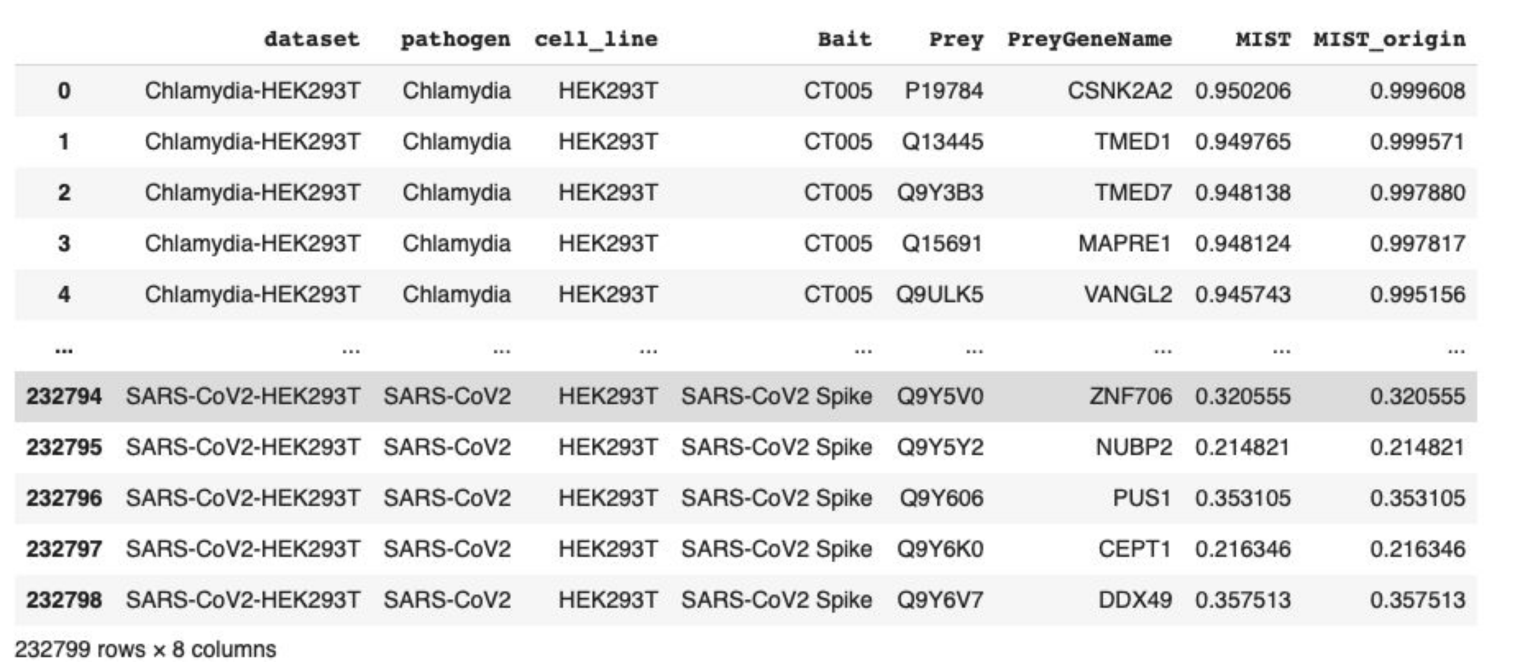
Positive and Negative Dataset

Training Data!

Sars-CoV-2 Sequencing and Scoring

Predictive Variables

PPI DATA

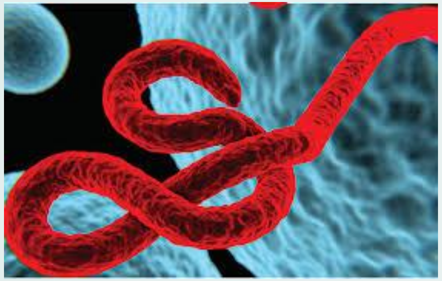


	dataset	pathogen	cell_line	Bait	Prey	PreyGeneName	MIST	MIST_origin
0	Chlamydia-HEK293T	Chlamydia	HEK293T	CT005	P19784	CSNK2A2	0.950206	0.999608
1	Chlamydia-HEK293T	Chlamydia	HEK293T	CT005	Q13445	TMED1	0.949765	0.999571
2	Chlamydia-HEK293T	Chlamydia	HEK293T	CT005	Q9Y3B3	TMED7	0.948138	0.997880
3	Chlamydia-HEK293T	Chlamydia	HEK293T	CT005	Q15691	MAPRE1	0.948124	0.997817
4	Chlamydia-HEK293T	Chlamydia	HEK293T	CT005	Q9ULK5	VANGL2	0.945743	0.995156
...	...	...	...	...	...	...	...	...
232794	SARS-CoV2-HEK293T	SARS-CoV2	HEK293T	SARS-CoV2 Spike	Q9Y5V0	ZNF706	0.320555	0.320555
232795	SARS-CoV2-HEK293T	SARS-CoV2	HEK293T	SARS-CoV2 Spike	Q9Y5Y2	NUBP2	0.214821	0.214821
232796	SARS-CoV2-HEK293T	SARS-CoV2	HEK293T	SARS-CoV2 Spike	Q9Y606	PUS1	0.353105	0.353105
232797	SARS-CoV2-HEK293T	SARS-CoV2	HEK293T	SARS-CoV2 Spike	Q9Y6K0	CEPT1	0.216346	0.216346
232798	SARS-CoV2-HEK293T	SARS-CoV2	HEK293T	SARS-CoV2 Spike	Q9Y6V7	DDX49	0.357513	0.357513

232799 rows × 8 columns

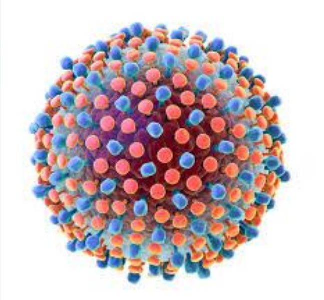
UTILIZING 6 DIFFERENT PATHOGENS

EBOLA

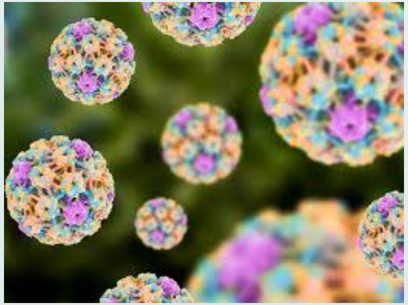


DENGUE

HCV

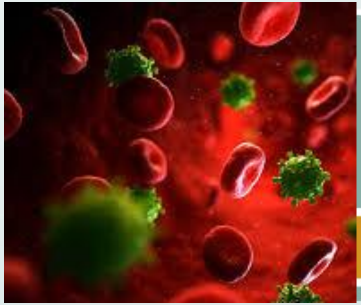


HPV



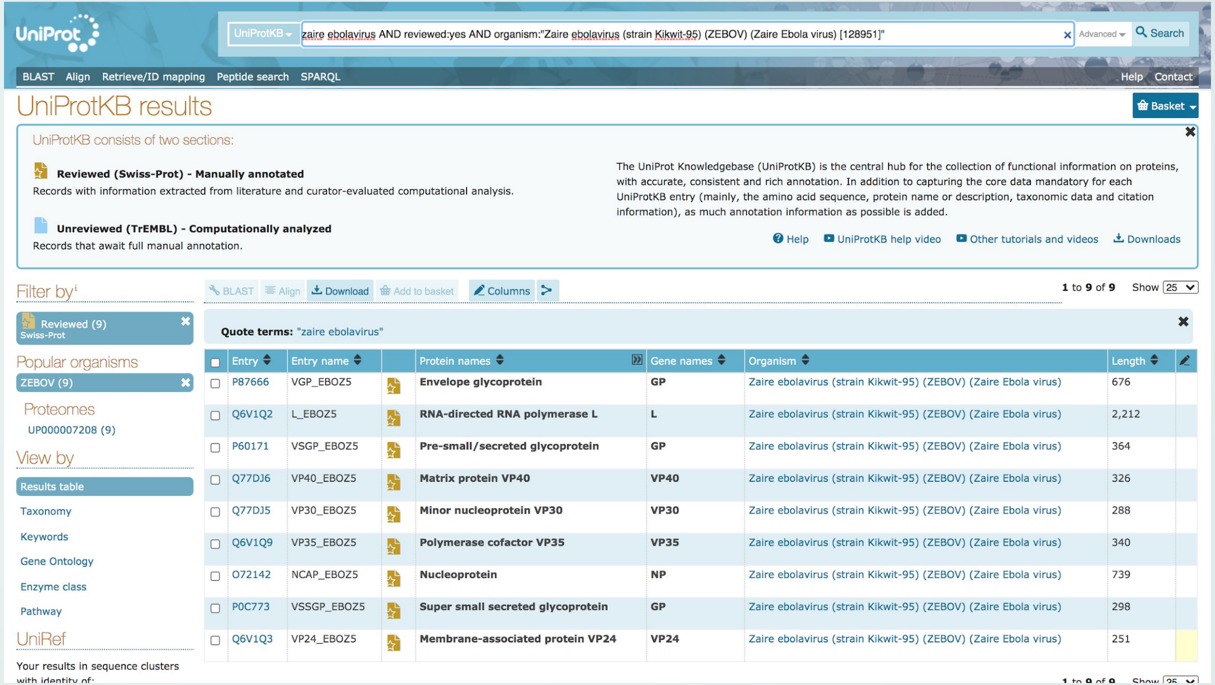
ZIKA

HIV



SEQUENCE SOURCING

We utilized fasta files containing protein ID and sequences for the diseases HIV, HCV, HPV Ebola, Dengue, Zika. To access the entire human proteome, we downloaded a fasta file from UniProt



zaire ebolavirus AND reviewed:yes AND organism:"Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus) [128951]"

UniProtKB results

UniProtKB consists of two sections:

Reviewed (Swiss-Prot) - Manually annotated
Records with information extracted from literature and curator-evaluated computational analysis.

Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

Filter by

Reviewed (9)
Swiss-Prot

Popular organisms

ZEBOV (9)

Proteomes

UP000007208 (9)

View by

Results table

Taxonomy

Keywords

Gene Ontology

Enzyme class

Pathway

UniRef

Your results in sequence clusters with identity of:

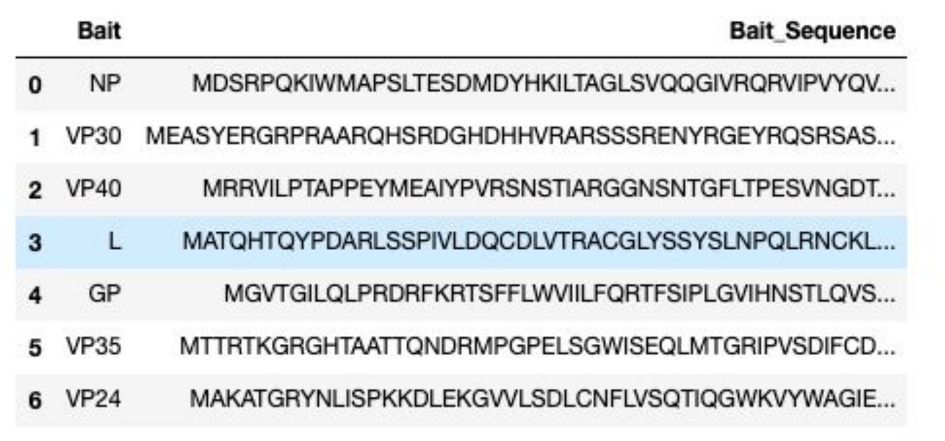
1 to 9 of 9

Quote terms: "zaire ebolavirus"

□	Entry	Entry name		Protein names	Gene names	Organism	Length
□	P87666	VGP_EBOZ5		**Envelope glycoprotein**	**GP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	676
□	Q6V1Q2	L_EBOZ5		**RNA-directed RNA polymerase L**	**L**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	2,212
□	P60171	VSGP_EBOZ5		**Pre-small/secreted glycoprotein**	**GP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	364
□	Q77DJ6	VP40_EBOZ5		**Matrix protein VP40**	**VP40**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	326
□	Q77DJ5	VP30_EBOZ5		**Minor nucleoprotein VP30**	**VP30**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	288
□	Q6V1Q9	VP35_EBOZ5		**Polymerase cofactor VP35**	**VP35**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	340
□	O72142	NCAP_EBOZ5		**Nucleoprotein**	**NP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	739
□	P0C773	VSSGP_EBOZ5		**Super small secreted glycoprotein**	**GP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	298
□	Q6V1Q3	VP24_EBOZ5		**Membrane-associated protein VP24**	**VP24**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	251

BAIT = PATHOGEN PROTEIN, PREY = HUMAN PROTEIN

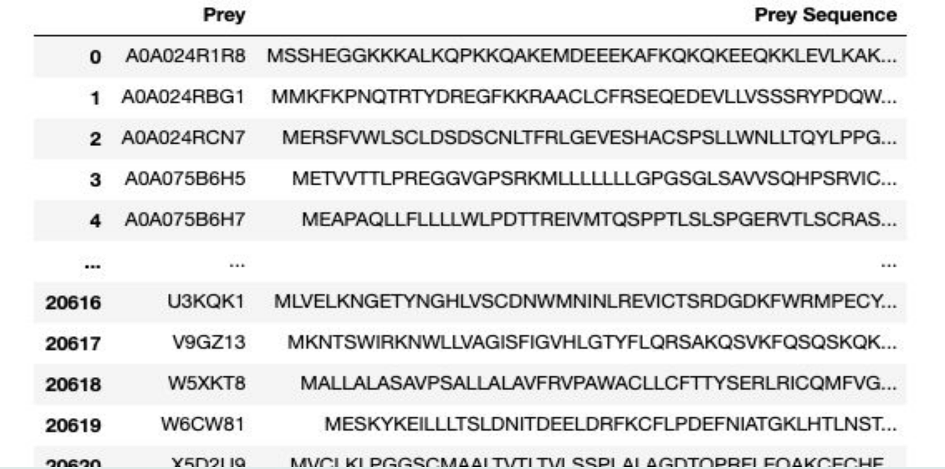
BAIT AND PREY



	Bait	Bait_Sequence
0	NP	MDSRPQKIWMAPSLTESDMDYHKILTAGLSVQQGIVRQRVIPVYQV...
1	VP30	MEASYERGRPRAARQHSRDGHDHHVRARSSSRENYRGEYRQSRSAS...
2	VP40	MRRVILPTAPPEYMEAIYPVRSNSTIARGGNSNTGFLTPESVNGDT...
3	L	MATQHTQYPDARLSSPIVLDQCDLVTRACGLYSSYSLNPQLRNCKL...
4	GP	MGVTGILQLPRDRFKRTSFFLWVIILFQRTFSIPLGVIHNSTLQVS...
5	VP35	MTTRTKGRGHTAATTQNDRMPGPELSGWISEQLMTGRIPVSDIFCD...
6	VP24	MAKATGRYNLISPKKDLEKGVVLSDLCNFLVSQTIQGWKVYWAGIE...

find proteins that interact

retrieve their amino acid sequence to study later on



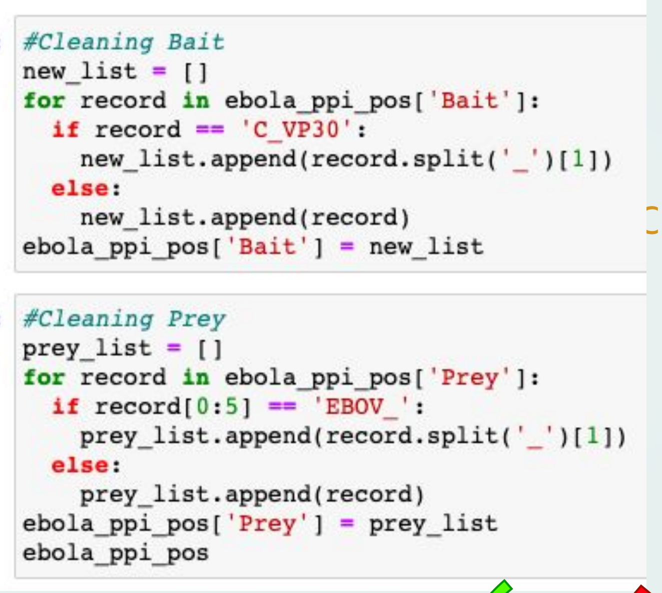
	Prey	Prey Sequence
0	A0A024R1R8	MSSHEGGKKKALKQPKKQAKEMDEEEKAFKQKQKEEQKKLEVLKAK...
1	A0A024RBG1	MMKFKPNQTRTYDREGFKKRAACLCFRSEQEDEVLLVSSSRYPDQW...
2	A0A024RCN7	MERSFVWLSCLDSDSCNLTFRLGEVESHACSPSLLWNLLTQYLPPG...
3	A0A075B6H5	METVVTTLPREGGVGPSRKMLLLLLLLGPGSGLSAVVSQHPSRVIC...
4	A0A075B6H7	MEAPAQLLFLLLLWLPDTTREIVMTQSPPTLSLSPGERVTLSCRAS...
...	...	...
20616	U3KQK1	MLVELKNGETYNGHLVSCDNWMNINLREVICTSRDGDKFWRMPECY...
20617	V9GZ13	MKNTSWIRKNWLLVAGISFIGVHLGTYFLQRSAKQSVKFQSQSKQK...
20618	W5XKT8	MALLALASAVPSALLALAVFRVPAWACLLCFTTYSERLRICQMFVG...
20619	W6CW81	MESKYKEILLLTSLDNITDEELDRFKCFLPDEFNIATGKLHTLNST...
20620	X5D2U9	MVCLKLPGGSCMAALTVTLTVLSSPLALAGDTQPRELEQAKCECHE

Data Cleaning

Uniformity

Eliminate mutations

Fill in NaN values for Prey

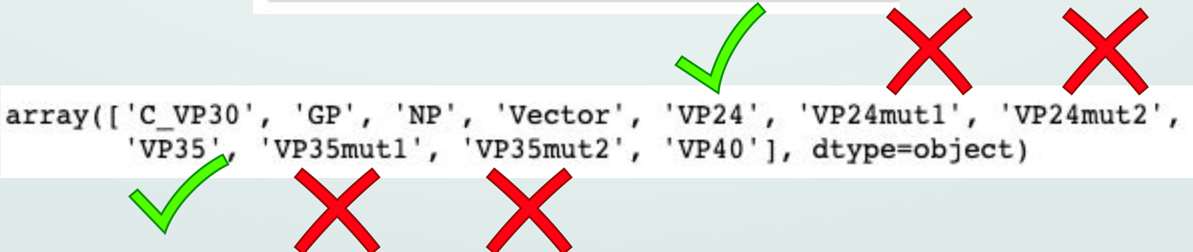


```
#Cleaning Bait
new_list = []
for record in ebola_ppi_pos['Bait']:
  if record == 'C_VP30':
    new_list.append(record.split('_')[1])
  else:
    new_list.append(record)
ebola_ppi_pos['Bait'] = new_list
```

```
#Cleaning Prey
prey_list = []
for record in ebola_ppi_pos['Prey']:
  if record[0:5] == 'EBOV_':
    prey_list.append(record.split('_')[1])
  else:
    prey_list.append(record)
ebola_ppi_pos['Prey'] = prey_list
ebola_ppi_pos
```

Ensure proteins came from people not mosquitos

Slice pathogen name



```
array(['C_VP30', 'GP', 'NP', 'Vector', 'VP24', 'VP24mut1', 'VP24mut2',
       'VP35', 'VP35mut1', 'VP35mut2', 'VP40'], dtype=object)
```

POSITIVE & NEGATIVE DATAFRAMES

POSITIVE

Criteria: MIST Score >= 0.75

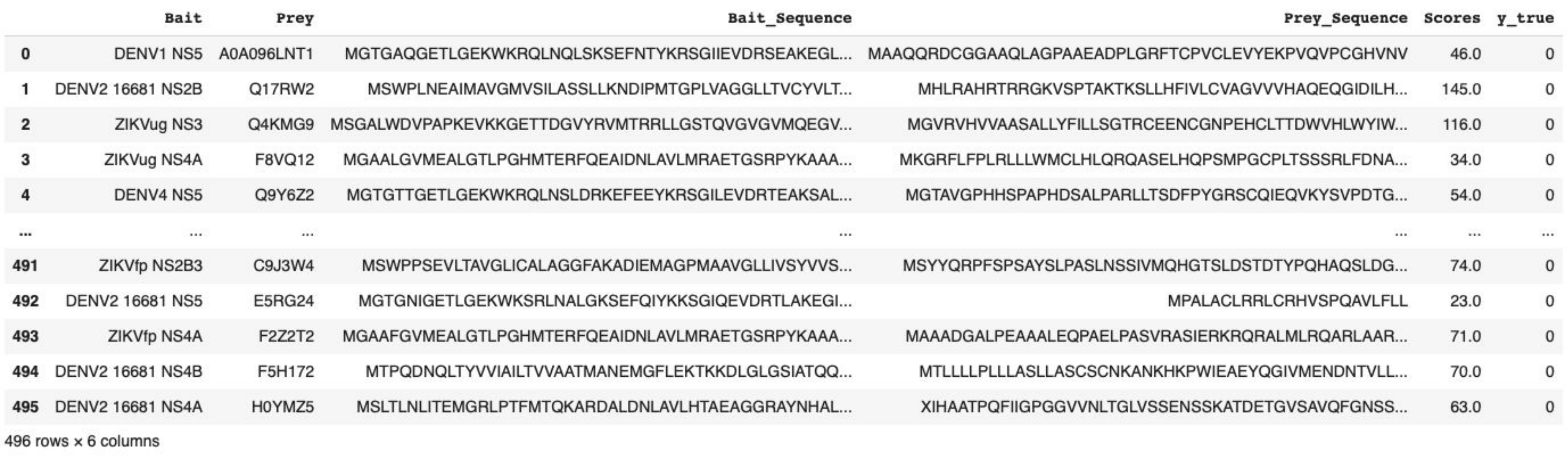
This dataframe's values were pulled from the original PPI dataset and also included the sequences of bait and prey, predictive variables that contribute to the classifier, and indication of a known interaction

NEGATIVE

Criteria: randomized bait and prey pairs

Random Bait and Prey were pulled from our pathogen and human proteome datasets to ensure that a known interaction was not guaranteed. Predictive variables were also included and indication was set to zero

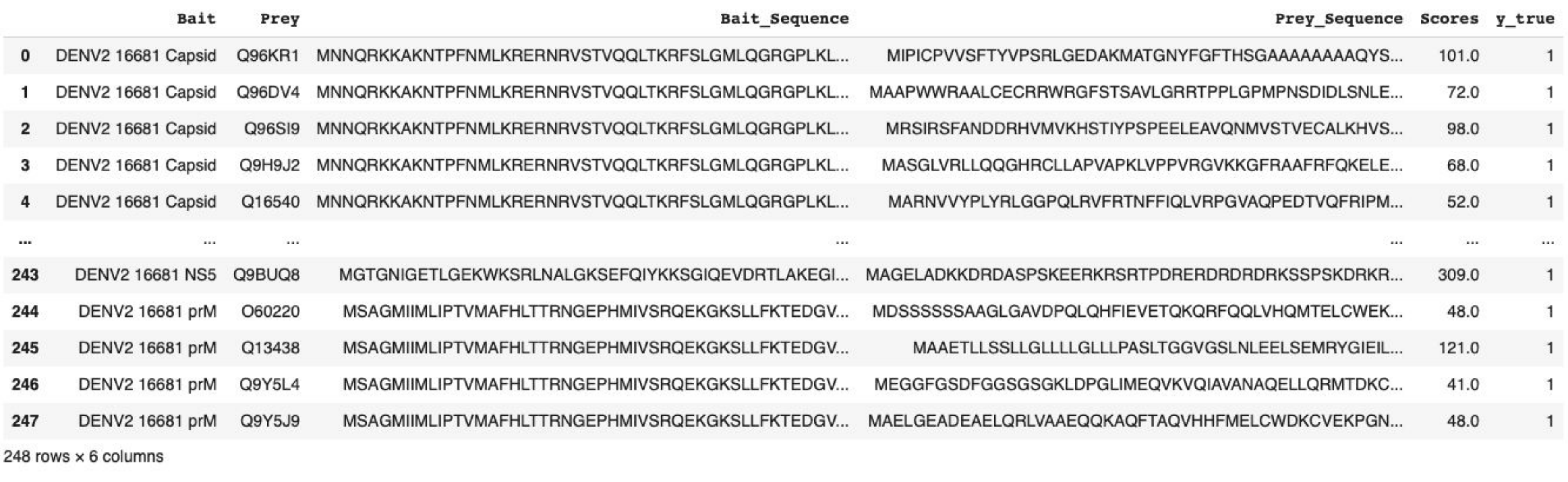
NEGATIVE DATASET



	Bait	Prey	Bait_Sequence	Prey_Sequence	Scores	y_true
0	DENV1 NS5	A0A096LNT1	MGTGAQGETLGEKWKRQLNQLSKSEFNTYKRSGIIEVDRSEAKEGL...	MAAQQRDCGGAAQLAGPAAEADPLGRFTCPVCLEVYEKPVQVPCGHVNV	46.0	0
1	DENV2 16681 NS2B	Q17RW2	MSWPLNEAIMAVGMVSILASSLLKNDIPMTGPLVAGGLLTVCYVLT...	MHLRAHRTRRGKVSPTAKTKSLLHFIVLCVAGVVVHAQEQGIDILH...	145.0	0
2	ZIKVug NS3	Q4KMG9	MSGALWDVPAPKEVKKGETTDGVYRVMTRRLLGSTQVGVGVMQEGV...	MGVRVHVVAASALLYFILLSGTRCEENCGNPEHCLTTDWVHLWYIW...	116.0	0
3	ZIKVug NS4A	F8VQ12	MGAALGVMEALGTLPGHMTERFQEAIDNLAVLMRAETGSRPYKAAA...	MKGRFLFPLRLLLWMCLHLQRQASELHQPSMPGCPLTSSSRLFDNA...	34.0	0
4	DENV4 NS5	Q9Y6Z2	MGTGTTGETLGEKWKRQLNSLDRKEFEEYKRSGILEVDRTEAKSAL...	MGTAVGPHHSPAPHDSALPARLLTSDFPYGRSCQIEQVKYSVPDTG...	54.0	0
...	...	...	...	...	...	...
491	ZIKVfp NS2B3	C9J3W4	MSWPPSEVLTAVGLICALAGGFAKADIEMAGPMAAVGLLIVSYVVS...	MSYYQRPFSPSAYSLPASLNSSIVMQHGTSLDSTDTYPQHAQSLDG...	74.0	0
492	DENV2 16681 NS5	E5RG24	MGTGNIGETLGEKWKSRLNALGKSEFQIYKKSGIQEVDRTLAKEGI...	MPALACLRRLCRHVSPQAVLFLL	23.0	0
493	ZIKVfp NS4A	F2Z2T2	MGAAFGVMEALGTLPGHMTERFQEAIDNLAVLMRAETGSRPYKAAA...	MAAADGALPEAAALEQPAELPASVRASIERKRQRALMLRQARLAAR...	71.0	0
494	DENV2 16681 NS4B	F5H172	MTPQDNQLTYVVIAILTVVAATMANEMGFLEKTKKDLGLGSIATQQ...	MTLLLLPLLLASLLASCSCNKANKHKPWIEAEYQGIVMENDNTVLL...	70.0	0
495	DENV2 16681 NS4A	H0YMZ5	MSLTLNLITEMGRLPTFMTQKARDALDNLAVLHTAEAGGRAYNHAL...	XIHAATPQFIIGPGGVVNLTGLVSSENSSKATDETGVSAVQFGNSS...	63.0	0

496 rows × 6 columns

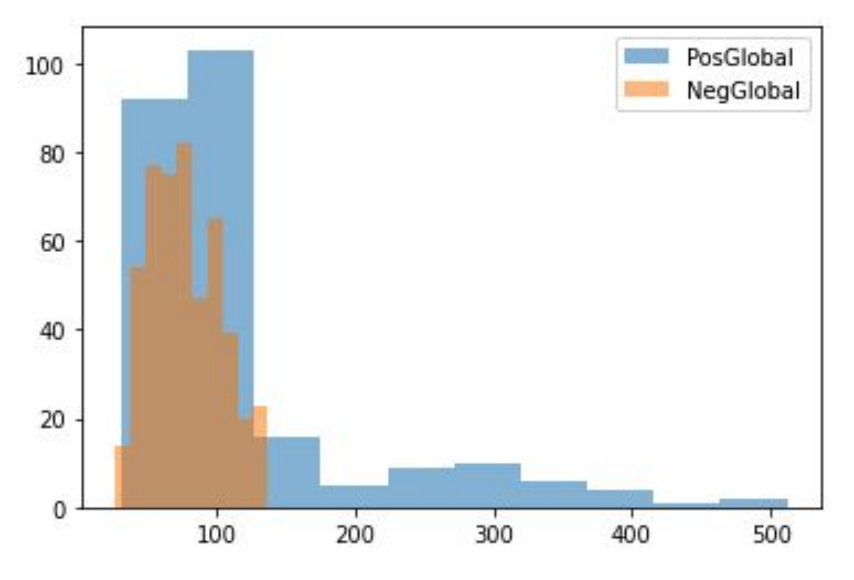
POSITIVE DATASET



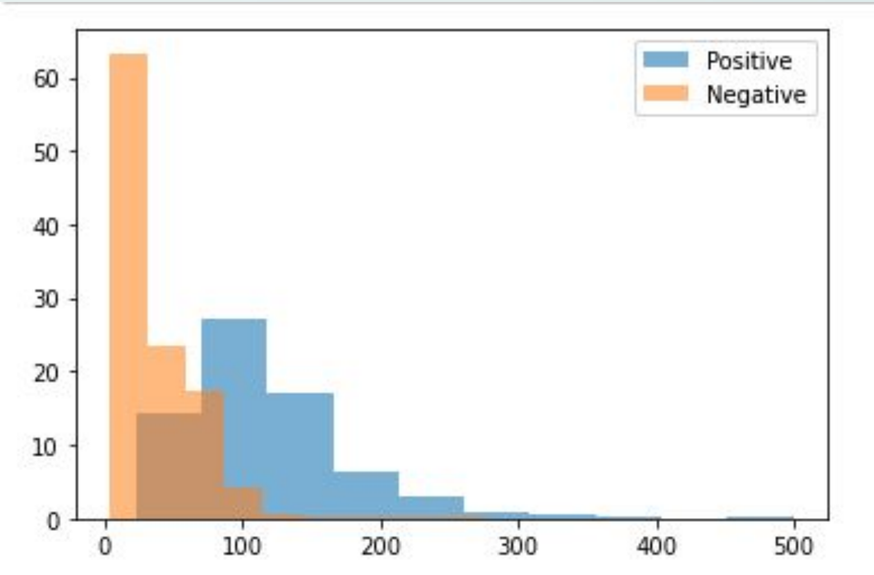
	Bait	Prey	Bait_Sequence	Prey_Sequence	Scores	y_true
0	DENV2 16681 Capsid	Q96KR1	MNNQRKKAKNTPFNMLKRERNRVSTVQQLTKRFSLGMLQGRGPLKL...	MIPICPVVSFTYVPSRLGEDAKMATGNYFGFTHSGAAAAAAAAQYS...	101.0	1
1	DENV2 16681 Capsid	Q96DV4	MNNQRKKAKNTPFNMLKRERNRVSTVQQLTKRFSLGMLQGRGPLKL...	MAAPWWRAALCECRRWRGFSTSAVLGRRTPPLGPMPNSDIDLSNLE...	72.0	1
2	DENV2 16681 Capsid	Q96SI9	MNNQRKKAKNTPFNMLKRERNRVSTVQQLTKRFSLGMLQGRGPLKL...	MRSIRSFANDDRHVMVKHSTIYPSPEELEAVQNMVSTVECALKHVS...	98.0	1
3	DENV2 16681 Capsid	Q9H9J2	MNNQRKKAKNTPFNMLKRERNRVSTVQQLTKRFSLGMLQGRGPLKL...	MASGLVRLLQQGHRCLLAPVAPKLVPPVRGVKKGFRAAFRFQKELE...	68.0	1
4	DENV2 16681 Capsid	Q16540	MNNQRKKAKNTPFNMLKRERNRVSTVQQLTKRFSLGMLQGRGPLKL...	MARNVVYPLYRLGGPQLRVFRTNFFIQLVRPGVAQPEDTVQFRIPM...	52.0	1
...	...	...	...	...	...	...
243	DENV2 16681 NS5	Q9BUQ8	MGTGNIGETLGEKWKSRLNALGKSEFQIYKKSGIQEVDRTLAKEGI...	MAGELADKKDRDASPSKEERKRSRTPDRERDRDRDRKSSPSKDRKR...	309.0	1
244	DENV2 16681 prM	O60220	MSAGMIIMLIPTVMAFHLTTRNGEPHMIVSRQEKGKSLLFKTEDGV...	MDSSSSSSAAGLGAVDPQLQHFIEVETQKQRFQQLVHQMTELCWEK...	48.0	1
245	DENV2 16681 prM	Q13438	MSAGMIIMLIPTVMAFHLTTRNGEPHMIVSRQEKGKSLLFKTEDGV...	MAAETLLSSLLGLLLLGLLLPASLTGGVGSLNLEELSEMRYGIEIL...	121.0	1
246	DENV2 16681 prM	Q9Y5L4	MSAGMIIMLIPTVMAFHLTTRNGEPHMIVSRQEKGKSLLFKTEDGV...	MEGGFGSDFGGSGSGKLDPGLIMEQVKVQIAVANAQELLQRMTDKC...	41.0	1
247	DENV2 16681 prM	Q9Y5J9	MSAGMIIMLIPTVMAFHLTTRNGEPHMIVSRQEKGKSLLFKTEDGV...	MAELGEADEAELQRLVAAEQQKAQFTAQVHHFMELCWDKCVEKPGN...	48.0	1

248 rows × 6 columns

THIS WAS A PROCESS...

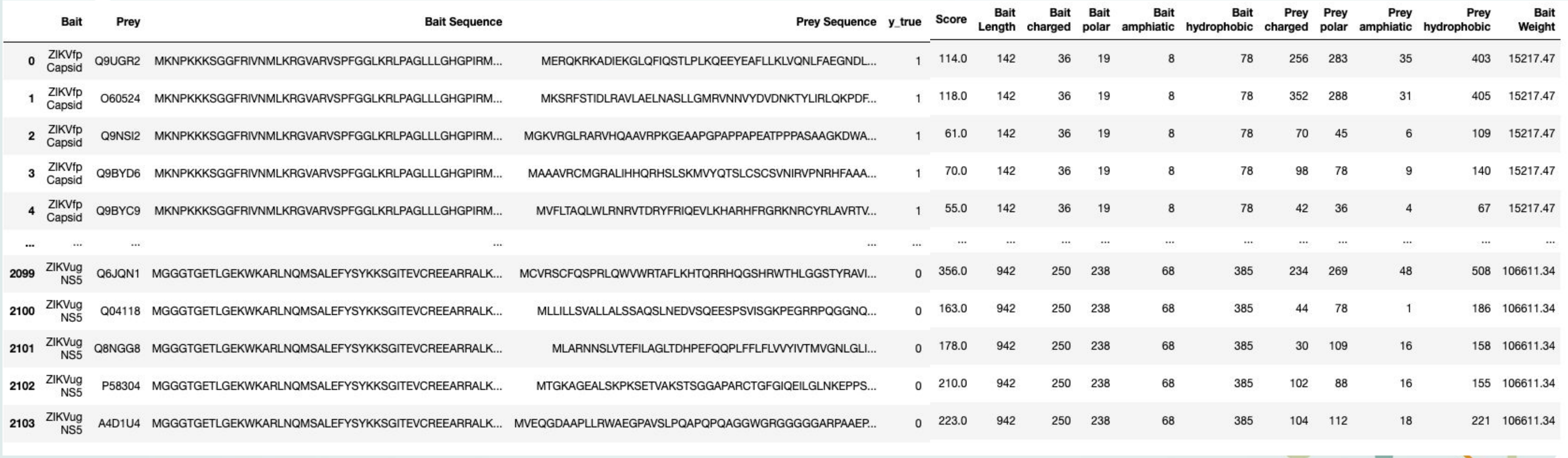


BEFORE



AFTER

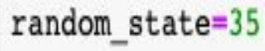
THE PRODUCT: TRAINING DATASET



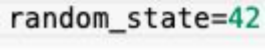
	Bait	Prey	Bait Sequence	Prey Sequence	y_true	Score	Bait Length	Bait charged	Bait polar	Bait amphiatic	Bait hydrophobic	Prey charged	Prey polar	Prey amphiatic	Prey hydrophobic	Bait Weight
0	ZIKVfp Capsid	Q9UGR2	MKNPKKKSGGFRIVNMLKRGVARVSPFGGLKRLPAGLLLGHGPIRM...	MERQKRKADIEKGLQFIQSTLPLKQEEYEAFLLKLVQNLFAEGNDL...	1	114.0	142	36	19	8	78	256	283	35	403	15217.47
1	ZIKVfp Capsid	O60524	MKNPKKKSGGFRIVNMLKRGVARVSPFGGLKRLPAGLLLGHGPIRM...	MKSRFSTIDLRAVLAELNASLLGMRVNNVYDVDNKTYLIRLQKPDF...	1	118.0	142	36	19	8	78	352	288	31	405	15217.47
2	ZIKVfp Capsid	Q9NSI2	MKNPKKKSGGFRIVNMLKRGVARVSPFGGLKRLPAGLLLGHGPIRM...	MGKVRGLRARVHQAAVRPKGEAAPGPAPPAPEATPPPASAAGKDWA...	1	61.0	142	36	19	8	78	70	45	6	109	15217.47
3	ZIKVfp Capsid	Q9BYD6	MKNPKKKSGGFRIVNMLKRGVARVSPFGGLKRLPAGLLLGHGPIRM...	MAAAVRCMGRALIHHQRHSLSKMVYQTSLCSCSVNIRVPNRHFAAA...	1	70.0	142	36	19	8	78	98	78	9	140	15217.47
4	ZIKVfp Capsid	Q9BYC9	MKNPKKKSGGFRIVNMLKRGVARVSPFGGLKRLPAGLLLGHGPIRM...	MVFLTAQLWLRNRVTDRYFRIQEVLKHARHFRGRKNRCYRLAVRTV...	1	55.0	142	36	19	8	78	42	36	4	67	15217.47
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2099	ZIKVug NS5	Q6JQN1	MGGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALK...	MCVRSCFQSPRLQWVWRTAFLKHTQRRHQGSHRWTHLGGSTYRAVI...	0	356.0	942	250	238	68	385	234	269	48	508	106611.34
2100	ZIKVug NS5	Q04118	MGGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALK...	MLLILLSVALLALSSAQSLNEDVSQEESPSVISGKPEGRRPQGGNQ...	0	163.0	942	250	238	68	385	44	78	1	186	106611.34
2101	ZIKVug NS5	Q8NGG8	MGGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALK...	MLARNNSLVTEFILAGLTDHPEFQQPLFFLFLVVYIVTMVGNLGLI...	0	178.0	942	250	238	68	385	30	109	16	158	106611.34
2102	ZIKVug NS5	P58304	MGGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALK...	MTGKAGEALSKPKSETVAKSTSGGAPARCTGFGIQEILGLNKEPPS...	0	210.0	942	250	238	68	385	102	88	16	155	106611.34
2103	ZIKVug NS5	A4D1U4	MGGGTGETLGEKWKARLNQMSALEFYSYKKSGITEVCREEARRALK...	MVEQGDAAPLLRWAEGPAVSLPQAPQPQAGGWGRGGGGGARPAAEP...	0	223.0	942	250	238	68	385	104	112	18	221	106611.34

Randomize, Balance, Split

- Randomize the positive and negative values into a mixed state



```
random_state=35
```



```
random_state=42
```

- Balance vs Unbalance dataset depending on how much your negative dataset interacts

Ratios: 1P : 2N and 1P : 1N

- Split the dataset into training and testing values for our input X, and our output Y

80% train

20% test

Input and Output

X_train, X_test

- Bait weight
- Bait length
- Number of Amino Acids with a Charge
- # of Polar Amino Acids
- # of Amphipathic Amino Acids
- # of Hydrophobic Amino Acids

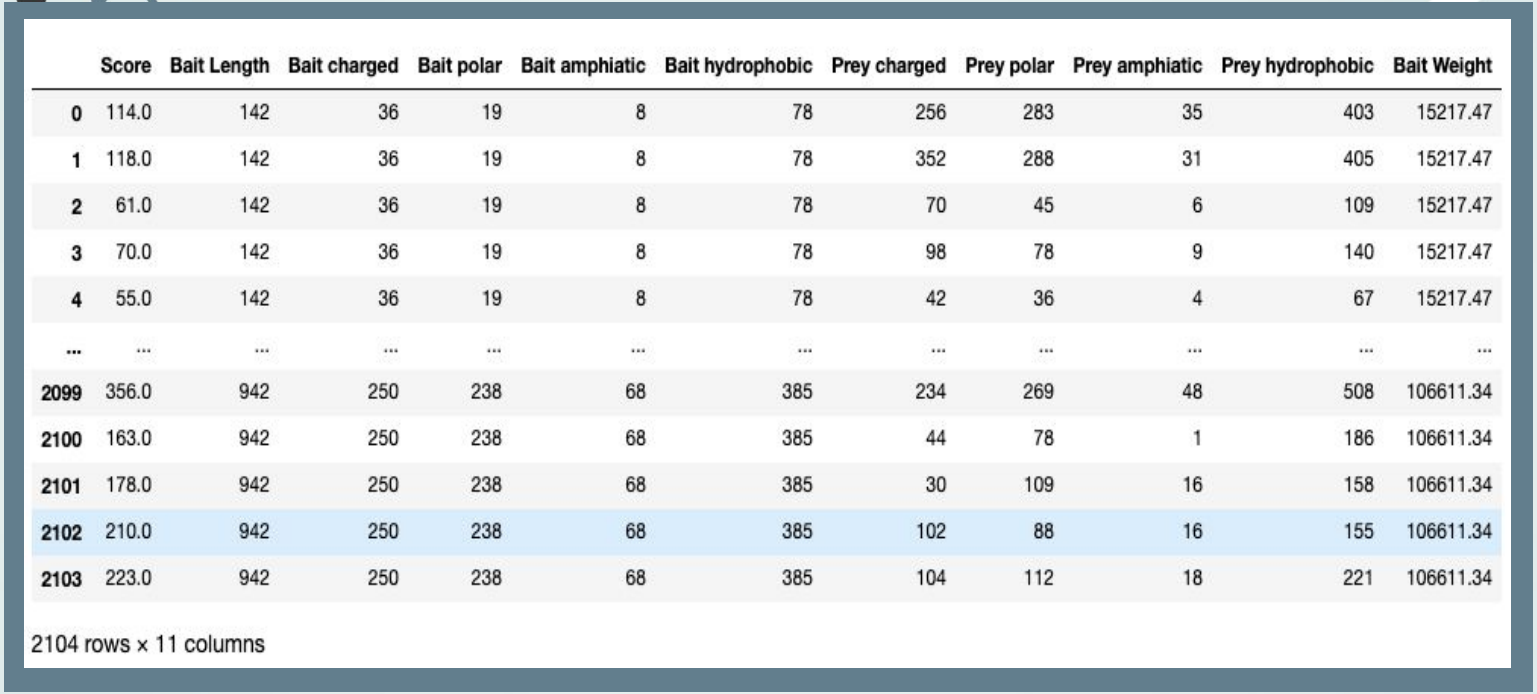
Y_train, Y_test

- y_True

Do the proteins interact or not?

THE CLASSIFIER

Data Features



	Score	Bait Length	Bait charged	Bait polar	Bait amphiatic	Bait hydrophobic	Prey charged	Prey polar	Prey amphiatic	Prey hydrophobic	Bait Weight
0	114.0	142	36	19	8	78	256	283	35	403	15217.47
1	118.0	142	36	19	8	78	352	288	31	405	15217.47
2	61.0	142	36	19	8	78	70	45	6	109	15217.47
3	70.0	142	36	19	8	78	98	78	9	140	15217.47
4	55.0	142	36	19	8	78	42	36	4	67	15217.47
...	...	...	...	...	...	...	...	...	...	...	...
2099	356.0	942	250	238	68	385	234	269	48	508	106611.34
2100	163.0	942	250	238	68	385	44	78	1	186	106611.34
2101	178.0	942	250	238	68	385	30	109	16	158	106611.34
2102	210.0	942	250	238	68	385	102	88	16	155	106611.34
2103	223.0	942	250	238	68	385	104	112	18	221	106611.34

2104 rows × 11 columns

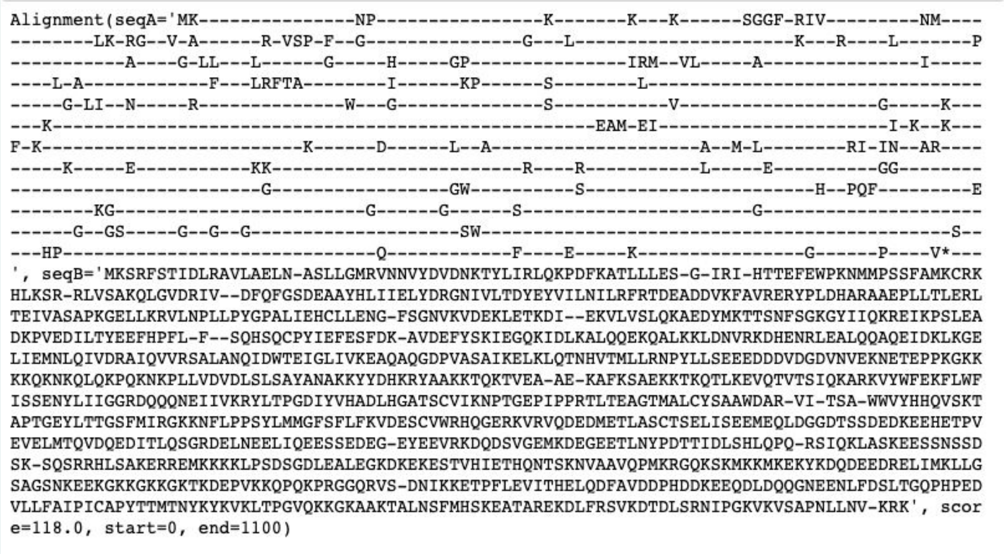
Example Protein Seq:

MKNPKKKSGGFRIVNMLKRGVARVNPLGGLKRLPAGLLLGHGPIRMVLAILAFLRFTAIKPSLGLINRWGSVGKKEAMEIIKKFKKDLAAMLRIINAR
KERKRRLEGGGGWSHPQFEKGGGSGGGSGGGSWSHPQFEKGPV

Alignment Score

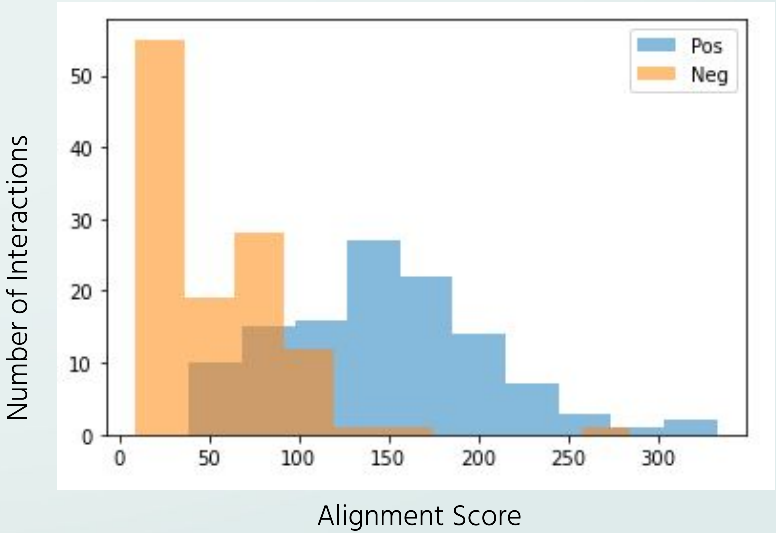
- Protein Sequence Alignment is used to find regions of similarity between two proteins
- We calculated the alignment score between each bait and prey protein and used it as a feature of the data
- Used a Global Pairwise Alignment from the BioPython module

Sequence Alignment Output



```
Alignment(seqA='MK---------------NP----------------K-------K---K------SGGF-RIV---------NM----
--------LK-RG--V-A------R-VSP-F--G---------------G---L---------------------K---R----L-------P
-----------A----G-LL---L------G-----H-----GP---------------IRM--VL-----A---------------I-----
----L-A------------F---LRFTA--------I------KP------S--------L--------------------------------
-----G-LI--N-----R--------------W---G--------------S-----------V-------------------G-----K---
---K----------------------------------------------------EAM-EI----------------------I-K--K---
F-K-------------------------K------D------L--A--------------------A--M-L--------RI-IN--AR----
-----K-----E-----------KK------------------------R----R-----------L-----E----------GG--------
------------------------G-----------------GW----------S----------------------H--PQF---------E
--------KG------------------------G------G------S----------------------G---------------------
------G--GS-----G--G--G--------------------SW---------------------------------------------S--
---HP------------------------------Q------------F----E-----K----------------G------P----V*---
', seqB='MKSRFSTIDLRAVLAELN-ASLLGMRVNNVYDVDNKTYLIRLQKPDFKATLLLES-G-IRI-HTTEFEWPKNMMPSSFAMKCRK
HLKSR-RLVSAKQLGVDRIV--DFQFGSDEAAYHLIIELYDRGNIVLTDYEYVILNILRFRTDEADDVKFAVRERYPLDHARAAEPLLTLERL
TEIVASAPKGELLKRVLNPLLPYGPALIEHCLLENG-FSGNVKVDEKLETKDI--EKVLVSLQKAEDYMKTTSNFSGKGYIIQKREIKPSLEA
DKPVEDILTYEEFHPFL-F--SQHSQCPYIEFESFDK-AVDEFYSKIEGQKIDLKALQQEKQALKKLDNVRKDHENRLEALQQAQEIDKLKGE
LIEMNLQIVDRAIQVVRSALANQIDWTEIGLIVKEAQAQGDPVASAIKELKLQTNHVTMLLRNPYLLSEEEDDDVDGDVNVEKNETEPPKGKK
KKQKNKQLQKPQKNKPLLVDVDLSLSAYANAKKYYDHKRYAAKKTQKTVEA-AE-KAFKSAEKKTKQTLKEVQTVTSIQKARKVYWFEKFLWF
ISSENYLIIGGRDQQQNEIIVKRYLTPGDIYVHADLHGATSCVIKNPTGEPIPPRTLTEAGTMALCYSAAWDAR-VI-TSA-WWVYHHQVSKT
APTGEYLTTGSFMIRGKKNFLPPSYLMMGFSFLFKVDESCVWRHQGERKVRVQDEDMETLASCTSELISEEMEQLDGGDTSSDEDKEEHETPV
EVELMTQVDQEDITLQSGRDELNEELIQEESSEDEG-EYEEVRKDQDSVGEMKDEGEETLNYPDTTIDLSHLQPQ-RSIQKLASKEESSNSSD
SK-SQSRRHLSAKERREMKKKKLPSDSGDLEALEGKDKEKESTVHIETHQNTSKNVAAVQPMKRGQKSKMKKMKEKYKDQDEEDRELIMKLLG
SAGSNKEEKGKKGKKGKTKDEPVKKQPQKPRGGQRVS-DNIKKETPFLEVITHELQDFAVDDPHDDKEEQDLDQQGNEENLFDSLTGQPHPED
VLLFAIPICAPYTTMTNYKYKVKLTPGVQKKGKAAKTALNSFMHSKEATAREKDLFRSVKDTDLSRNIPGKVKVSAPNLLNV-KRK', scor
e=118.0, start=0, end=1100)
```

Ebola Alignment Scores Histogram

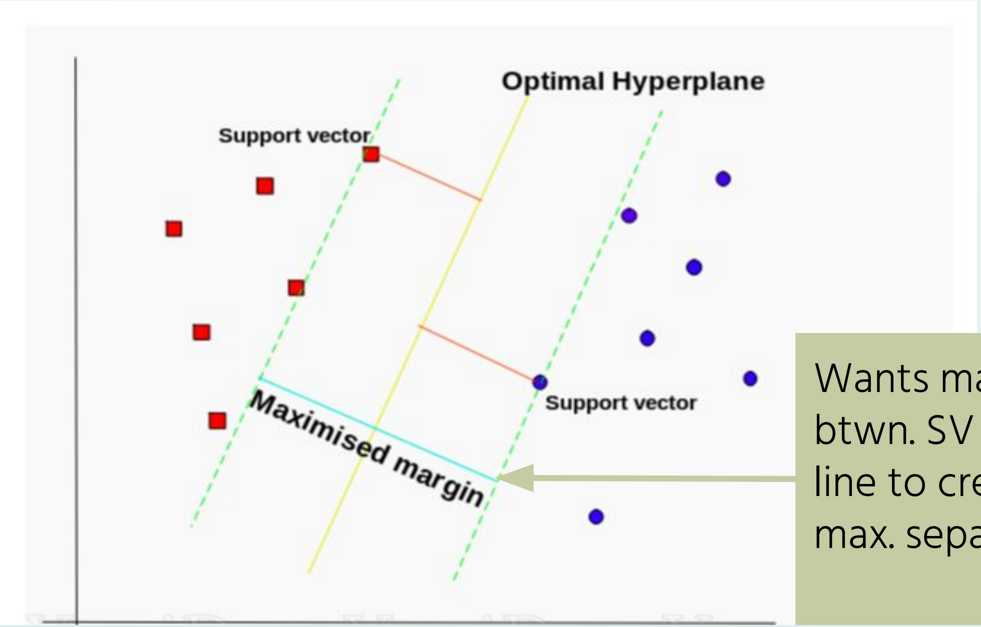


The Model!

- Supervised Learning: Data was labeled
- Binary Classification Problem: Wanted an output of 0 or 1, indicating whether or not the two sequences would interact
- Support Vector Machine!

How does the SVM work?

- Tries to find separating line for interacting and non interacting protein pairs
- To find the line that best separates data it uses SUPPORT VECTORS (SV)



Wants max dist. btwn. SV and line to create max. separation

Two Tuning Parameters:

- **C** = Want a smooth boundary or curved one to fit points?
- **Gamma** = Do you take in distance from close points or far ones too?

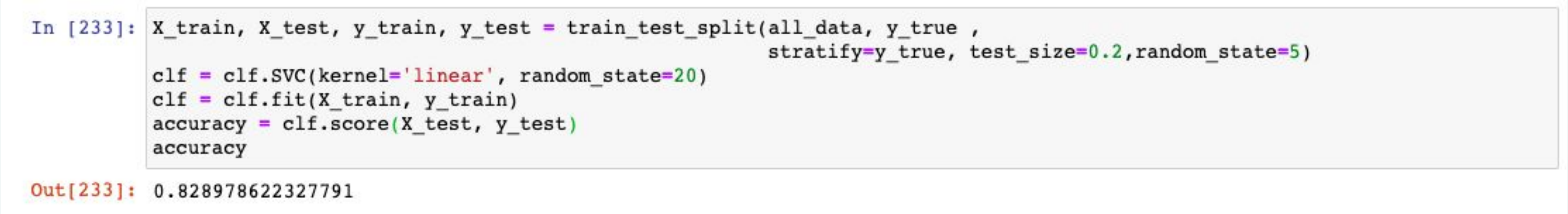
Training and Testing

Part One:

- Each person trained their model using 80% of their individual virus data
- Tested on the remaining 20%

Part Two:

- Trained using all of our individual virus data
- Used the model to predict the interaction between each SARS-CoV-2 protein and each human protein



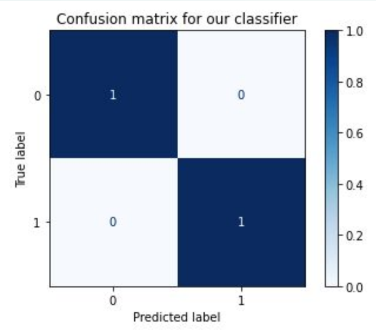
```
In [233]: X_train, X_test, y_train, y_test = train_test_split(all_data, y_true ,
                                                          stratify=y_true, test_size=0.2,random_state=5)
          clf = clf.SVC(kernel='linear', random_state=20)
          clf = clf.fit(X_train, y_train)
          accuracy = clf.score(X_test, y_test)
          accuracy

Out[233]: 0.828978622327791
```

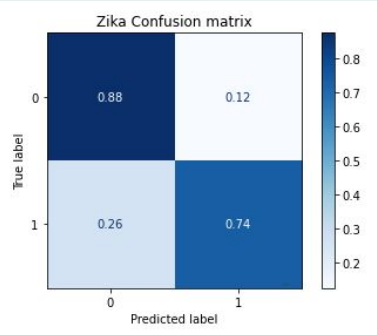
Part One: Training / Testing on Zika

RESULTS

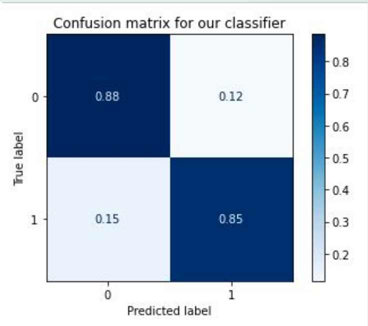
Model Test on Viruses



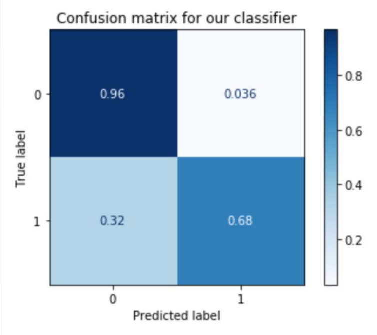
Dengue



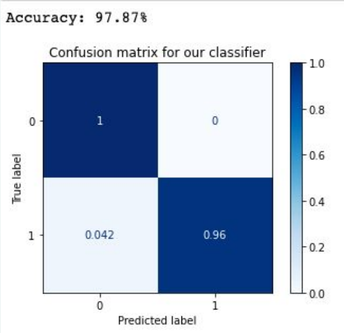
Zika



HCV



HIV



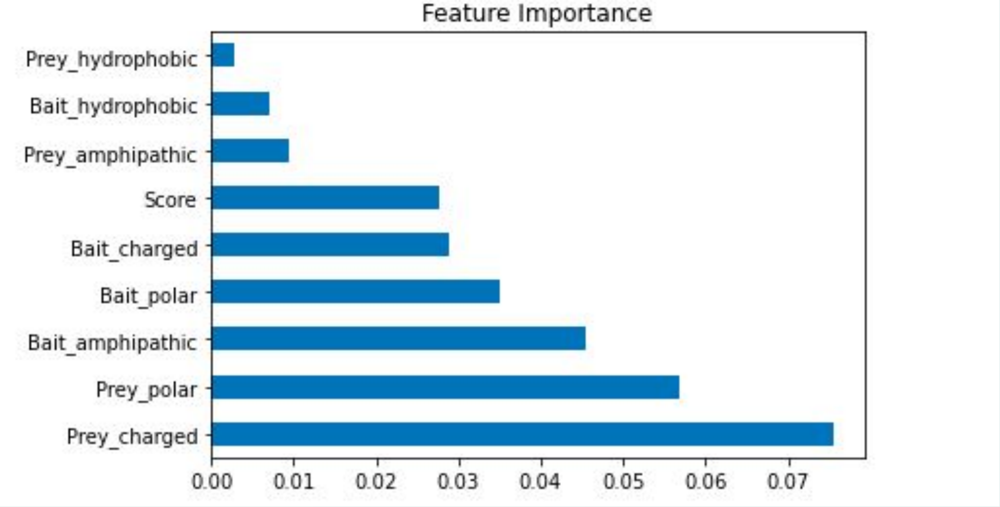
```
Accuracy: 97.87%
```

Ebola

Testing our model on SARS-CoV-2 PPI!

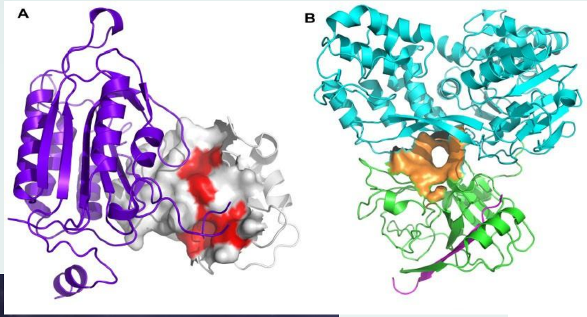
Ebola trained SVM tested on Sars-Cov-2 Accuracy:

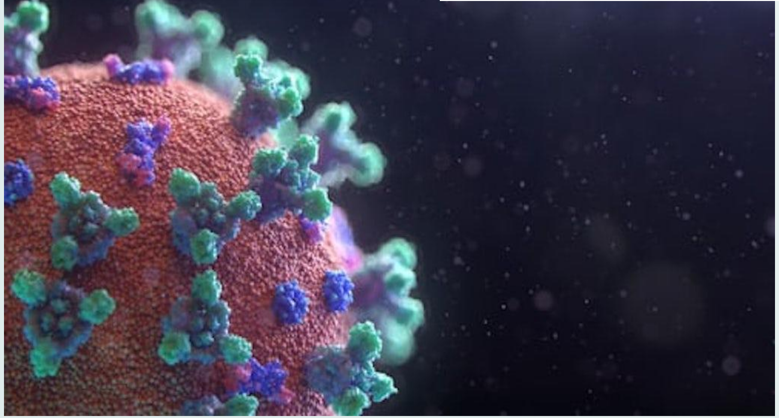
42.0%

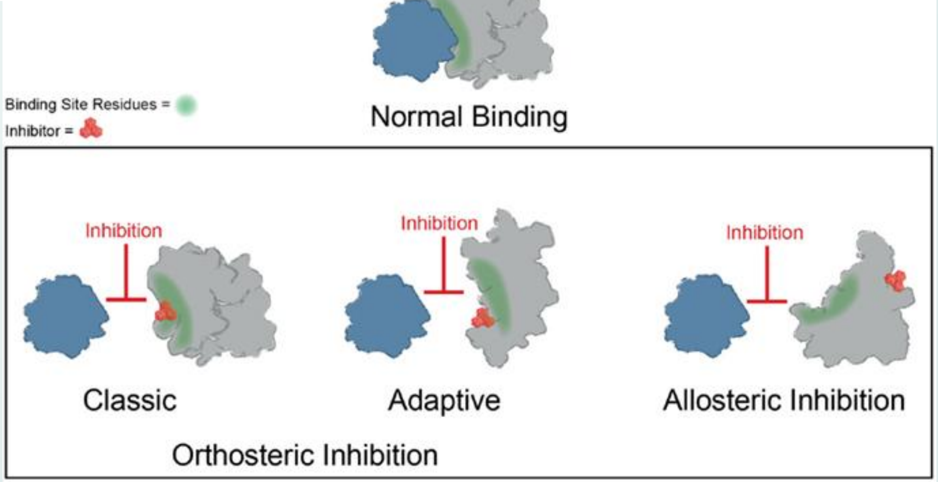


IMPACT

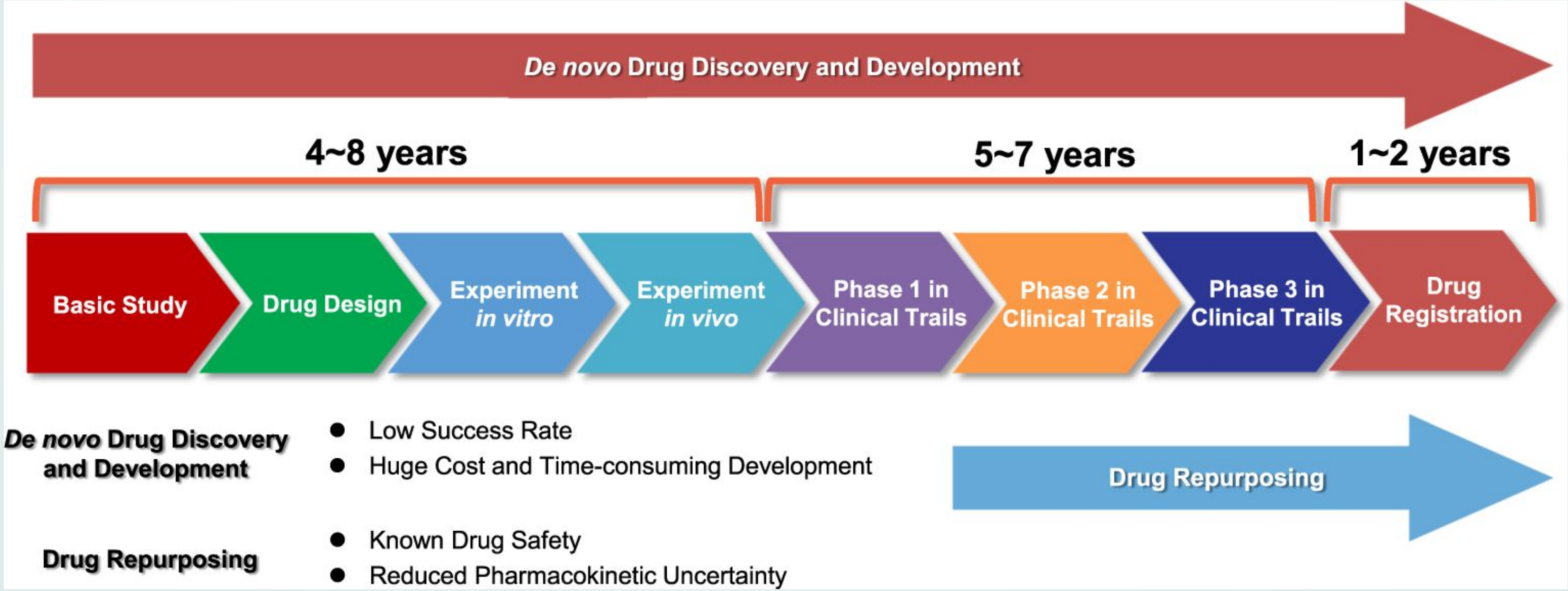
Impact of Experimental Results







Drug Repurposing



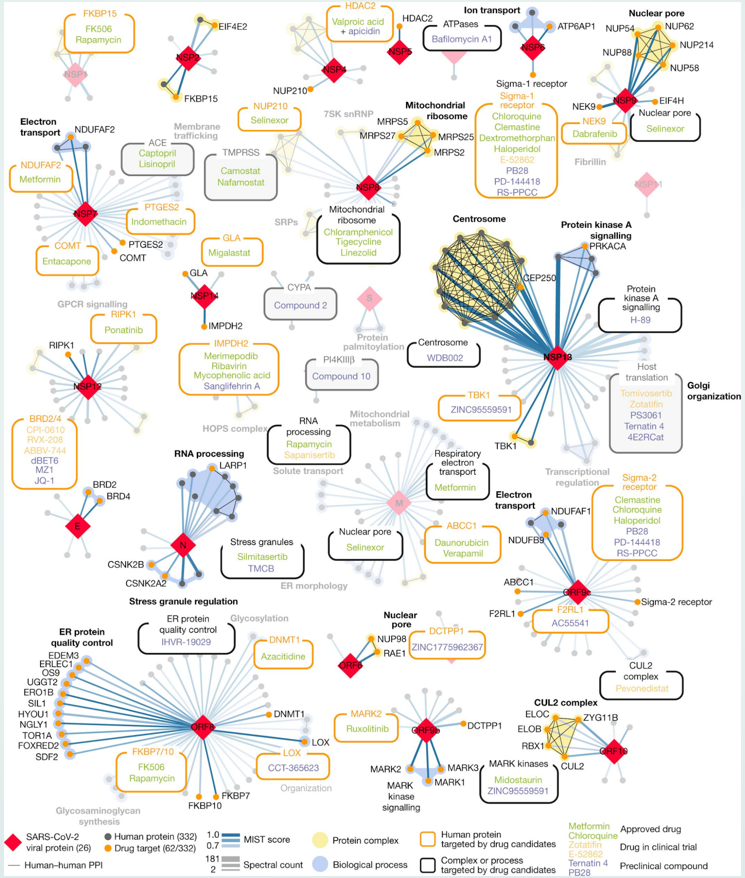
***De novo* Drug Discovery and Development**

- Low Success Rate
- Huge Cost and Time-consuming Development

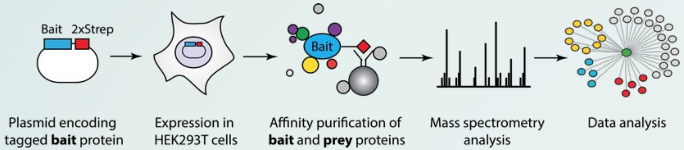
Drug Repurposing

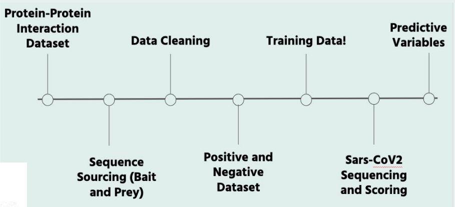
- Known Drug Safety
- Reduced Pharmacokinetic Uncertainty

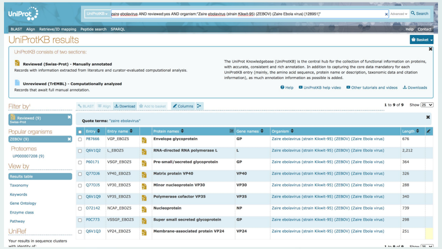
Drug-Human Target Network



Future Research & Closing Remarks


Bait
2xStrep
Plasmid encoding tagged **bait** protein
Expression in HEK293T cells
Bait
Affinity purification of **bait** and **prey** proteins
Mass spectrometry analysis
Data analysis


Protein-Protein Interaction Dataset
Data Cleaning
Training Data!
Predictive Variables
Sequence Sourcing (Bait and Prey)
Positive and Negative Dataset
Sars-CoV2 Sequencing and Scoring



zaire ebolavirus AND reviewed:yes AND organism:"Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus) [128951]"

UniProtKB results

UniProtKB consists of two sections:

Reviewed (Swiss-Prot) - Manually annotated
Records with information extracted from literature and curator-evaluated computational analysis.

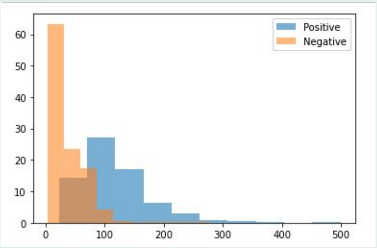
Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.

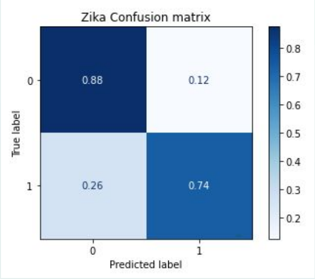
The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data and citation information), as much annotation information as possible is added.

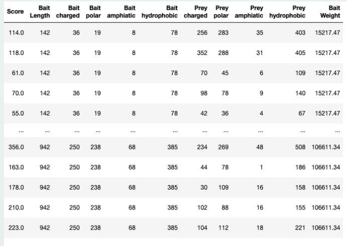
Your results in sequence clusters with identity of:

Quote terms: "zaire ebolavirus"

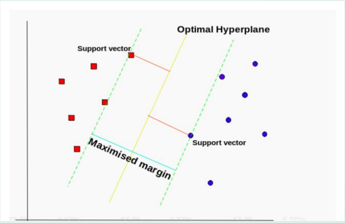
Entry	Entry name	Protein names	Gene names	Organism	Length
P87666	VGP_EBOZ5	**Envelope glycoprotein**	**GP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	676
Q6V1Q2	L_EBOZ5	**RNA-directed RNA polymerase L**	**L**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	2,212
P60171	VSGP_EBOZ5	**Pre-small/secreted glycoprotein**	**GP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	364
Q77DJ6	VP40_EBOZ5	**Matrix protein VP40**	**VP40**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	326
Q77DJ5	VP30_EBOZ5	**Minor nucleoprotein VP30**	**VP30**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	288
Q6V1Q9	VP35_EBOZ5	**Polymerase cofactor VP35**	**VP35**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	340
O72142	NCAP_EBOZ5	**Nucleoprotein**	**NP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	739
P0C773	VSSGP_EBOZ5	**Super small secreted glycoprotein**	**GP**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	298
Q6V1Q3	VP24_EBOZ5	**Membrane-associated protein VP24**	**VP24**	Zaire ebolavirus (strain Kikwit-95) (ZEBOV) (Zaire Ebola virus)	251

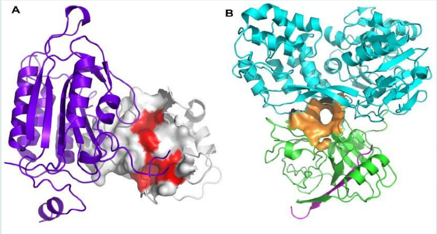


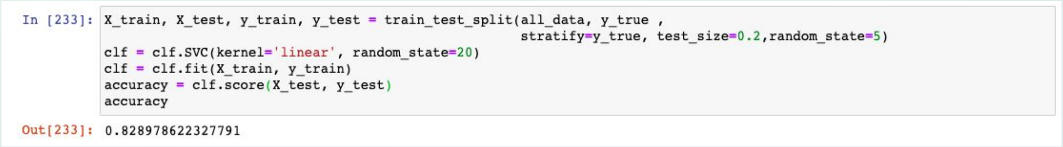




Score	Bait Length	Bait charged	Bait polar	Bait amphiatic	Bait hydrophobic	Prey charged	Prey polar	Prey amphiatic	Prey hydrophobic	Bait Weight
114.0	142	36	19	8	78	256	283	35	403	15217.47
118.0	142	36	19	8	78	352	288	31	405	15217.47
61.0	142	36	19	8	78	70	45	6	109	15217.47
70.0	142	36	19	8	78	98	78	9	140	15217.47
55.0	142	36	19	8	78	42	36	4	67	15217.47
...	...	...	...	...	...	...	...	...	...	...
356.0	942	250	238	68	385	234	269	48	508	106611.34
163.0	942	250	238	68	385	44	78	1	186	106611.34
178.0	942	250	238	68	385	30	109	16	158	106611.34
210.0	942	250	238	68	385	102	88	16	155	106611.34
223.0	942	250	238	68	385	104	112	18	221	106611.34







```
In [233]: X_train, X_test, y_train, y_test = train_test_split(all_data, y_true ,
                                                              stratify=y_true, test_size=0.2,random_state=5)
          clf = clf.SVC(kernel='linear', random_state=20)
          clf = clf.fit(X_train, y_train)
          accuracy = clf.score(X_test, y_test)
          accuracy

Out[233]: 0.828978622327791
```

Acknowledgements

UCSF AI4ALL Directors
Marina Sirota, PhD
Tomiko Oskotsky, MD

Lead TA
Snow Naing

Alumnae TA
Isha Karim

Student Researchers:
Ami Baid
Arhana Aatresh
Esha Gohil
Joyce Yang
Valerie Kwek
Yomn Hammad

Miscellaneous
Google
StackQuest Man (Josh)
StackOverflow
Stupid Questions

PPI

Questions?

Or comments…