Frederick National Laboratory
for Cancer Research

sponsored by the National Cancer Institute

COVID-19 Research and Proposal Ideas

BIDS
FNLCR

May 29, 2020

DEPARTMENT OF HEALTH AND HUMAN SERVICES • National Institutes of Health • National Cancer Institute

Frederick National Laboratory is a Federally Funded Research and Development Center operated by Leidos Biomedical Research, Inc., for the National Cancer Institute

Agenda/Project ideas for discussion

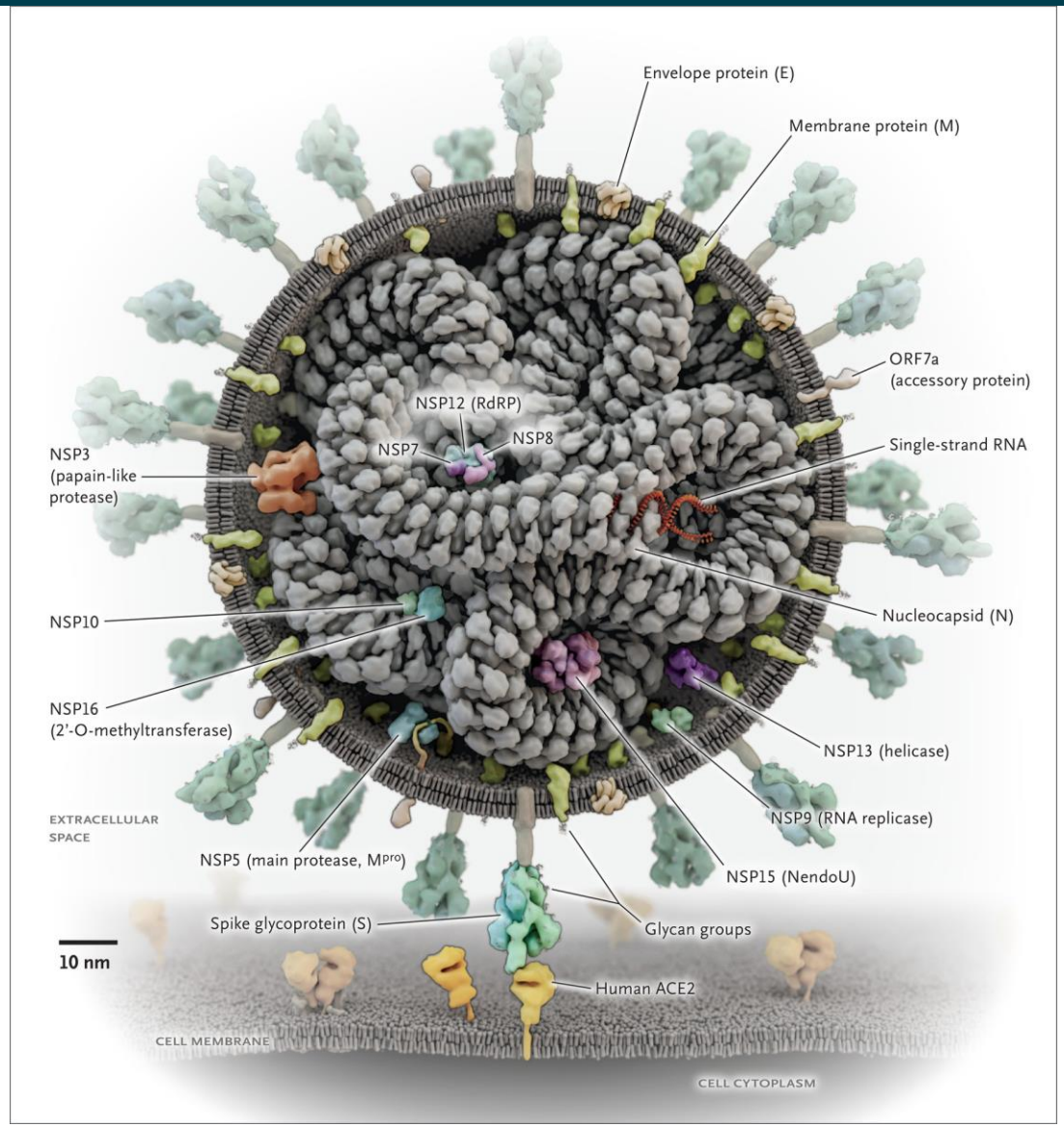
- **Preliminary information**
  - Introduction to the Sars Cov-2 System; genome description; protein information; vaccine development; data types/sources
- **Project-1 (a, b and c): Receptor-focused study (experimental and modeling)**
- **Project-2: Receptor flexibility and MD simulations to identify conformations**
- **Project-3 (3, 3a): Modeling Receptor-Drug interactions**
  - Docking simulations (may be using the conformations identified by Project-2)
  - Commercial/Open-source small-molecule libraries
  - Custom small-molecule library construction based on COVID-19 drug-discovery publications
- **Project-4: ML: Drug-Repurposing**
- **Project-5: Experimental assays for quantifying COVID-19 viral activity**

The System: SARS-CoV-2 Virion and Its Proteins

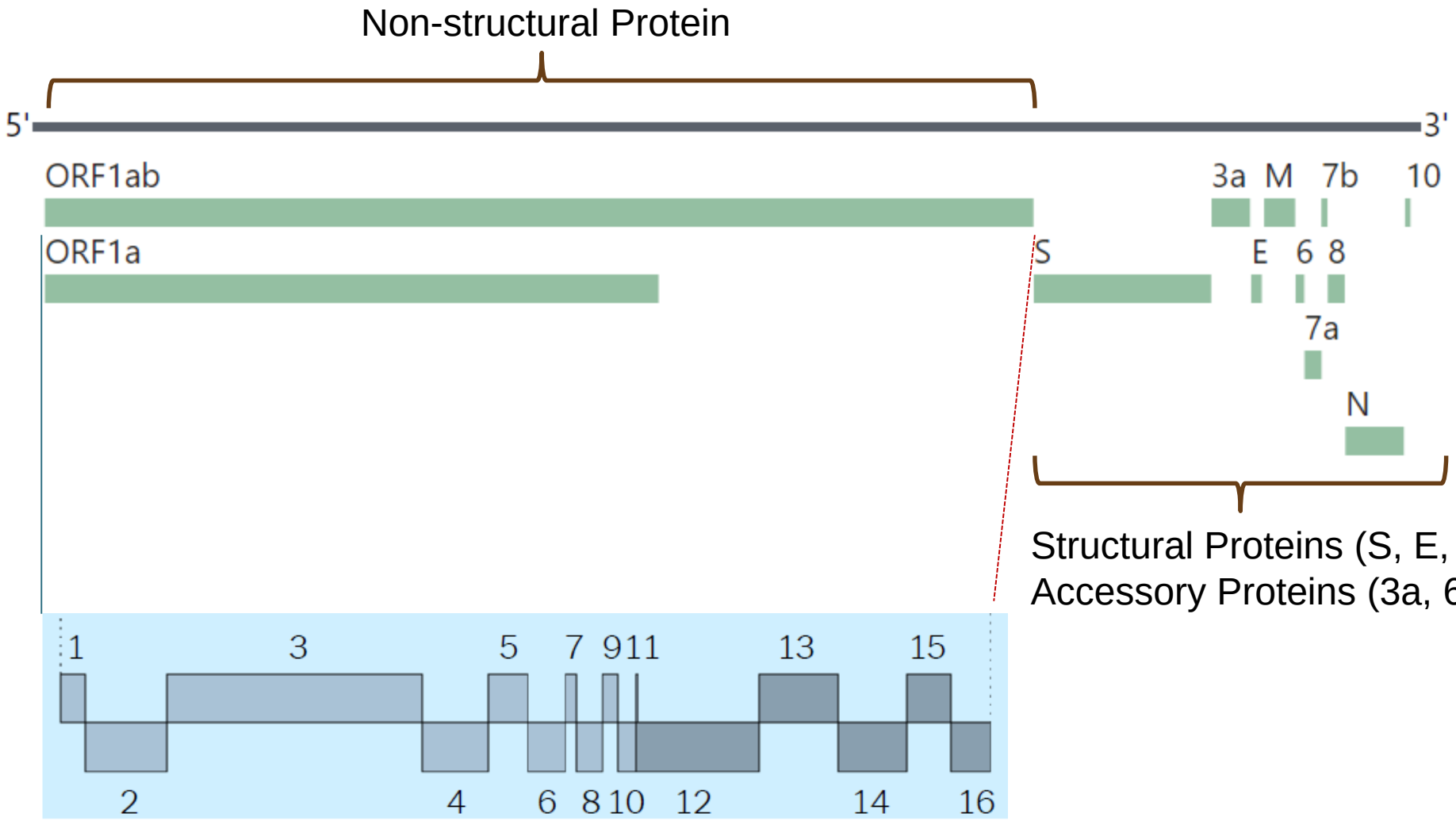
Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

The figure is from the following NEJM paper (figure 1)

https://www.nejm.org/doi/full/10.1056/NEJMcibr2007042



The System: SARS-CoV-2 Virion and Its Proteins



Structural Proteins (S, E, M, N)
Accessory Proteins (3a, 6, 7a/b, 8, 10)

What SARS-CoV-2 information is available?

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

- **Sequences**
  - NCBI
  - Ensembl
  - UCSC

https://www.ncbi.nlm.nih.gov/

https://covid-19.ensembl.org/index.html

https://genome.ucsc.edu/covid19.html

- **3D structures**
  - Experimental: PDB (US, Europe and Japan)
  - Modeling
    - I-TASSER
    - ROSETTA
    - SwissModel
    - Phyre2

https://www.rcsb.org/
https://www.ebi.ac.uk/pdbe/
https://pdbj.org/

https://zhanglab.ccmb.med.umich.edu/COVID-19/

https://www.rosettacommons.org/software

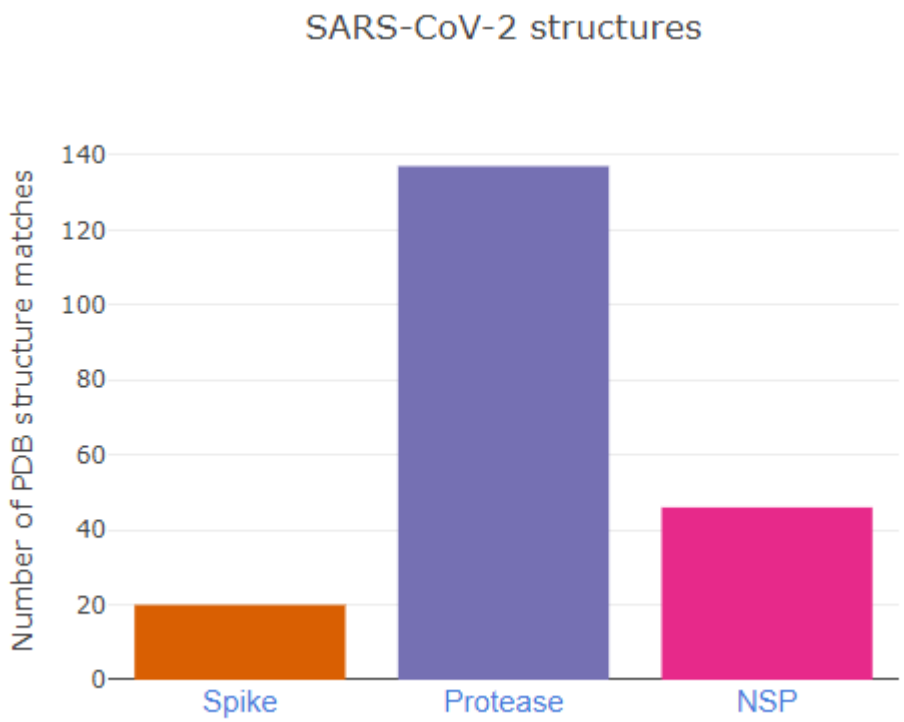
https://swissmodel.expasy.org/

http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index

What SARS-CoV-2 3D information is available?

- https://cov3d.ibbr.umd.edu

05/28/2020



Experimental 3D structure information is common to most projects

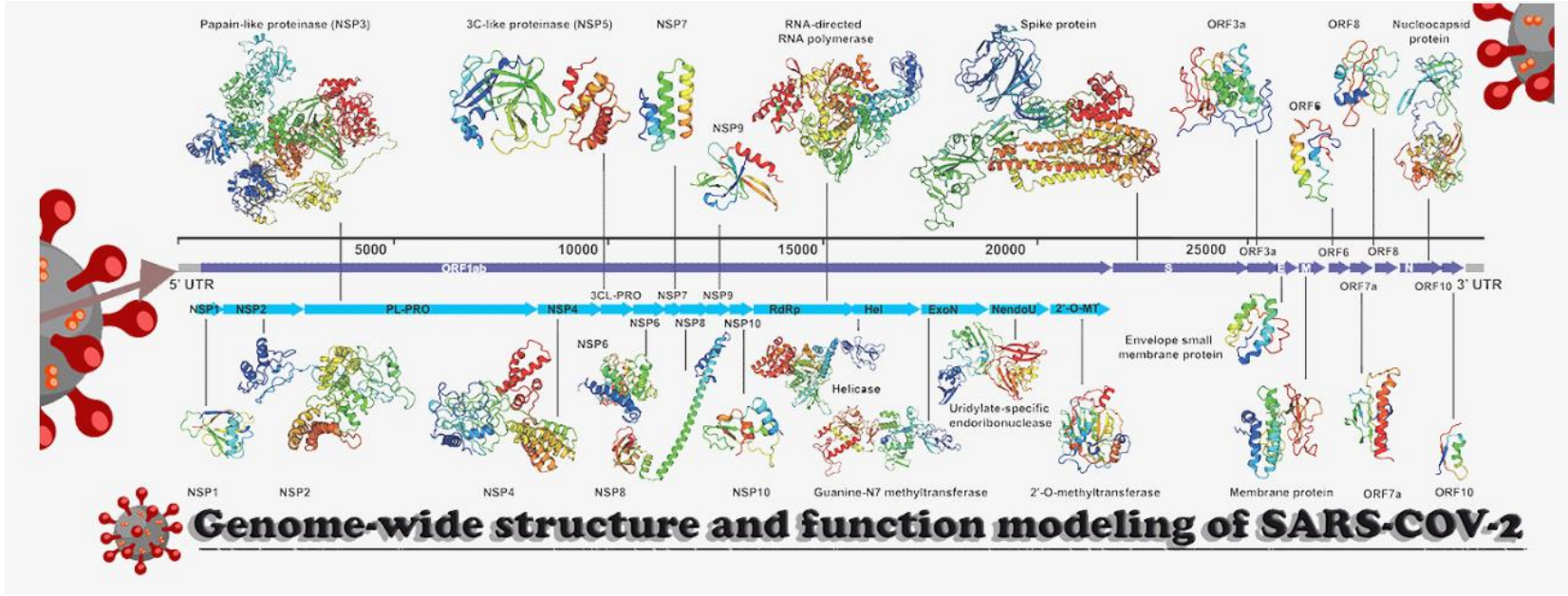
What if there are no 3D structures?

- **If experimental structures are not available. Structure-based modeling can help**

- **I-TASSER**
  - All COVID-19 protein structure homology models are available
  - Why choose I-TASSER over other modeling software?

- **SwissModel**
  - All COVID-19 protein structures are available

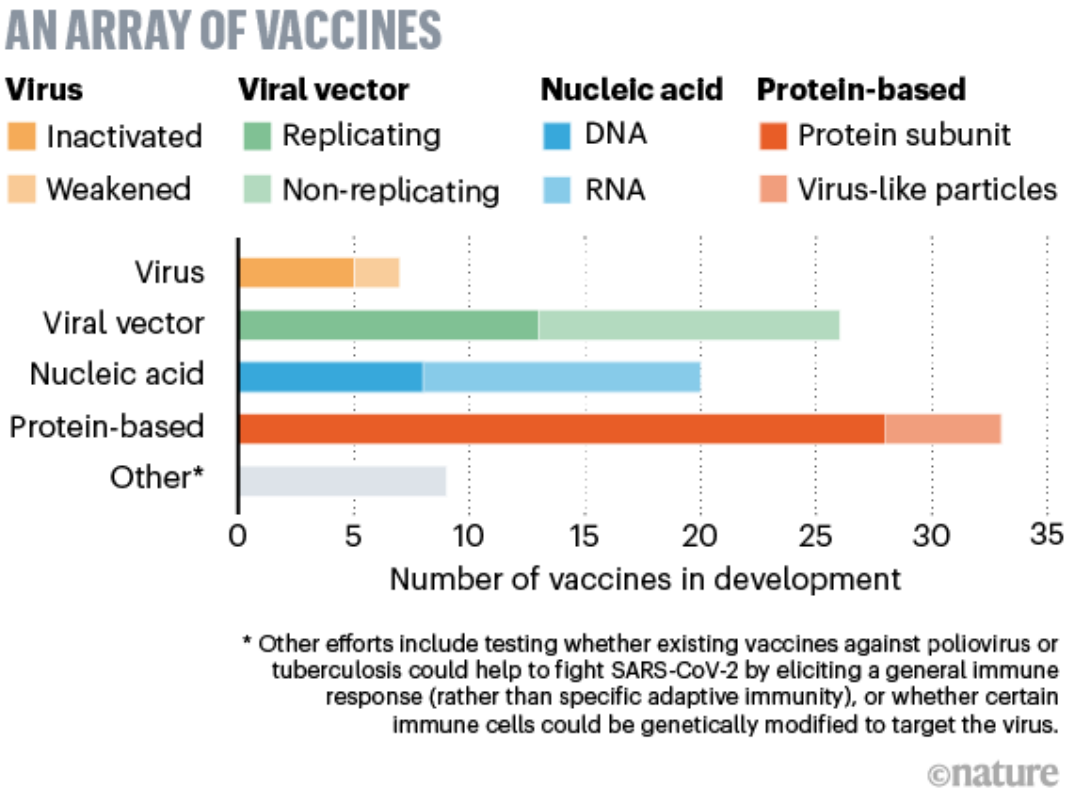
- **Phyre2**

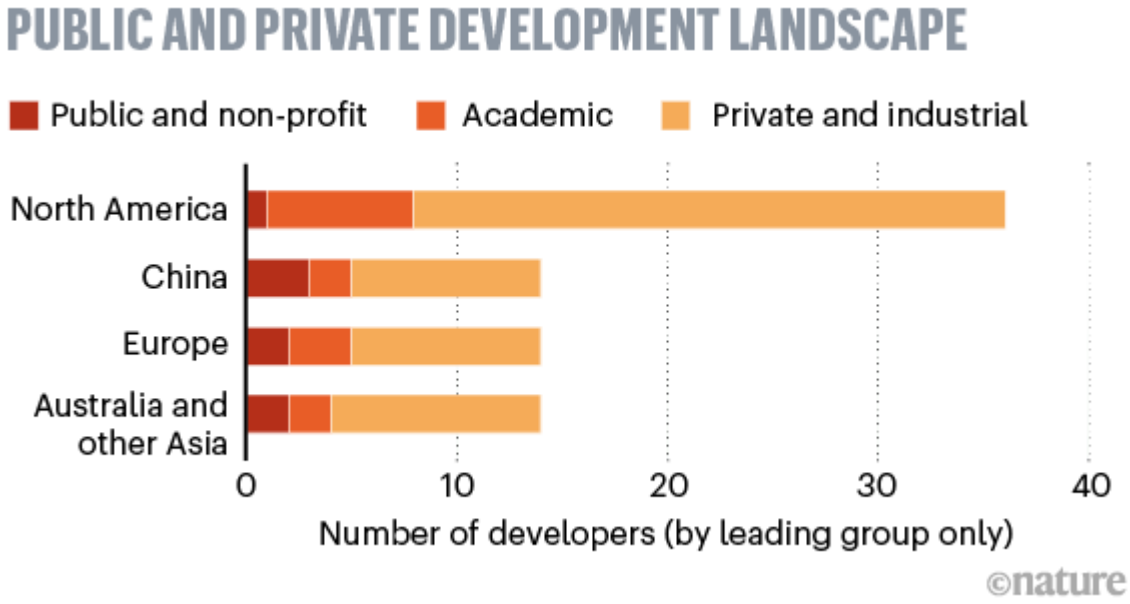


What can vaccine developmental landscape teach us?

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

What receptor will be suitable? Can we get help from ongoing Vaccine development?





<u>Looks like protein-based efforts are gaining ground</u>

Sources: *Nature* analysis based on: WHO COVID-19 Vaccine Landscape/Milken Institute COVID-19 Treatment and Vaccine Tracker/T. Thanh Le *et al. Nature Rev. Drug. Disc.* http://doi.org/ggrnbr (2020)/F. Amanat & F. Krammer *Immunity* **52**, 583–589 (2020)/W. Shang *et al. npj Vaccines* **5**, 18 (2020).

Project-1a: Sars CoV-2 Druggable Receptor focused search

- **Sars CoV-2 receptor choices?**

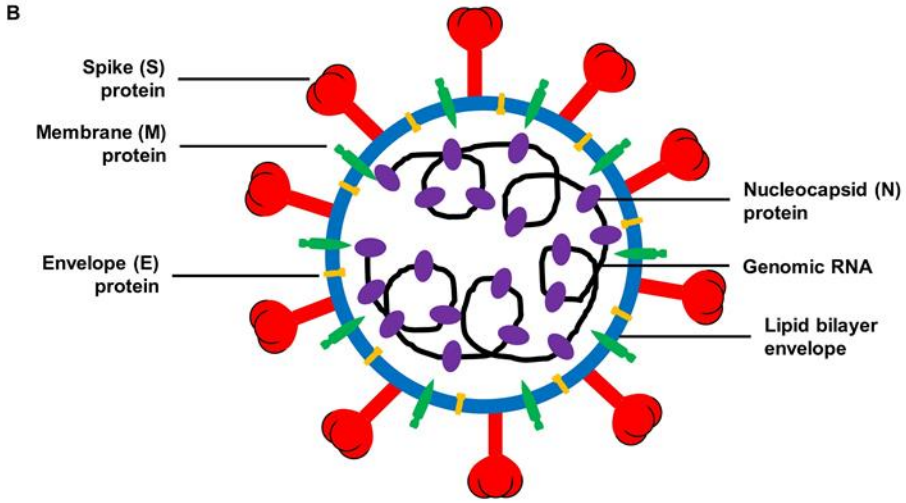
- **Virus entry**
  - Spike protein (experimental structure availability)

- **Protease**
  - Mpro (Most PDB structures)

- **Viral replication**
  - RdRp (polymerase)

- **Virus fusion**
  - Targeting the Fusion of the vesicle and virion

Spike Protein: Active form Trimer



Project-1b: Impact of mutation on the choice of receptors

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

- Most drug developments focus on nsp
  - nsp: non-structural protein(s)
- Most conserved protein is Envelope protein
- Second most conserved in M-Pro and no human-protein is similar to this protein
  - So less toxic

Table 5: Protein-specific statistics of SARS-CoV-2 single mutations. Length refers to the number of codons in the genome associated with a specific protein.

how many freq

Protein	Length	# of mutations	mutation ratio	Mutation h-index
Spike protein	1273	385	0.30	16
Main protease	306	68	0.22	9
Papain-like protease	1945	599	0.31	15
RNA polymerase	932	223	0.24	13
Endoribo-nuclease	346	87	0.25	9
Envelope protein	75	13	0.17	5
Membrane protein	222	63	0.28	9
Nucleocapsid protein	419	235	0.56	27

Mutation Ratio: # of mutations/residue

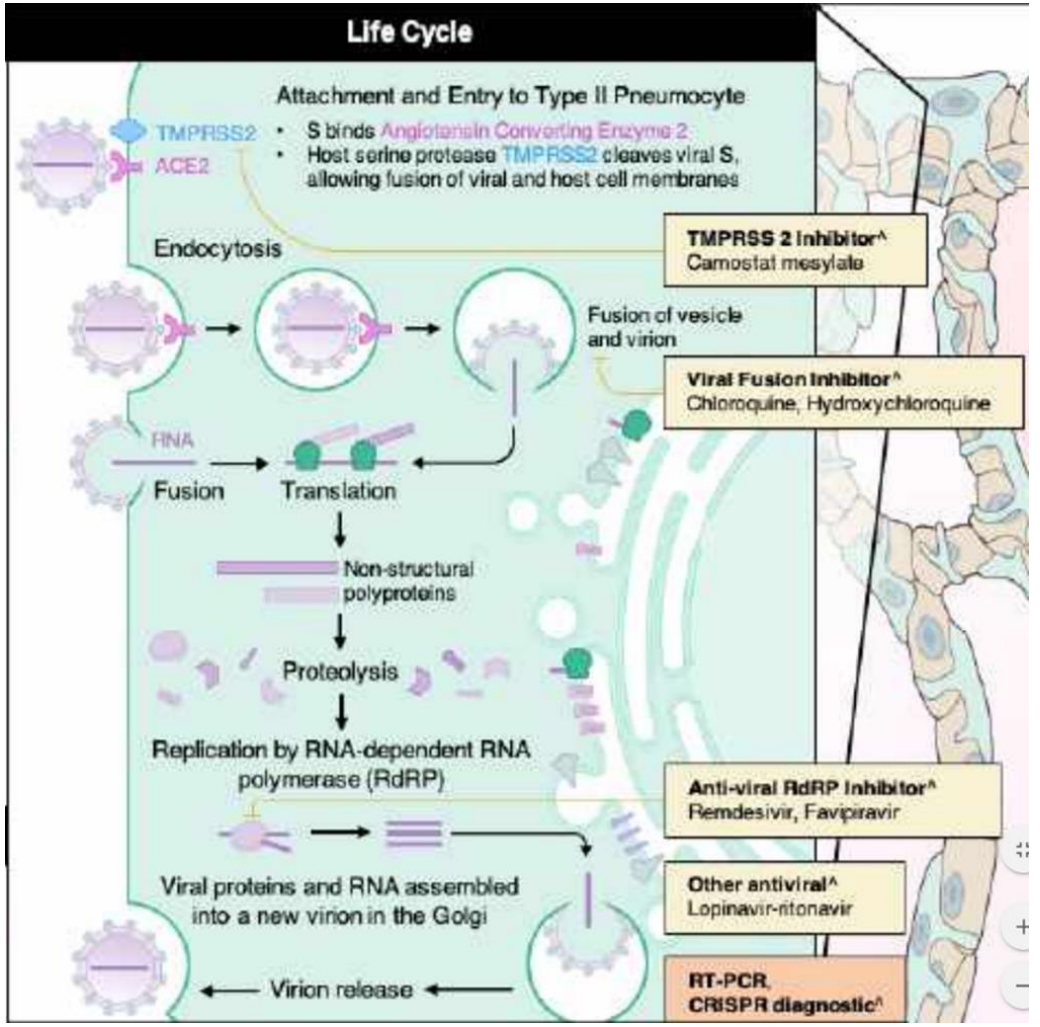
Data collection: Jan 05 to Apr 24, 2020

https://arxiv.org/pdf/2004.14114.pdf

Project-1c: Sars CoV-2 Host Druggable Receptor focused search

- **Virus entry** ***(drugs described here are in market)***
  - TMPRSS2
  - ACE2
  - Furin

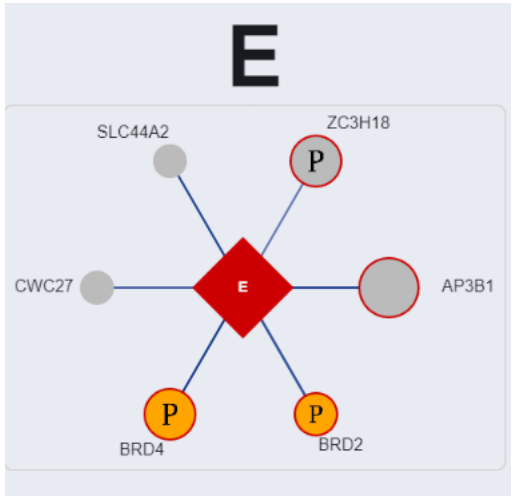
- **Cytokine storm**
  - Mitigate hyper immune response (or immune system hyper-activity)
  - Overproduction of proinflammatory cytokines (tumour necrosis factor [TNF], IL-6, and IL-1β)
  - Targeting thrombin, coagulation factor Xa or PAR-1

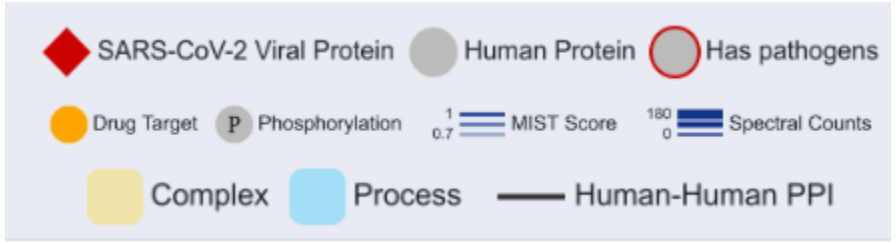


DOI: 10.1016/j.cell.2020.04.013; CELL

Project-1c: Choice of host receptors gleaned via P-P network

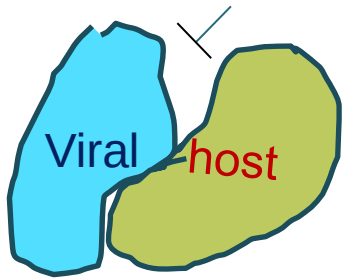
- **D. E. Gordon *et al.* Preprint at bioRxiv https://doi.org/10.1101/2020.03.22.002386; 2020).**
  - "cloned, tagged and expressed 26 of the 29 viral proteins in human cells and identified the human proteins physically associated with each using affinity-purification mass spectrometry (AP-MS)"
  - we identify 67 druggable human proteins or host factors targeted by 69 existing FDA-approved drugs, drugs in clinical trials and/or preclinical compounds,





https://ppi.zoiclabs.io/#/

Here is an example of how targeting BRD2/4 can possibly interrupt virus infections



Drugs that can target either one of them

Project-1a/b/c: Modeling Software Commonly Used

- **Structure-based modeling**
  - ITASSER
  - SwissModel
  - Rosetta *(for small peptides; not so-friendly system to locally setup)*
  - Phyre2
  - Schrodinger (basic protein preparation; commercial)

Modeling Software Commonly Used (based on publications)

- **Visualization and preparation**
  - VMD
  - Schrodinger Maestro (commercial)

Project-2: Simulations of targets to identify suitable conformation(s)

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

- Spike protein is conformationally flexible
  - **Distinct conformational states of SARS-CoV-2 spike protein**
    - **doi:** https://doi.org/10.1101/2020.05.16.099317
    - https://doi.org/10.26434/chemrxiv.11871402.v3
      (repurposing therapeutics for COVID-19; Smith et al)

Here we combine restrained temperature replica-exchange molecular dynamics (restrained T-REMD) simulations with virtual high-throughput screening in an ensemble docking campaign to identify well-characterized drugs, metabolites, and/or natural products that may disrupt S-protein:ACE2 receptor interface stability or the ability of the S-protein to recognize the

- GROMACS/NAMD/AMBER for MD simulations
  - https://chemrxiv.org/articles/Repurposing_Therapeutics_for_the_Wuhan_Coronavirus_nCov-2019_Supercomputer-Based_Docking_to_the_Viral_S_Protein_and_Human_ACE2_Interface/11871402/4
- The low-energy conformations can be used for docking

Project-2: Modeling Software Commonly Used

- **Software**
  - NAMD
  - GROMACS
  - AMBER

Modeling Software Commonly Used (based on publications)

- **Analysis/visualization**
  - **VMD**
  - **Biovia (discovery studio visualizer)**

Project-3: Modeling Protein-Drug/Inhibitor Interactions

- **Where can we get the compounds?**
  - PubChem
  - DrugBank
  - ZINC
  - DrugBank
  - FDA approved drug database (commercial/free versions)
  - ChEMBL
  - SWETLEAD ;
  - DRUGCentral
  - SuperDRUG2
  - Natural-compounds library

Databases Commonly Used (based on publications)

Detailed information from Github repository (will be available soon)

- **Receptor structures could come from Project-2 or from PDB (conformationally stable)**

Project-3: Modeling Protein-Drug/Inhibitor Interactions

- **Questions**
  - What receptors for docking?
  - Conformational effects important?
  - What domain to model?
    - Spike: RBD or whole protein
    - Biologically relevant complex (trimer, dimer etc.)
  - What compound libraries are important?
    - FDA approved; Drugs of a certain class (anti-viral)
  - What if IC50 or ki or kd are not available?
    - Can we estimate them using binding affinity?
    - Scoring and ranking?
  - Natural compound library screening (a separate effort?)

Databases Commonly Used (based on publications)

Detailed information from Github repository (will be available soon)

Project-3a: Carrying out docking using the receptor conformations that had been identified in Project-2

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

- **We have created a custom small-molecule set (~ 300 compounds)**
  - Collected from COVID-19 publications (reference; pubchem id and name included)
  - Contains inhibitor targets: Spike, M-pro, ACE2, RdRp, Viral replication/activity, Cytokine storm, TMPRSS2 etc
  - Experimental binding information (Ex. IC50, EC50 etc.)

PubChem	Name	IC50 (mM: Micro Molar)	Reference	Inhibitor Target
3194	Ebselen	0.67 +/- 0.09 mM	https://doi.org/10.1038/s41586-020-2223-y	M(pro) Protease
3117	Disulfiram	9.35 pm 0.18 mM	https://doi.org/10.1038/s41586-020-2223-y	M(pro) Protease
11313622	Tideglusib	1.55 pm 0.30 mM	https://doi.org/10.1038/s41586-020-2223-y	M(pro) Protease
2577	Carmofur	1.82 pm 0.06 mM	https://doi.org/10.1038/s41586-020-2223-y	M(pro) Protease
479503	Shikonin	15.75 pm 8.22 mM	https://doi.org/10.1038/s41586-020-2223-y	M(pro) Protease

An accompanying Jupyter Notebook to analyze and display compounds

COVID19-disruptors.ipynb

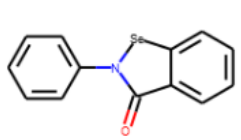
Code Python 3

```
df4 = pcp.compounds_to_frame(cs, properties=['isomeric_smiles','xlogp', 'rotatable_bond_count'])
# df4
# df4.dtypes
```

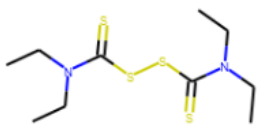
[190]:

```
# show the molecules
core = Chem.MolFromSmiles( 'C1=CC=CC2=C1C=CC=C2' )

#amide
core1 = Chem.MolFromSmiles( 'C(N[H])=O' )

csm_list = df4.index.values.astype('str').tolist()

csm = [Chem.MolFromSmiles(x) for x in df4.isomeric_smiles.to_list()]
Draw.MolsToGridImage( csm[:24], molsPerRow = 4, subImgSize=(400, 300),legends = csm_list[:24])
```

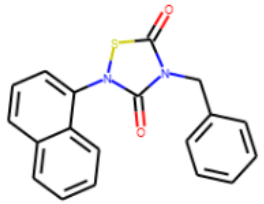
[190]:



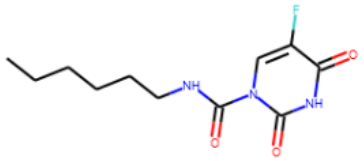
3194



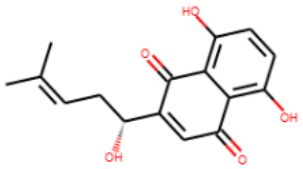
3117

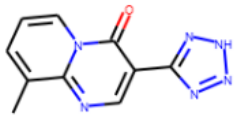


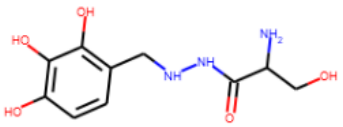
11313622



2577







Project-4: Machine-learning or Deep-learning methodology

- **Drug repurposing**
- "*Drug repurposing is a strategy for identifying new uses for approved or investigational drugs that are outside the scope of the original medical indication*" Nat. Rev. 18, 41, 2019

- Data for Modeling and other questions?
  - Assay or experimental data?
  - How many compounds in the data(set)?
    - From where this data came from?
  - How many classes of outcome and are they balanced?

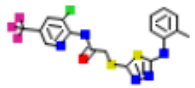
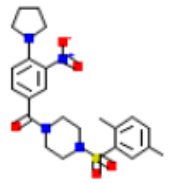
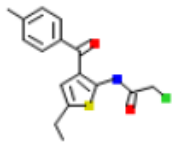
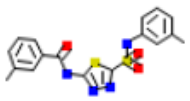
Project-4: Can a related Bioassay be used for COVID-19 3CL-Pro

BIOASSAY RECORD

QFRET-based primary biochemical high throughput screening assay to identify inhibitors of the SARS coronavirus 3C-like Protease (3CLPro)

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

https://pubchem.ncbi.nlm.nih.gov/bioassay/1706

Tested Substance			Activity	Score	Inhibition, %
Structure	CID	SID			
	4175307	22406679	Active	24	19.74
	24819855	49828046	Active	15	12.84
	859639	17508646	Active	24	19.51
	16017527	49722098	Active	15	12.42

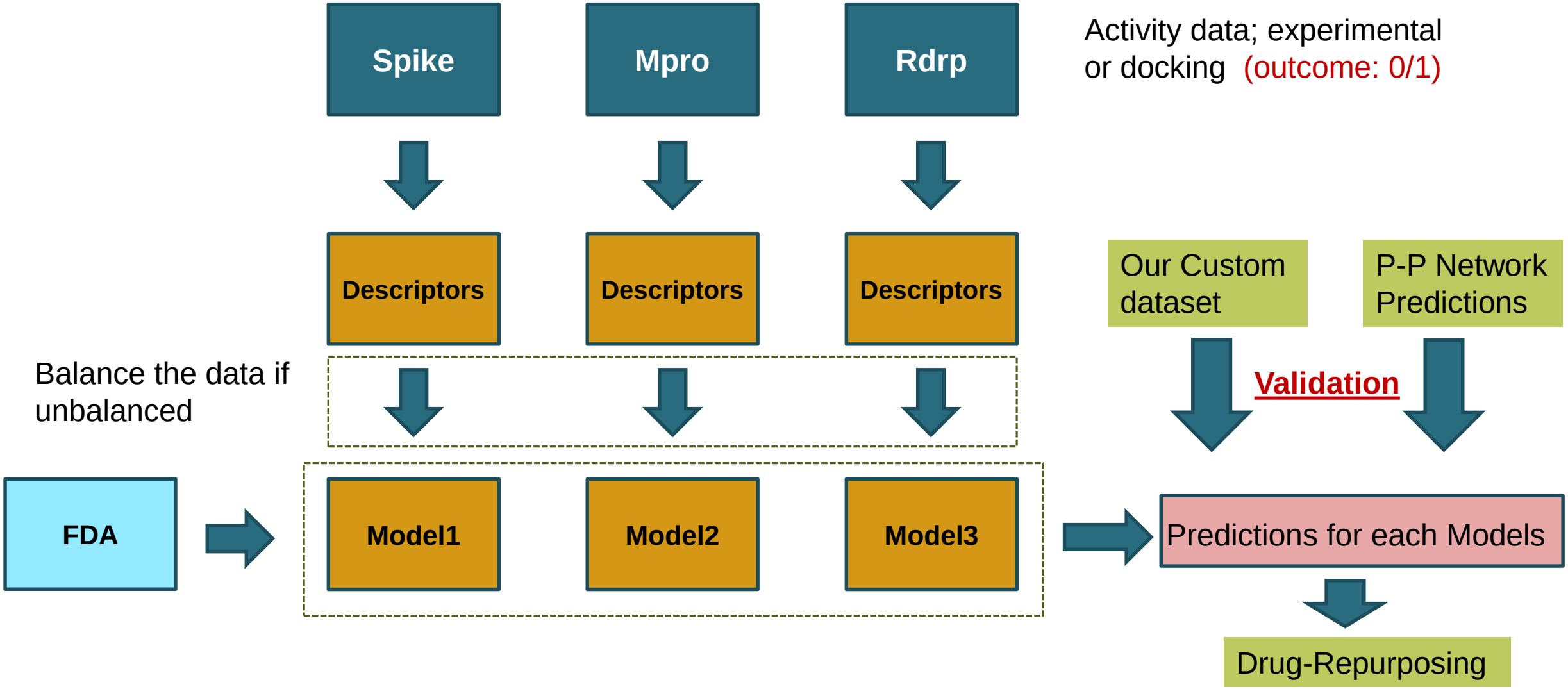
Balance the dataset

FDA Dataset

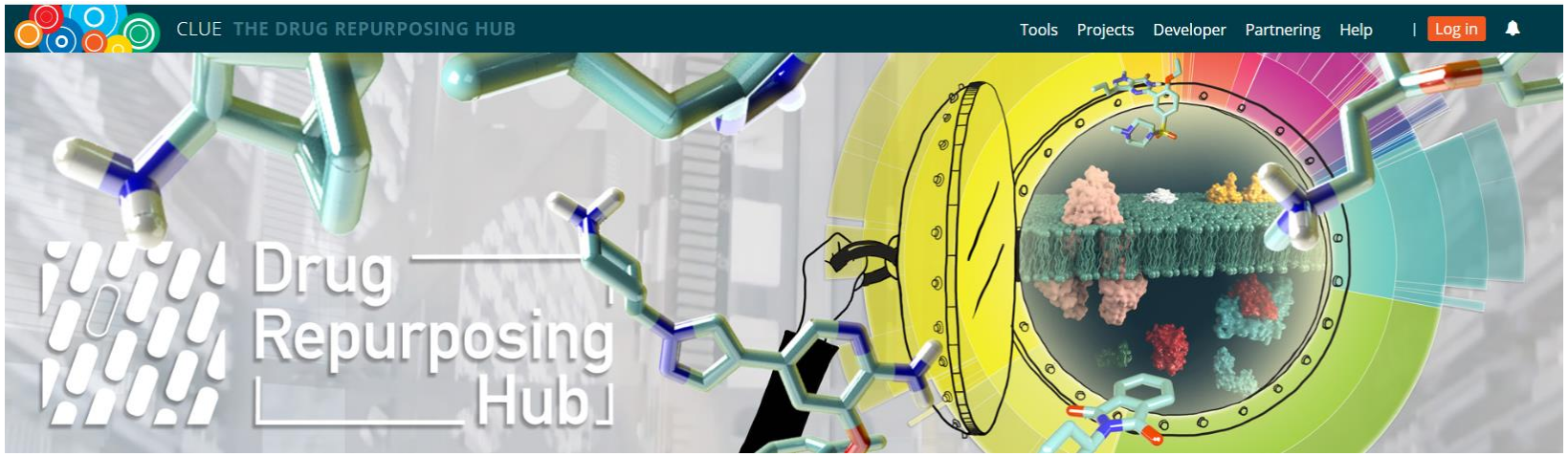
Model for Protease inhibitors

What compound from FDA dataset is active for 3CL-Pro?

Drug Repurposing ML Pipeline



Important site for ML Drug repurposing effort



What does the drug database have?

Contains annotation-level information including compound name, clinical phase, mechanism of action, and protein target.

BROAD INSTITUTE

13,553 TOTAL SAMPLES | 2,183 PROTEIN TARGETS | 6,798 UNIQUE COMPOUNDS | 670 DRUG INDICATIONS

pert_iname	clinical_phase	moa	target	disease_area	indication
(R)-(-)-apomorphine	Launched	dopamine receptor agonist	ADRA2A|ADRA2B|ADRA2C|CALY|[	neurology/psychiatry	Parkinson's Disease
(R)-(-)-rolipram	Phase 1	phosphodiesterase inhibitor	PDE4A|PDE4B|PDE4C|PDE4D|PDE[		
(R)-baclofen	Phase 3	benzodiazepine receptor agonist	GABBR1|GABBR2		
(S)-(+)-rolipram	Phase 1	phosphodiesterase inhibitor	PDE4B|PDE4D		
[sar9,met(o2)11]-substance-p	Preclinical	tachykinin antagonist	TACR1		
A-1070722	Preclinical	glycogen synthase kinase inhibitor	GSK3A|GSK3B		
A-1120	Preclinical	retinoid receptor ligand	RBP4		
A-317491	Preclinical	purinergic receptor antagonist	P2RX3		
A-33903	Phase 2				
A-366	Preclinical	histone lysine methyltransferase inhibitor	EHMT1|EHMT2		
A-381393	Preclinical	dopamine receptor antagonist			

6798 Compounds

Broad Drug Repurposing Compounds Data (13, 563)

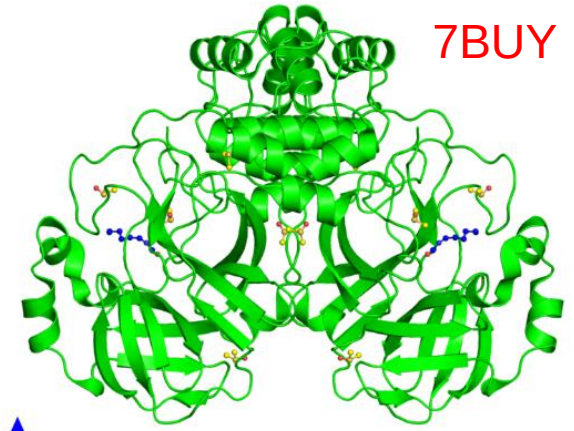
Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

broad_id	pert_iname	qc_incompatible	purity	vendor	catalog_no	vendor_name	expected_mass	smiles	pubchem_cid	deprecated_broad_id
BRD-K76022557-003-28-9	(R)-(-)-apomorphine	0	98.9	MedChemEx	HY-12723A	Apomorphine (hydrochloride hemihydrate)	267.126	CN1CCc2cccc-3c2[C@H]1Cc1ccc(O)c(O)c-31	6005	
BRD-K76022557-003-02-7	(R)-(-)-apomorphine	0	97.34	Tocris	2073	(R)-(-)-Apomorphine hydrochloride	267.126	CN1CCc2cccc-3c2[C@H]1Cc1ccc(O)c(O)c-31	6005	
BRD-K76022557-003-29-9	(R)-(-)-apomorphine	0	97.36	Tocris	2073	(R)-(-)-Apomorphine hydrochloride	267.126	CN1CCc2cccc-3c2[C@H]1Cc1ccc(O)c(O)c-31	6005	
BRD-K76022557-001-03-9	(R)-(-)-apomorphine	0	95.8	Selleck	S4350	R-(-)-Apomorphine HCl Hemihydrate	267.126	CN1CCc2cccc-3c2[C@H]1Cc1ccc(O)c(O)c-31	6005	
BRD-K75516118-001-04-1	(R)-(-)-rolipram	0	93.92	Tocris	1349	(R)-(-)-Rolipram	275.152	COc1ccc(cc1OC1CCCC1)[C@@H]1CNC(=O)C1	448055	
BRD-K75516118-001-05-9	(R)-(-)-rolipram	0	93.75	Tocris	1349	(R)-(-)-Rolipram	275.152	COc1ccc(cc1OC1CCCC1)[C@@H]1CNC(=O)C1	448055	
BRD-K75516118-001-03-3	(R)-(-)-rolipram	0	97.48	Tocris	1349	(R)-(-)-Rolipram	275.152	COc1ccc(cc1OC1CCCC1)[C@@H]1CNC(=O)C1	448055	

broad_id
pert_iname
qc_incompatible
purity
vendor
catalog_no
vendor_name
expected_mass
smiles
pubchem_cid
deprecated_broad_id

Project-5: Experimental study of the compounds identified in Project 3/4

- **Taken from https://www.nature.com/articles/s41586-020-2286-9_reference.pdf**
  - More information on page 4 (Antiviral activity of host-directed drugs and compounds)
  - "medium-throughput immunofluorescence-based assay (detecting the viral NP protein) to screen 37 compounds for inhibition of SARS-CoV-2 infection in the Vero E6 cell line."
  - "viral RNA was monitored using qRT-PCR upon treatment with 44 drugs and compounds"
  - "TCID50 assays on supernatants from infected cells treated with PB28 (IC90 0.278 μM) and zotatifin (IC90 0.037 μM) revealed a more potent inhibition than was observed in the NP-staining assay"
  - Antiviral effect
    - "*To better understand the mechanism by which these inhibitors exert their antiviral effects, we performed a time course assay where the drugs were added at different times relative to infection*"



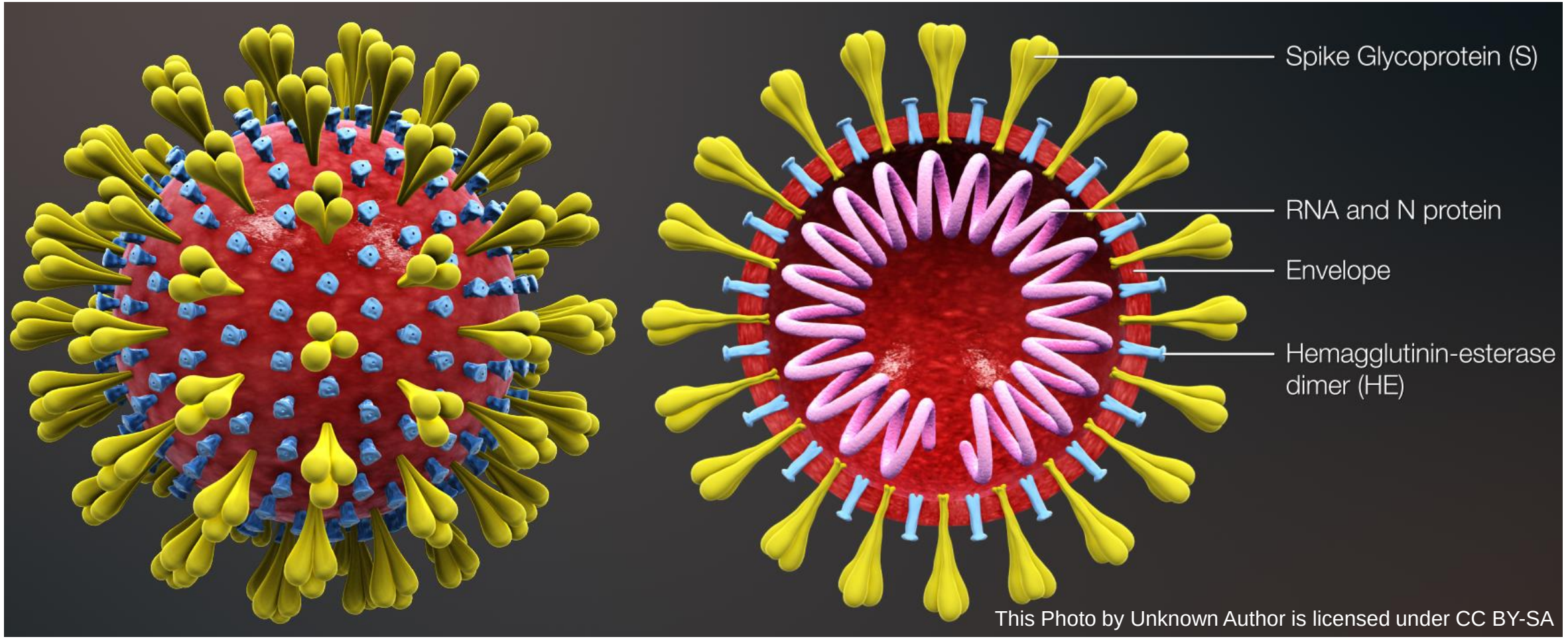
Project-5: Experimental study of the compounds identified in Project 3

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

- **Nature Communications**
  - https://www.nature.com/articles/s41594-020-0440-6; Structural basis for the inhibition of SARS-CoV-2 main protease by antineoplastic drug carmofur
    - pre-seeded Vero E6 cells used with qRT-PCR analysis, while cells were fixed and subjected to immunofluorescence to monitor intracellular NP level as described previously
    - For cytotoxicity assays, Vero E6 cells were suspended in growth medium in 96-well plates. The next day, appropriate concentrations of carmofur were added to the medium. After 24 h, the relative numbers of surviving cells were measured using a Cell Counting Kit-8 (CCK8, Beyotime) assay in accordance with the manufacturer's instructions. All experiments were performed in triplicate, and all the infection experiments were performed at biosafety level 3 (BSL-3).

Basic Biology of COVID-19

Structural protein: surface glycoprotein: Spike Protein

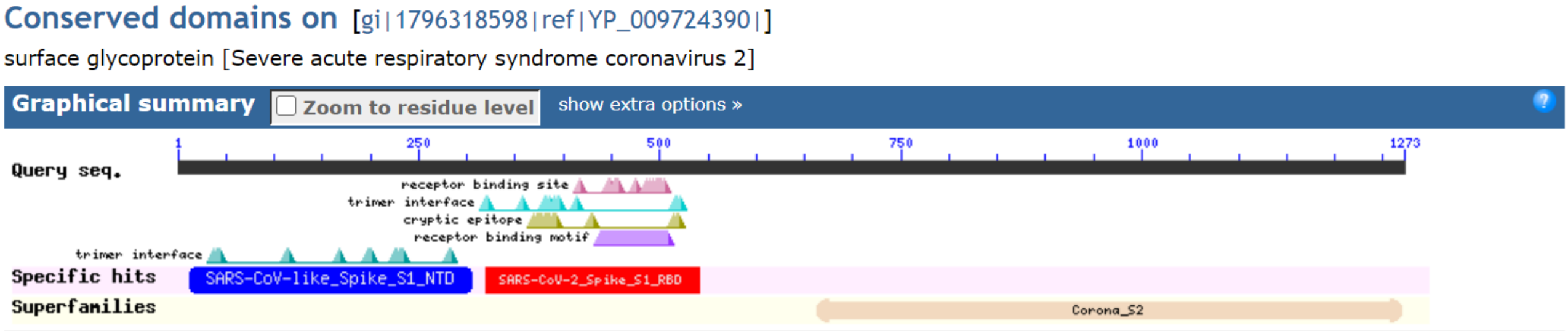


https://covid-19.uniprot.org/uniprotkb/P0DTC2

Frederick National Laboratory for Cancer Research
sponsored by the National Cancer Institute

Spike Protein: Conserved Domains

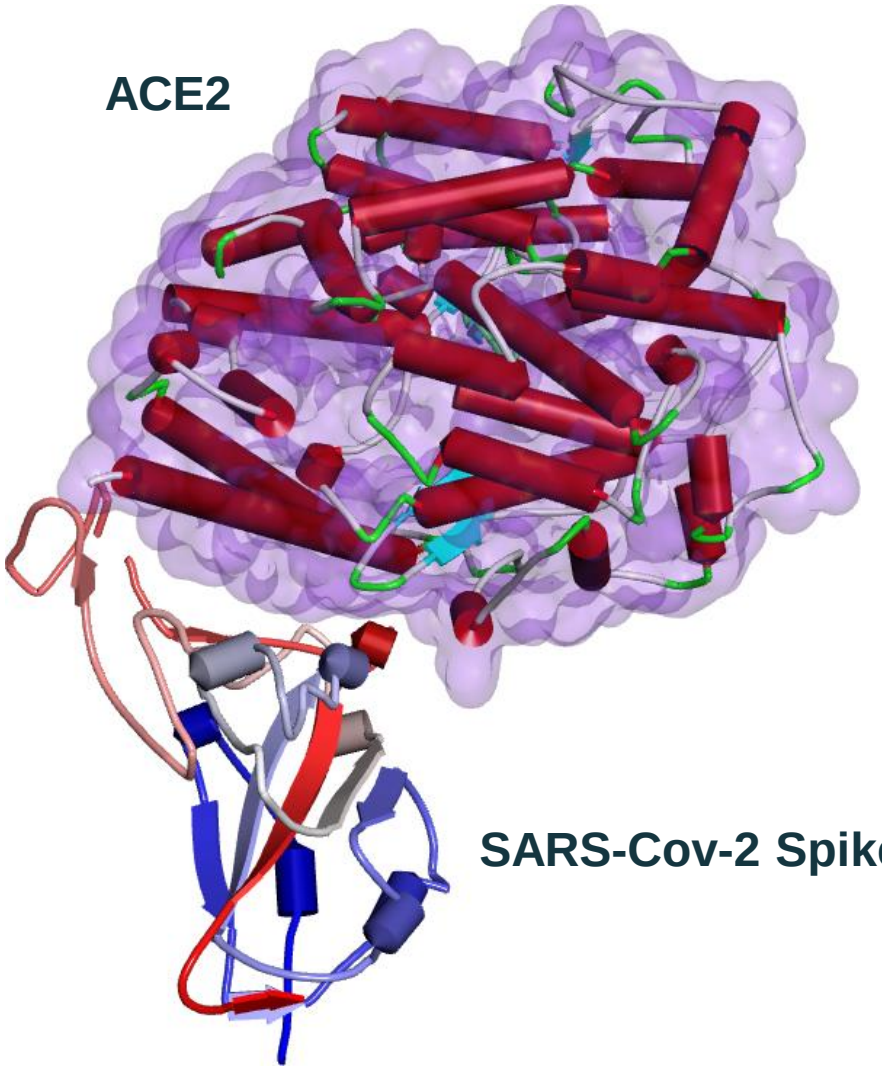
Coronavirus gets its name from the Spike proteins. Spikie proteins are located on the surface. They usually occur as trimers



N-terminal domain of the S1 subunit of the Spike (S) protein

Receptor-binding domain of the S1 subunit

Crystal structure of SARS-CoV-2 spike receptor-binding domain bound with ACE2 (PDB: 6M0J)



PDBe-KB: 21 experimental 3D structures of either S-protein alone or with other proteins (05/27/2020)

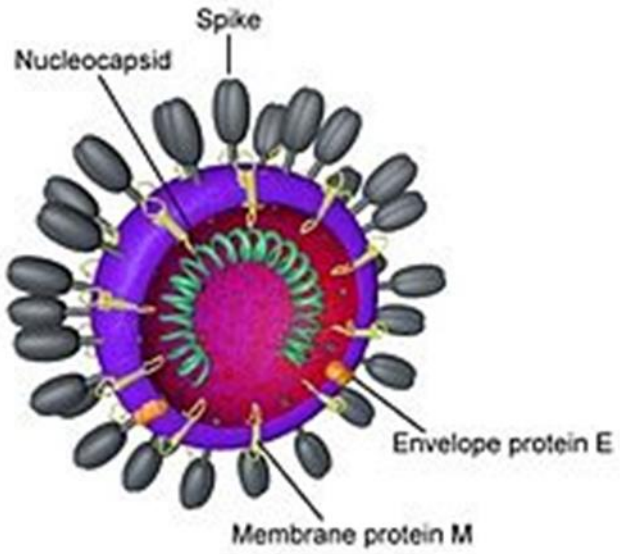
https://www.uniprot.org/uniprot/P0DTC2

Structural protein: surface glycoprotein: Spike Protein

https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P0DTC2#structures

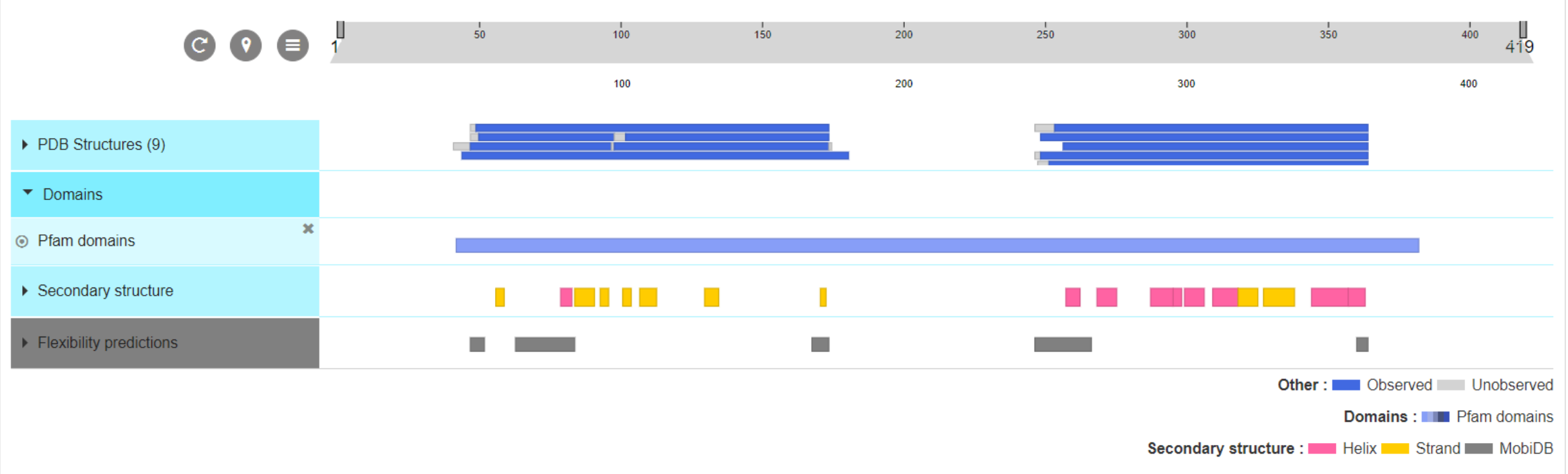
Nucleocapsid phosphoprotein

- https://www.ncbi.nlm.nih.gov/gene/43740575
- https://www.uniprot.org/uniprot/P0DTC9
- Safe-guards/keeps the virus stable RNA
- There are many N-proteins that are linked in a spiral and these are often wrap and coil around the RNA
- https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P0DTC9
- **419 aa length**



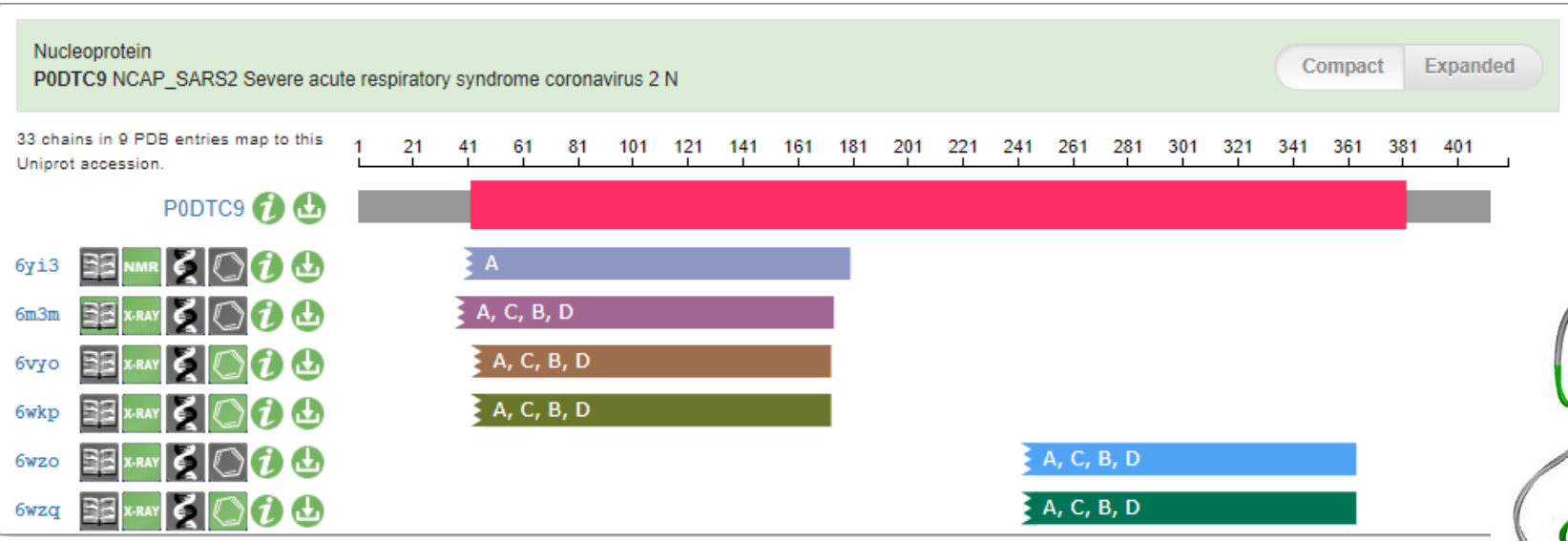
Nucleocapsid phosphoprotein

- **9 PDB structures (PDBe-kb; date: 05/27/2020)**
- **3D structure coverage**

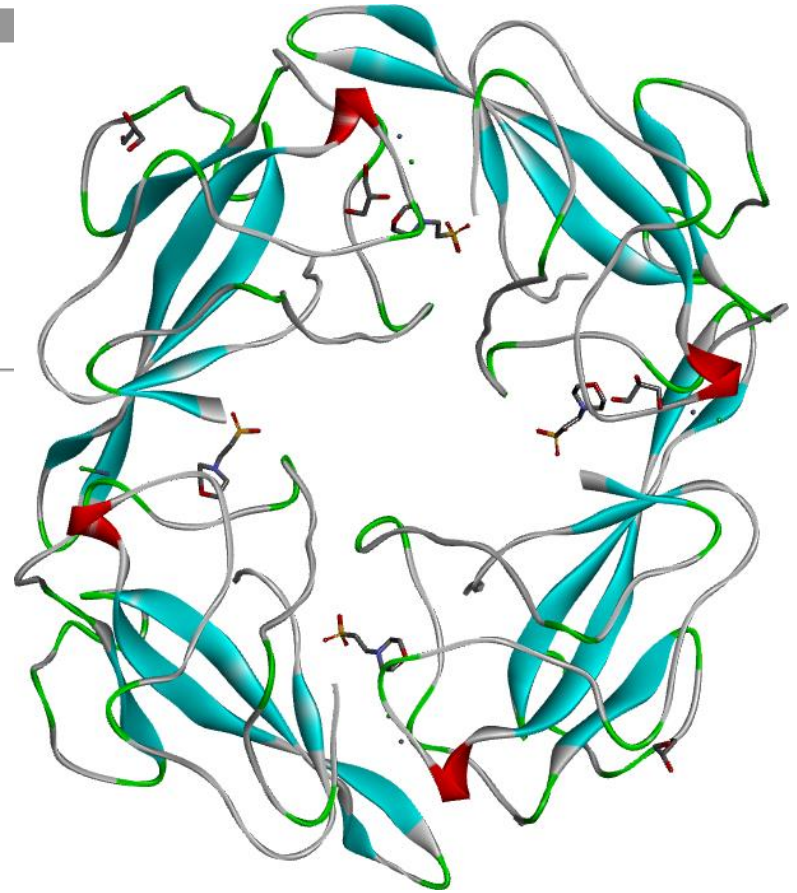


https://www.ebi.ac.uk/pdbe/pdbe-kb/proteins/P0DTC9#structures

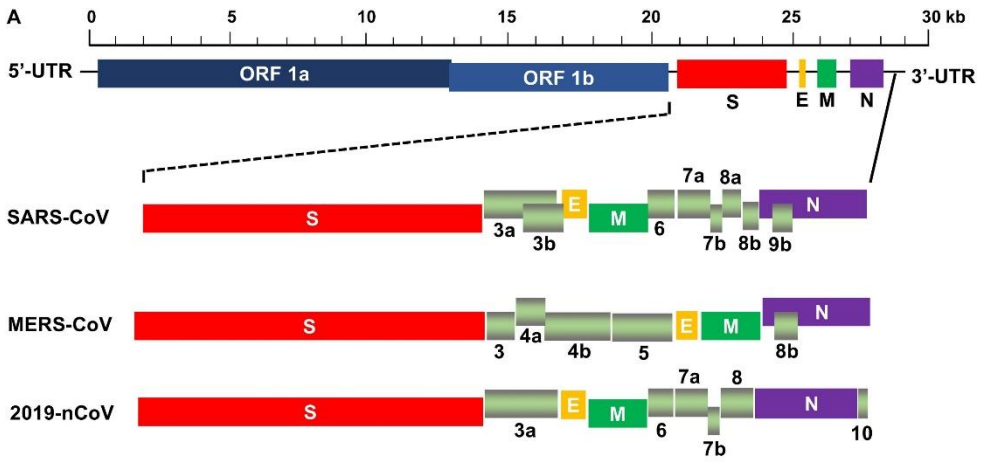
Crystal structure of RNA binding domain of nucleocapsid phosphoprotein from SARS coronavirus 2 (PDB: 6VYO)

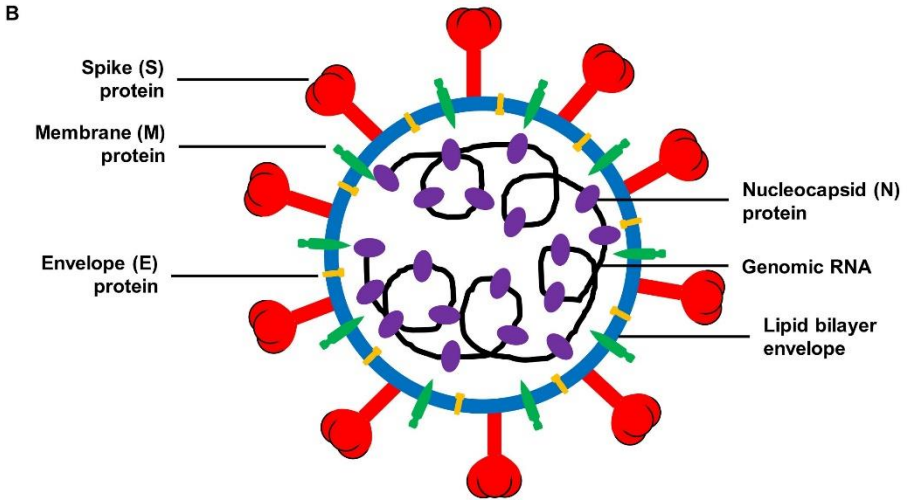


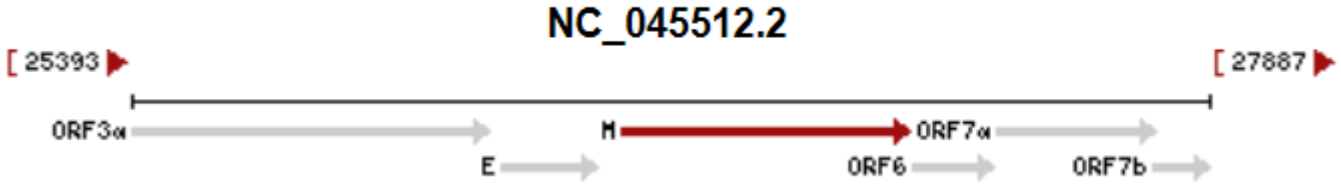
Sequence-coverage for 6VYO and other proteins



M membrane glycoprotein





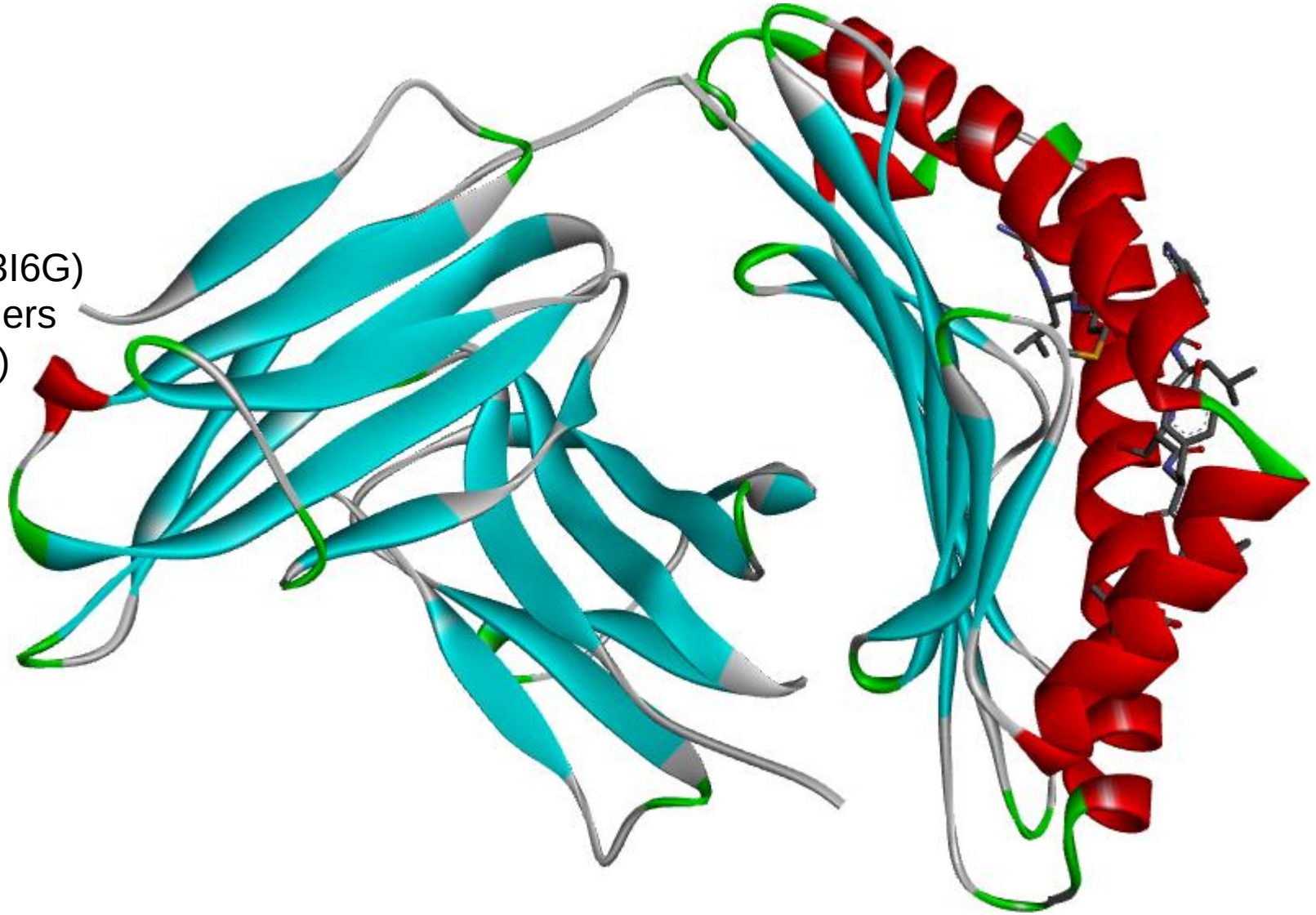


The highlighted area is the only region for which 3D structure data is available, PDB: 3I6G

```
        10         20         30         40         50
MADSNGTITV EELKKLLEQW NLVIGFLFLT WICLLQFAYA NRNRFLYIIK
        60         70         80         90         100
LIFLWLLWPV TLACFVLAAV YRINWITGGI AIAMACLVGL MWLSYFIASF
       110        120        130        140        150
RLFARTRSMW SFNPETNILL NVPLHGTILT RPLLESELVI GAVILRGHLR
       160        170        180        190        200
IAGHHLGRCD IKDLPKEITV ATSRTLSYYK LGASQRVAGD SGFAAYSRYR
       210        220
IGNYKLNTDH SSSSDNIALL VQ
```

Newly identified epitope Mn2 from SARS-CoV M protein complexed withHLA-A*0201 (PDB: 3I6G)

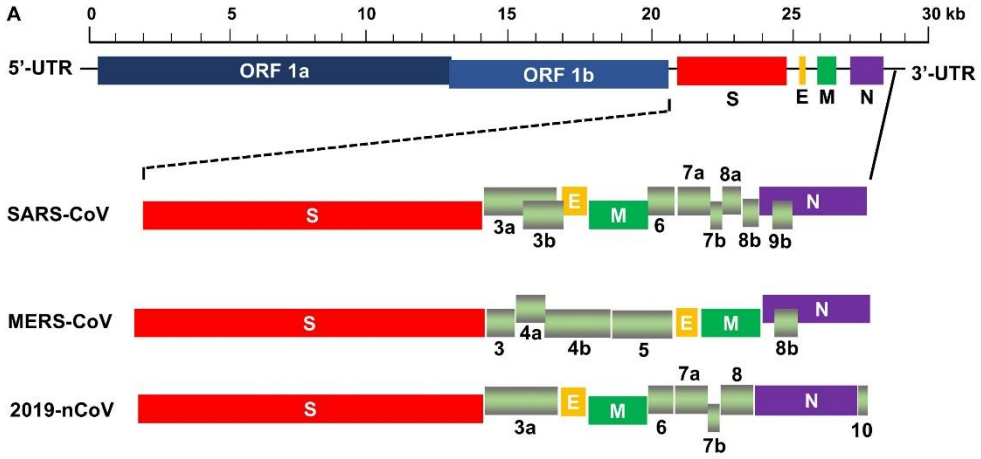
Structural protein
Forms the outer coat of the virus
https://www.uniprot.org/uniprot/P59596
221 amino acids
RCSB: One experimental 3D structure (3I6G)
PDBe-KB: three experimental 3D structuers
(see below; as of 05/27/2020)

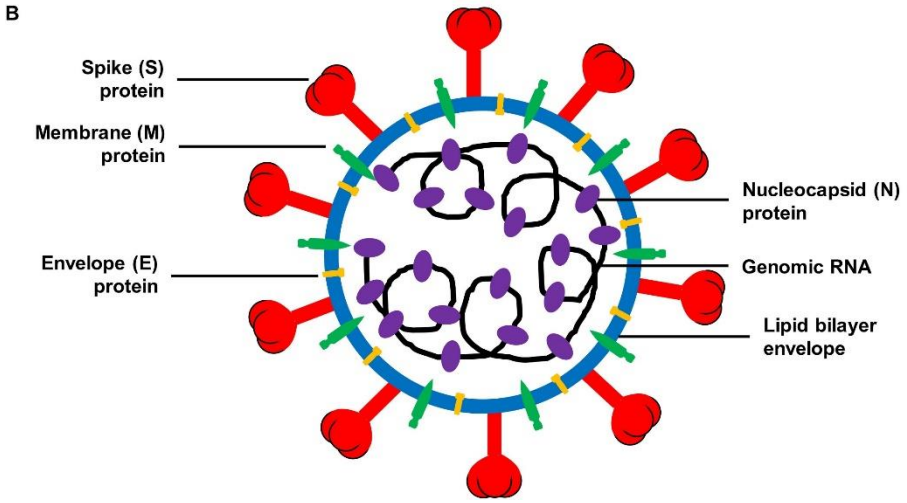


PDB Structures (3)

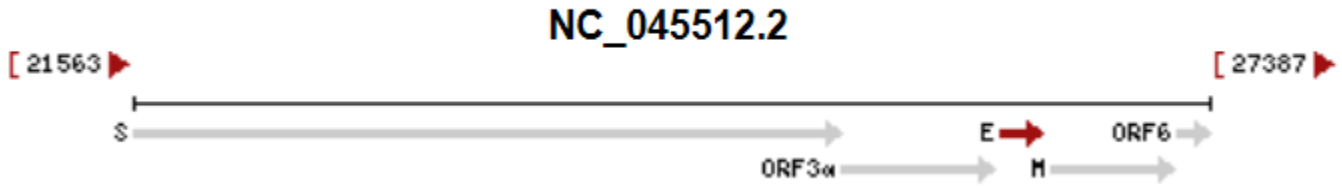
3to2	2.6Å
3i6k	2.8Å
3i6g	2.2Å

E Envelope Protein





- https://www.ncbi.nlm.nih.gov/gene/43740570
- https://www.uniprot.org/uniprot/P0DTC4
- **75 amino acid**
- **No 3D structure**

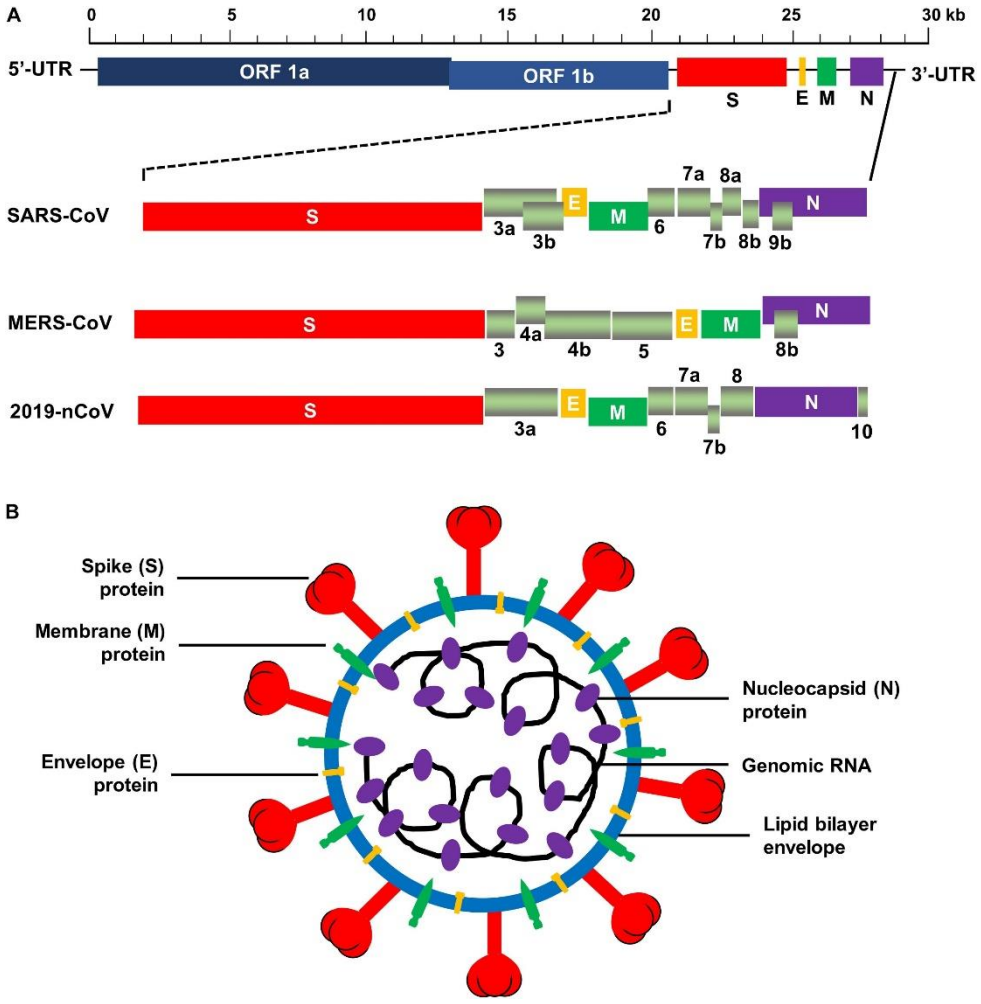


E protein

- **Function**
- The envelope protein is a structural protein that helps form the oily bubble of the virus. It may also have jobs to do once the virus is inside the cell. Researchers have found that it latches onto proteins that help turn our own genes on and off. It’s possible that pattern changes when the E protein interferes.

Important nonstructured proteins

Coronavirus Main Proteinase (3CL^{pro}) (M-pro or nsp5 or main protease)



- nsp5 makes most of the cuts that free other nsp proteins to carry out their jobs.

3C-like proteinase (EC:3.4.22.-)

Short name:
3CL-PRO

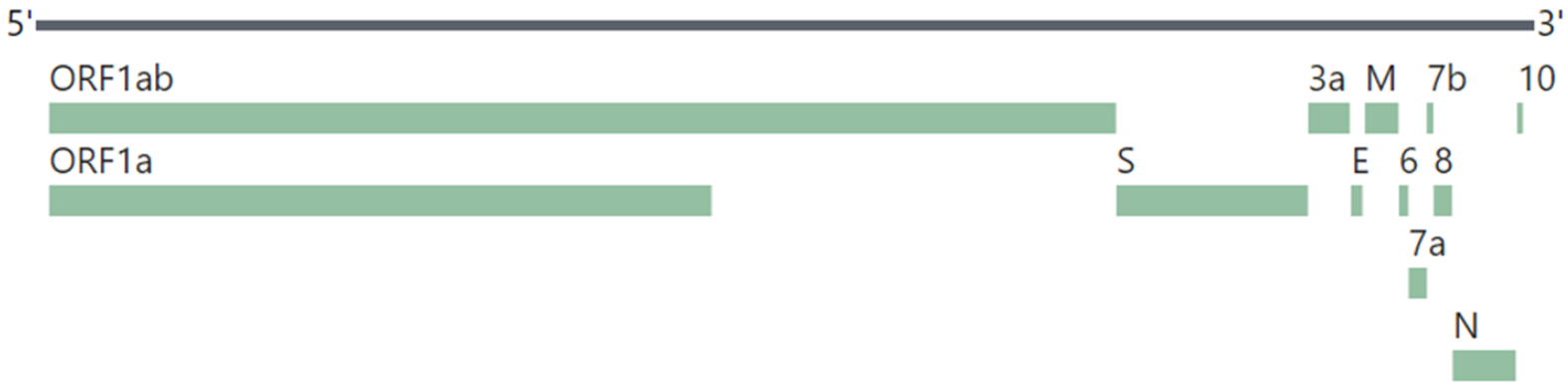
Short name:
3CLp

Alternative name(s):
M-PRO
nsp5
p34

Mpro function

Protein	Submitted name: **Orf1ab polyprotein**
Gene	**PP1ab**
Organism	*Human SARS coronavirus (SARS-CoV) (Severe acute respiratory syndrome coronavirus)*

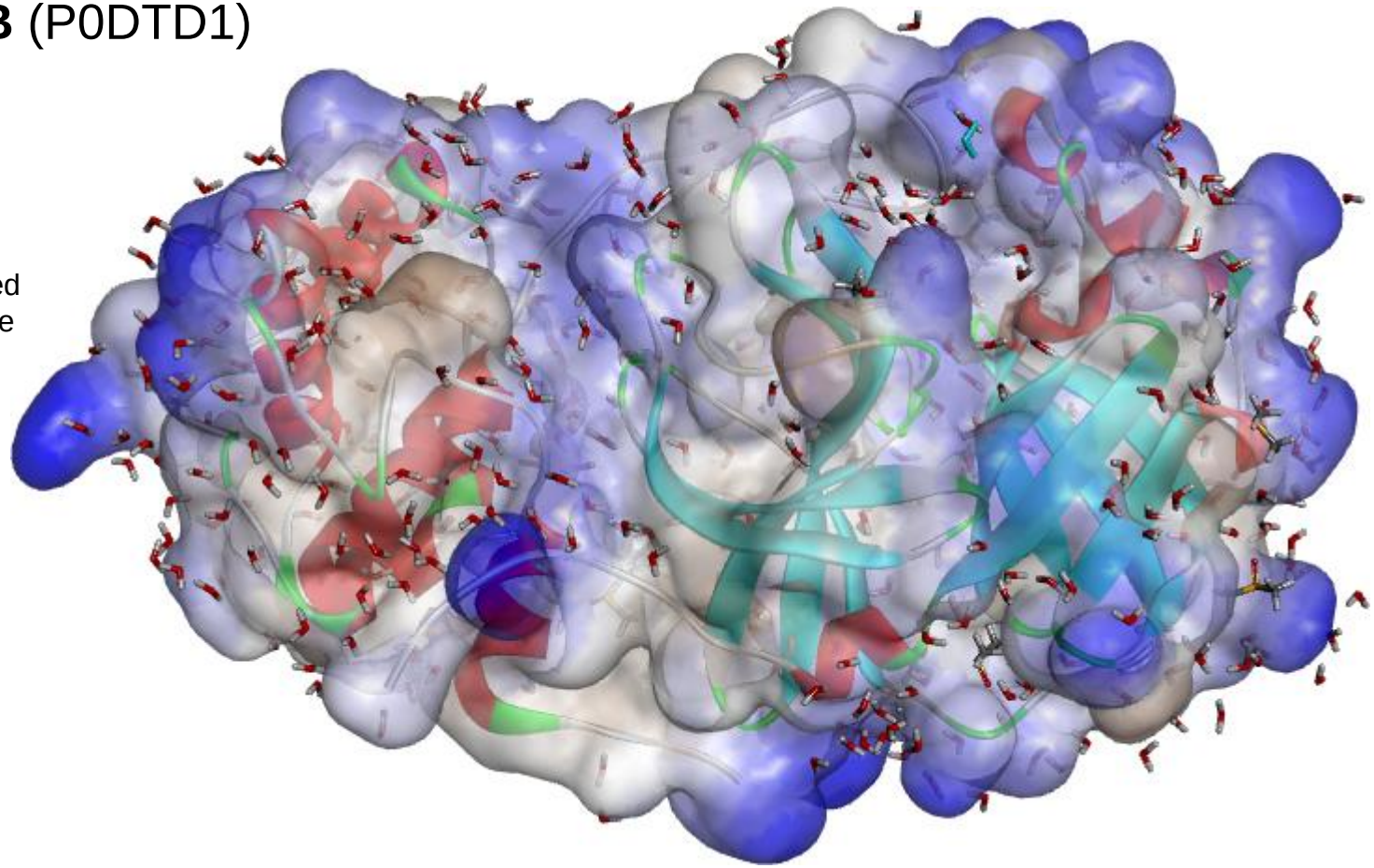
- **Brief summary from https://www.nature.com/articles/s41586-020-2223-y**
- **"The COVID-19 virus genome is comprised of ~30,000 nucleotides; its replicase gene encodes two overlapping polyproteins, pp1a and pp1ab, required for viral replication and transcription. The functional polypeptides are released from the polyproteins by extensive proteolytic processing, predominantly by a 33.8-kDa main protease (Mpro), also referred to as the 3C-like protease. Mpro digests the polyprotein at no less than 11 conserved sites, starting with the autolytic cleavage of this enzyme itself from pp1a and pp1ab. The functional importance of Mpro in the viral life cycle, together with the absence of closely related homologues in humans, identify the Mpro as an attractive target for antiviral drug design"**



Main Protease 3D structures

- Polymer: 1
  Length: 306 residues
  Chain Type: polypeptide(L)
  Reference: **UniProtKB** (P0DTD1)

134 experimental structures as of 05/27/2020 (from RCSB-PDB)



SARS-CoV-2 main protease with unliganded active site (2019-nCoV, coronavirus disease 2019, COVID-19; PDB: 6YB7)

Thanks

Questions
Contact ravichandrans@mail.nih.gov