Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

David Gifford
Lecture 20

April 21, 2020

COVID-19
Machine Learning Designed Therapeutics

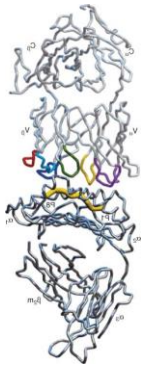
http://mit6874.github.io

Overview of today's lecture

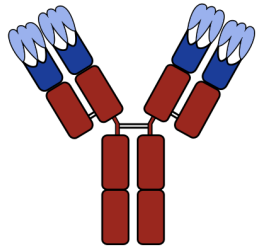
- COVID-19 and SARS-CoV-2 overview
- COVID-19 epidemiology
- COVID-19 testing
- Vaccines for COVID-19
- Antibody therapeutics for COVID-19

Today’s deep learning methods

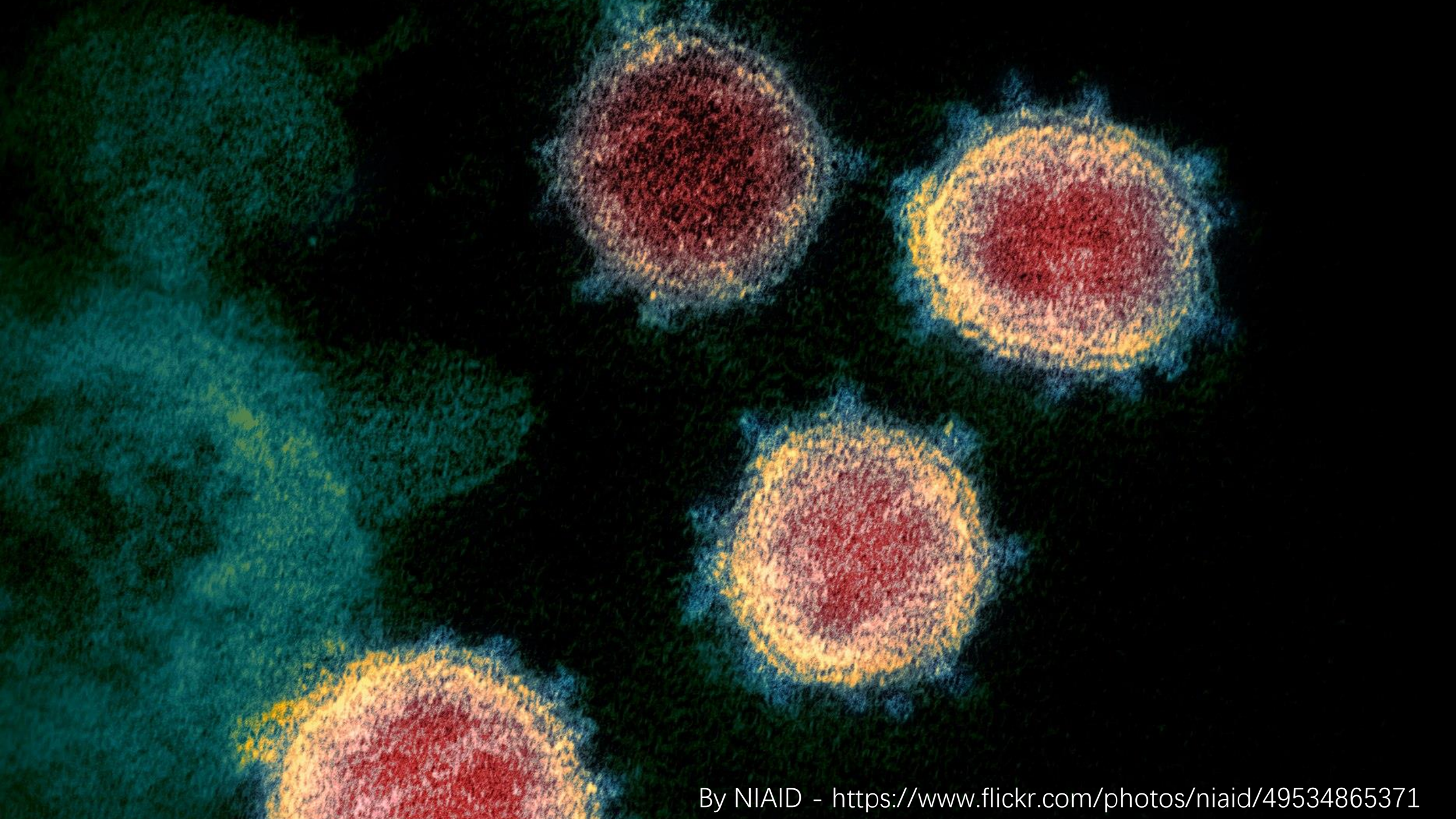
- Vaccine design



- Antibody discovery and improvement

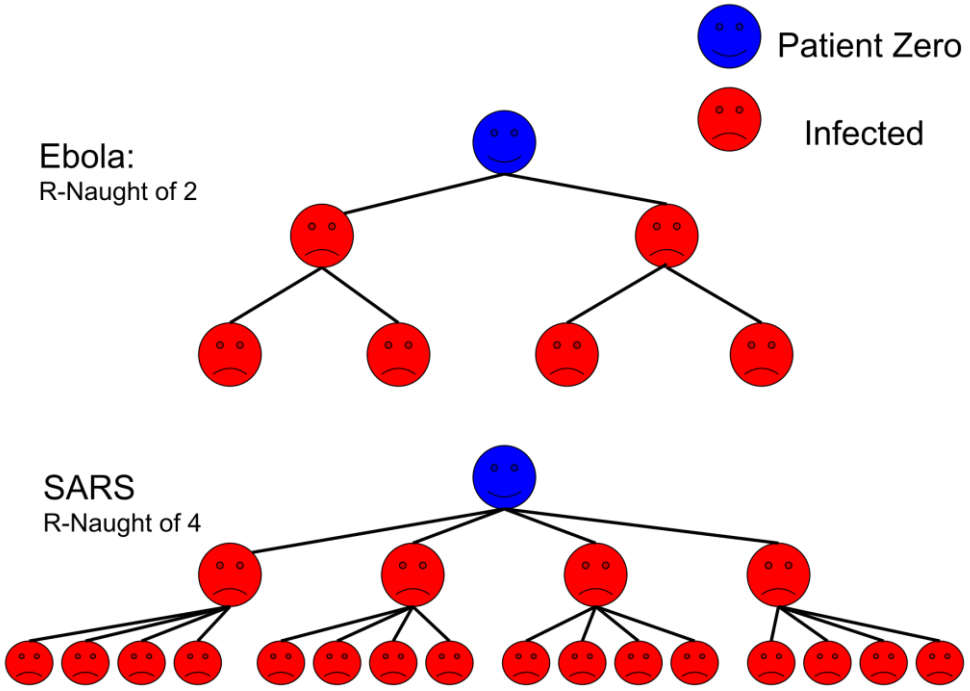


COVID-19 (the disease)
SARS-CoV-2 (the virus)



By NIAID - https://www.flickr.com/photos/niaid/49534865371

The basic reproduction number R_0 describes number of secondary infections from one individual



Values of R_0 of well-known infectious diseases[1]

Disease	Transmission	R_0
Measles	Airborne	12–18[2]
Chickenpox (varicella)	Airborne	10–12[3]
Mumps	Airborne droplet	10–12[4]
Polio	Fecal–oral route	5–7
Rubella	Airborne droplet	5–7
Pertussis	Airborne droplet	5.5[5]
Smallpox	Airborne droplet	3.5–6[6]
COVID-19	Airborne droplet	1.4–5.7[7][8][9][10]
HIV/AIDS	Body fluids	2–5
SARS	Airborne droplet	2–5[11]
Common cold	Airborne droplet	2–3[12]
Diphtheria	Saliva	1.7–4.3[13]
Influenza (1918 pandemic strain)	Airborne droplet	1.4–2.8[14]
Ebola (2014 Ebola outbreak)	Body fluids	1.5–2.5[*citation needed*]
Influenza (2009 pandemic strain)	Airborne droplet	1.4–1.6[15]
Influenza (seasonal strains)	Airborne droplet	0.9–2.1[15]
MERS	Airborne droplet	0.3–0.8[16]

https://en.wikipedia.org/wiki/Basic_reproduction_number

High Contagiousness and Rapid Spread of Severe Acute Respiratory Syndrome Coronavirus 2

Steven Sanche[1], Yen Ting Lin[1], Chonggang Xu, Ethan Romero-Severson, Nick Hengartner, and Ruian Ke✉
Author affiliations: Los Alamos National Laboratory, Los Alamos, New Mexico, USA
Main Article

Figure 5

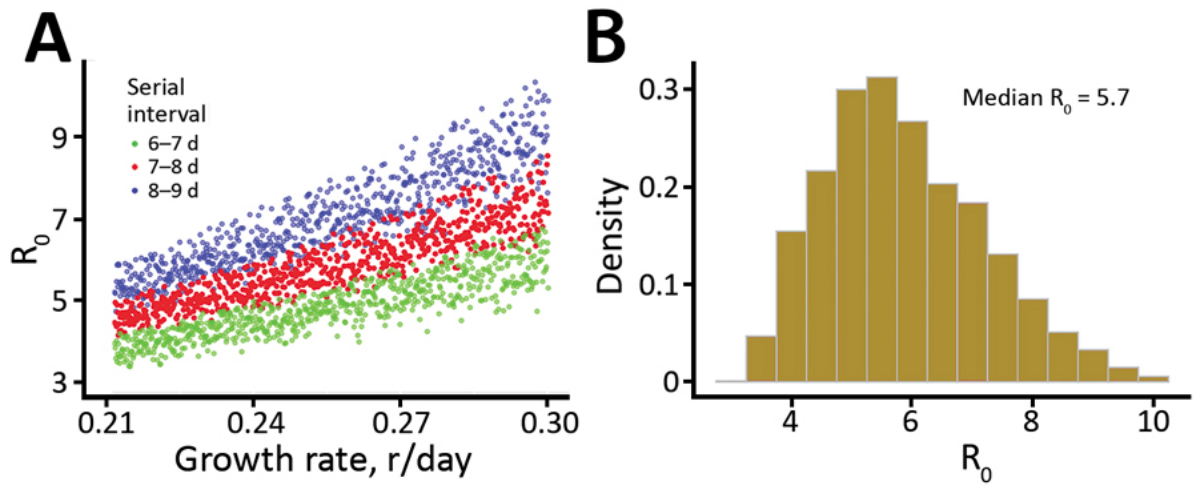


Figure 5. Estimation of the basic reproductive number (R_0), derived by integrating uncertainties in parameter values, during the 2019 novel coronavirus disease outbreak in China. A) Changes in R_0 based on different growth rates and serial intervals. Each dot represents a calculation with mean latent period (range 2.2–6 days) and mean infectious periods (range 4–14 days). Only those estimates falling within the range of serial intervals of interests were plotted. B) Histogram summarizing the estimated R_0 of all dots in panel A (i.e., serial interval ranges of 6–9 days). The median R_0 is 5.7 (95% CI 3.8–8.9).

ACE2 bound to the 2019-nCoV S ectodomain with ~15 nM affinity, which is ~10- to 20-fold higher than ACE2 binding to SARS-CoV S.

10.1126/science.abb2507

Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020

Weekly / March 27, 2020 / 69(12);343-346

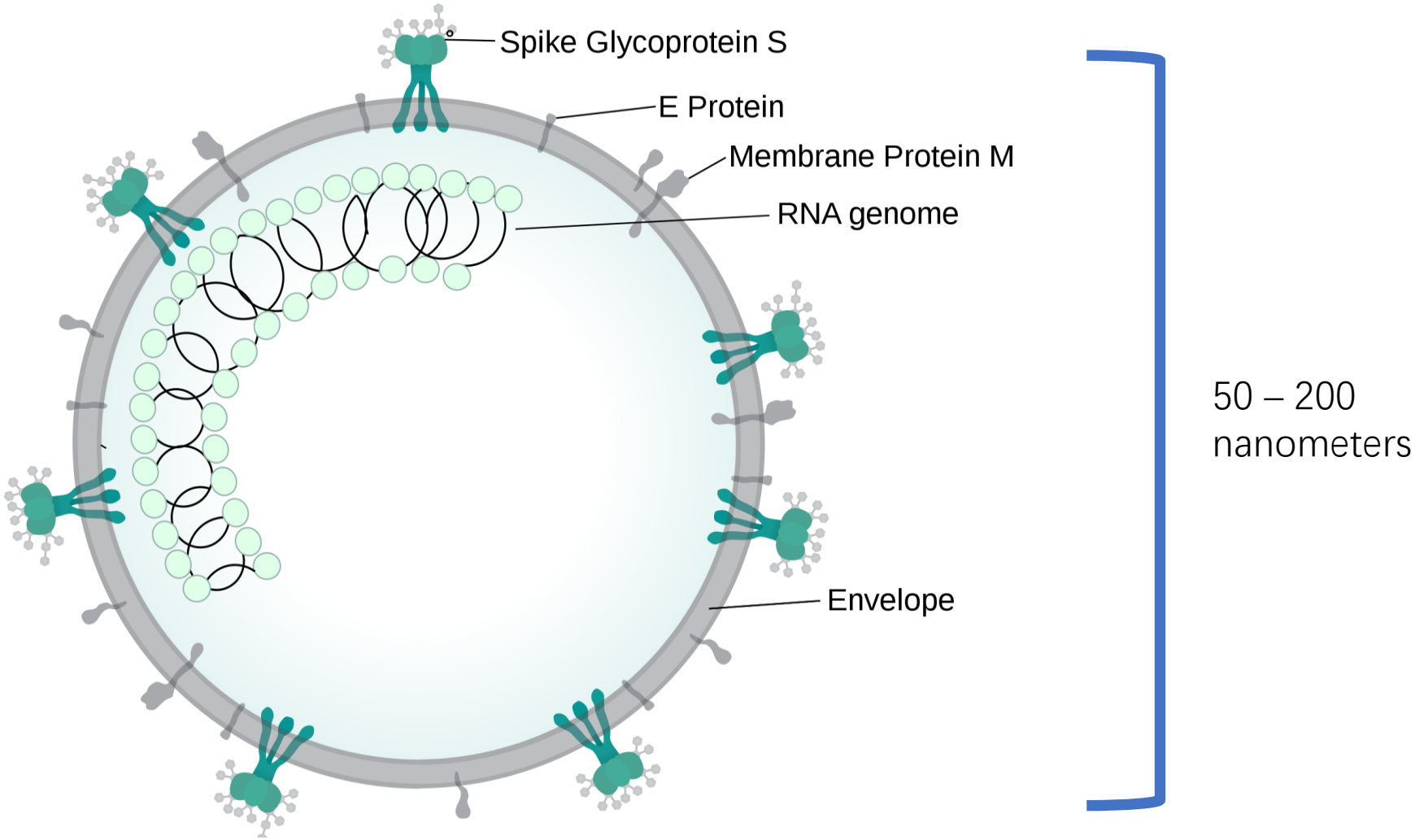
On March 18, 2020, this report was posted online as an MMWR *Early Release.*

Please note: *This report has been corrected.*

CDC COVID-19 Response Team (View author affiliations)

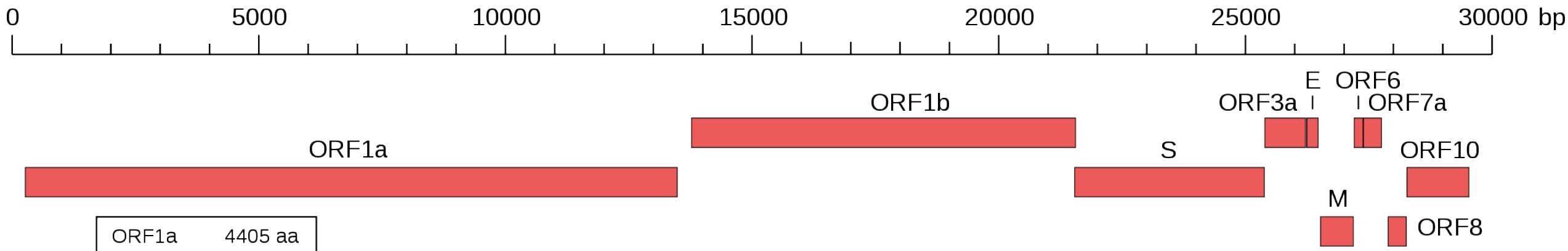
This first preliminary description of outcomes among patients with COVID-19 in the United States indicates that fatality was highest in persons aged ⩾85, ranging from 10% to 27%, followed by 3% to 11% among persons aged 65–84 years, 1% to 3% among persons aged 55-64 years, <1% among persons aged 20–54 years, and no fatalities among persons aged ⩽19 years.

SARS-CoV-2 is a positive-sense single-stranded RNA virus that causes COVID-19



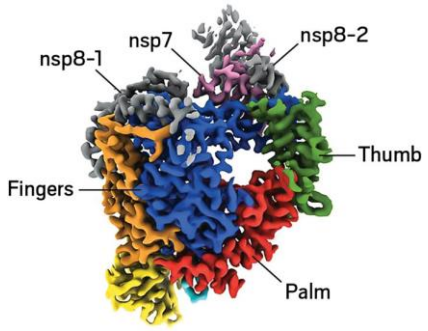
https://en.wikipedia.org/wiki/Coronavirus_disease_2019

SARS-CoV-2 is 29,903 bases and encodes 4 structural proteins (spike, envelope, membrane, nucleocapsid)



ORF1a	4405 aa
ORF1b	2595 aa
S	1282 aa
ORF3a	275 aa
E	75 aa
M	222 aa
ORF6	61 aa
ORF7a	121 aa
ORF8	121 aa
ORF10	419 aa

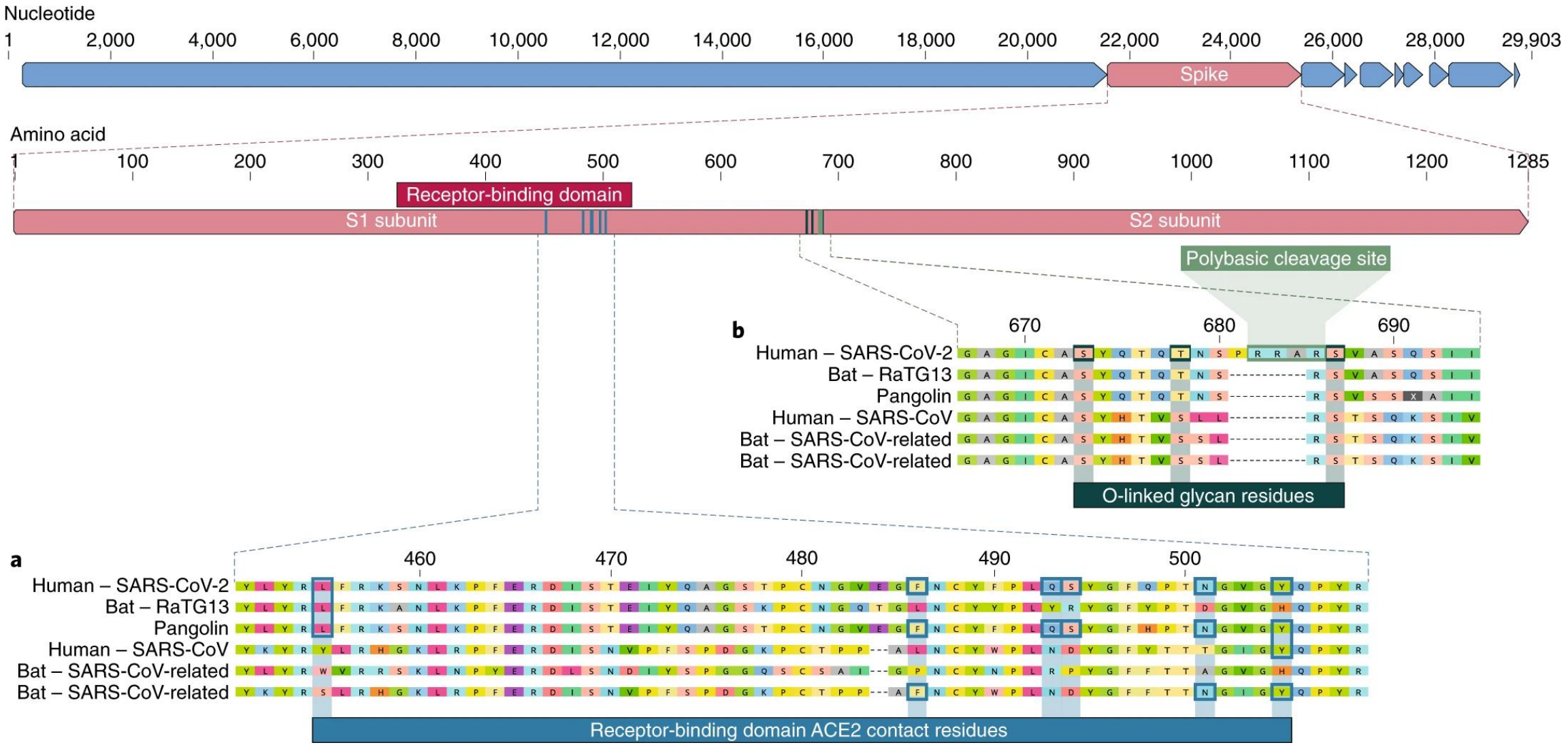
Wuhan-Hu-1 (GenBank MN908947)



RNA dependent RNA polymerase
ORF1a and ORF1b (Remdesivir target)

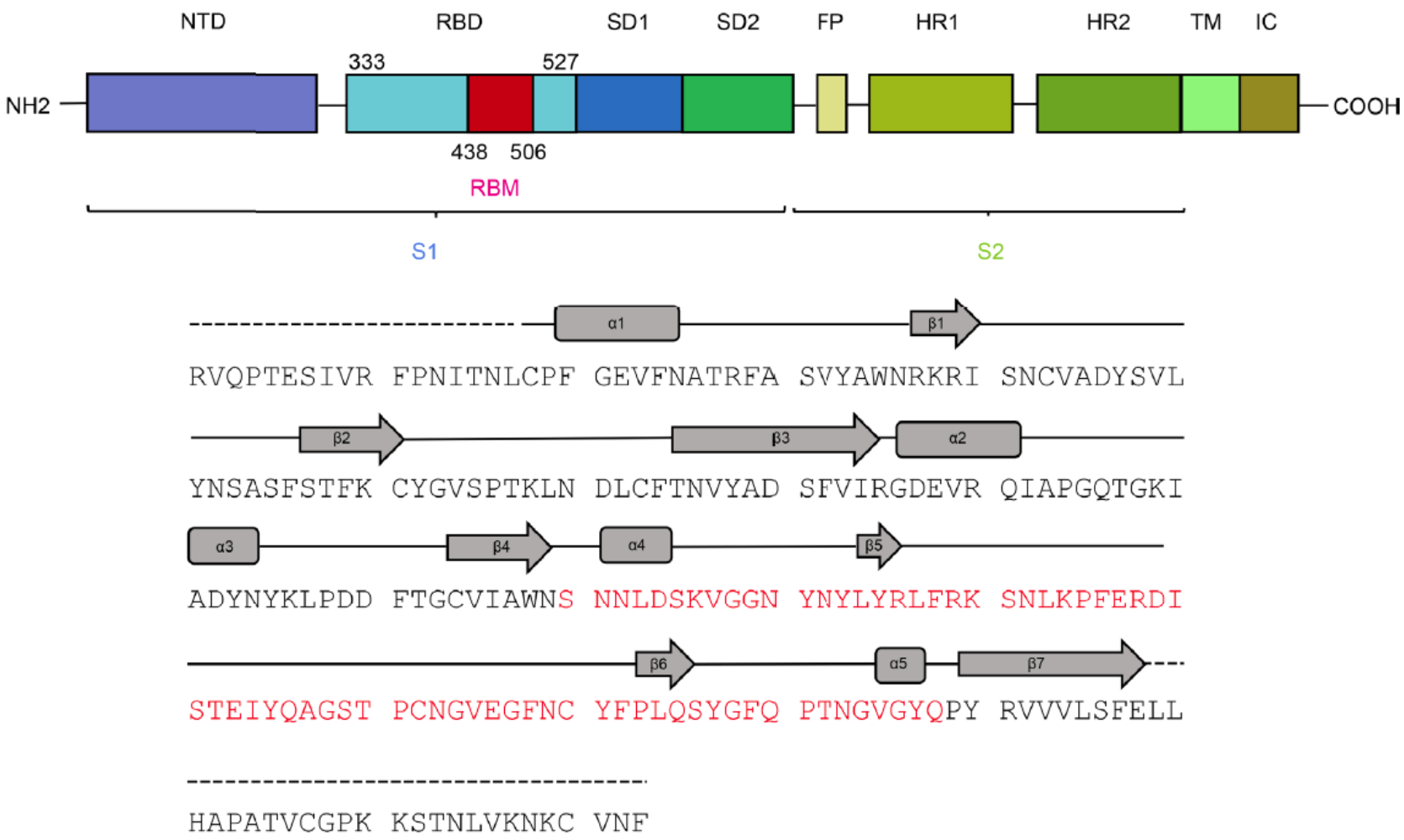
https://en.wikipedia.org/wiki/Severe_acute_respiratory_syndrome_coronavirus_2

Viral genome data suggests SARS-CoV-2 came from an animal



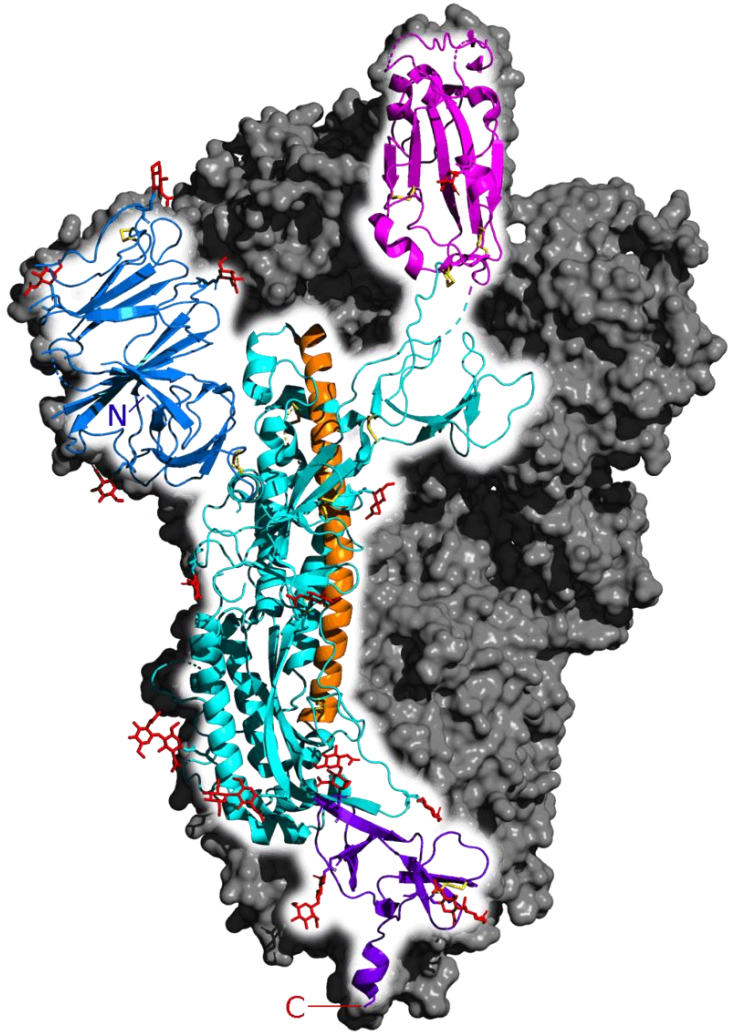
https://doi.org/10.1038/s41591-020-0820-9

The receptor binding domain (RBD) of the spike protein is a primary therapeutic target

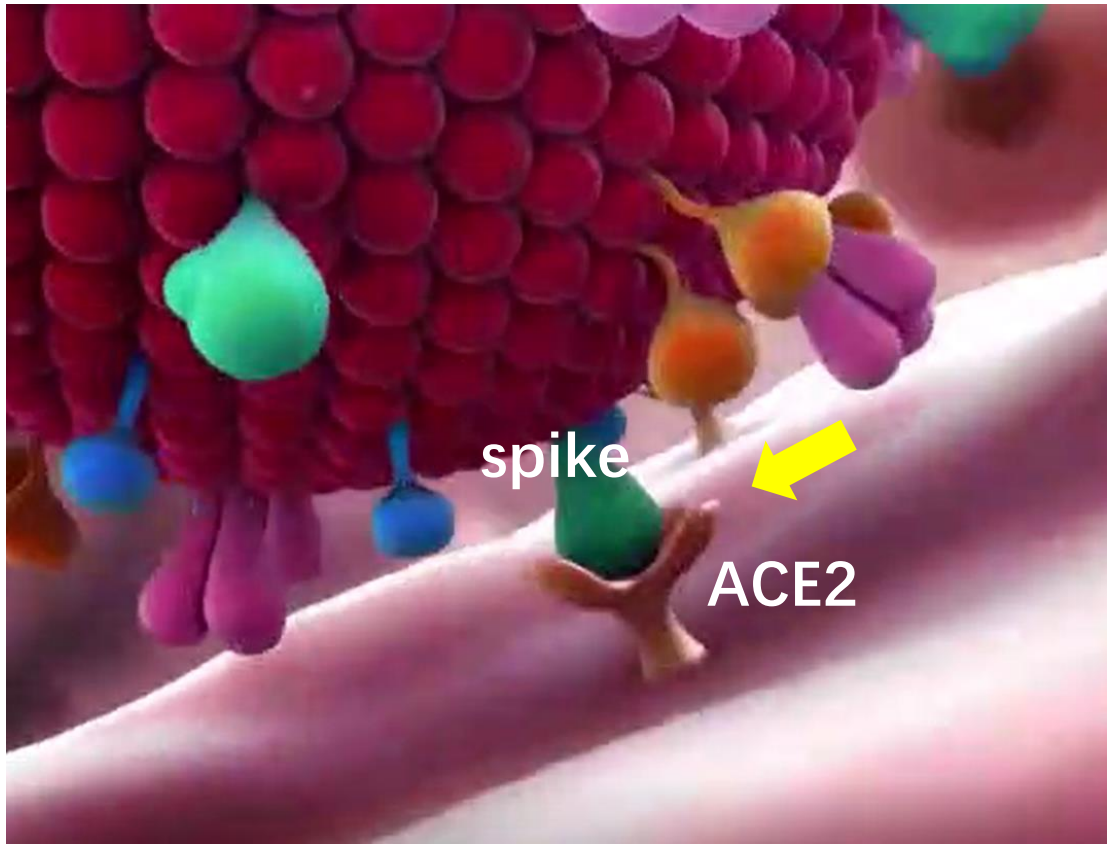


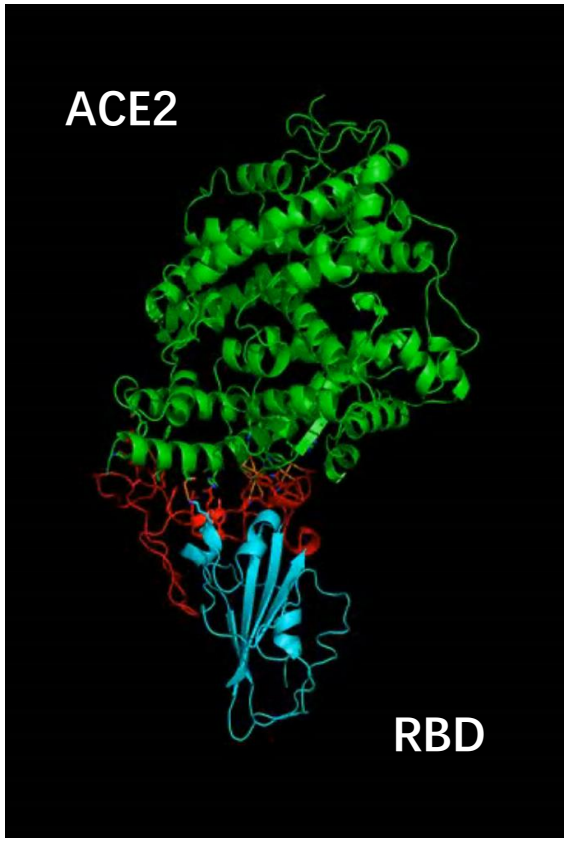
https://doi.org/10.1038/s41586-020-2179-y

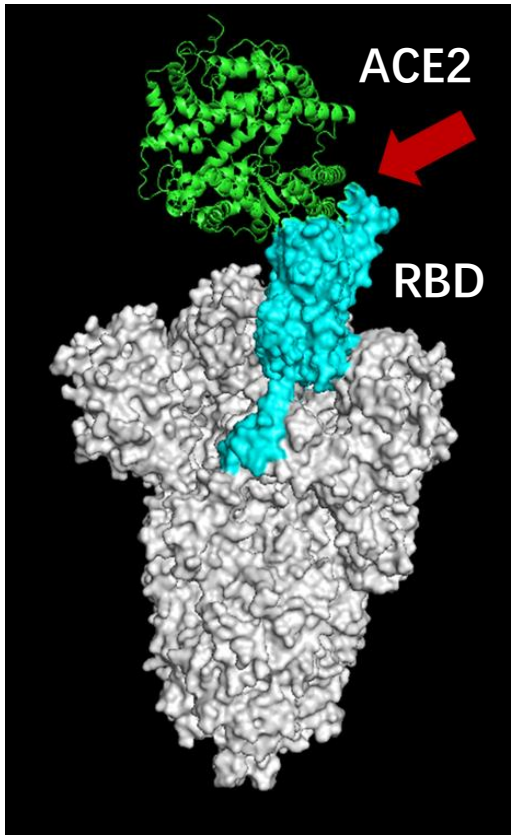
The spike protein (S) trimer ”up” component interacts with ACE2 on host cells



Interaction between RBD of spike and ACE2







Lan et al., Nature March 30, 2020

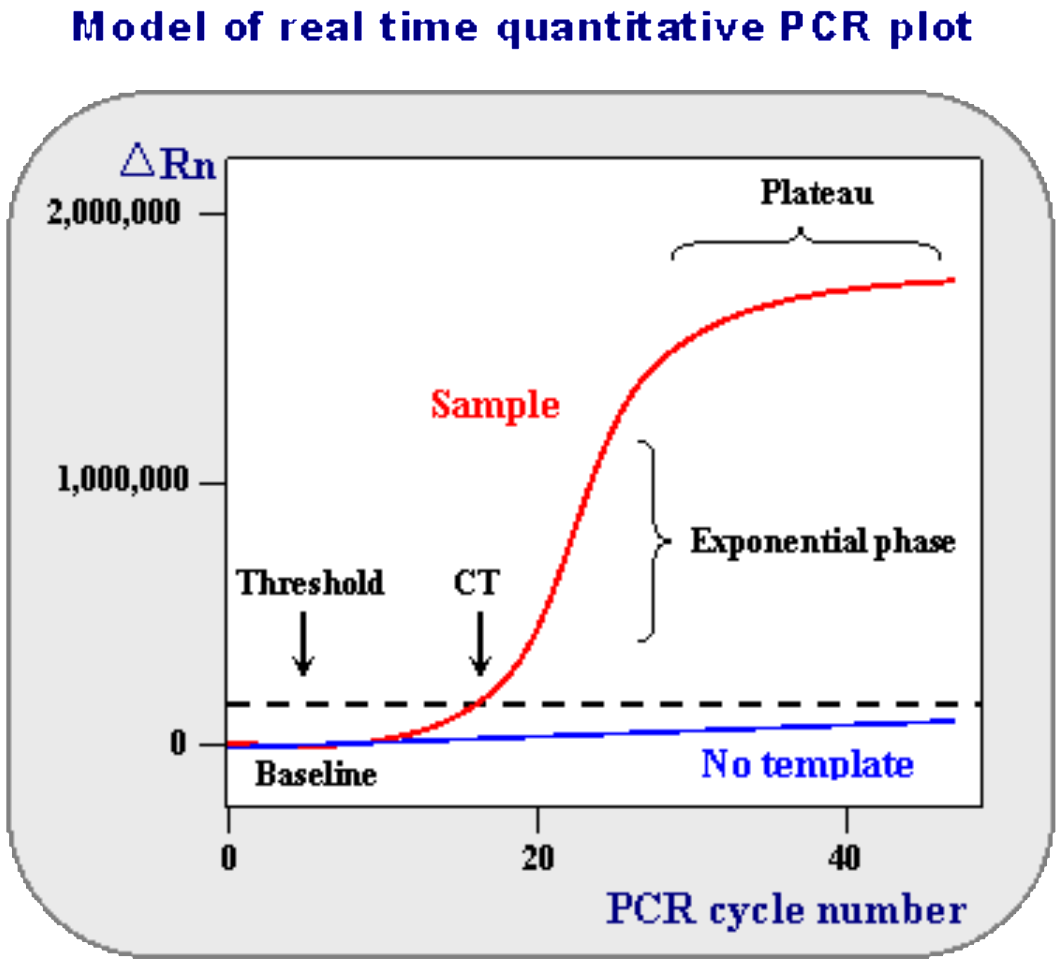
Massachusetts Consortium on Pathogen Readiness

Boston
March 2, 2020

Guangzhou Institute of Respiratory Health (GIRH)
Zhong Nanshan, Director

COVID-19 testing

Real-time quantitative PCR (RT qPCR) is used with primers specific to SARS-CoV-2

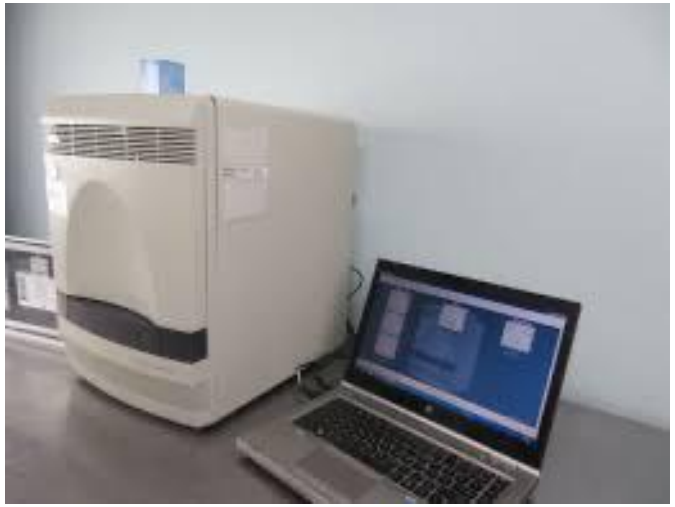


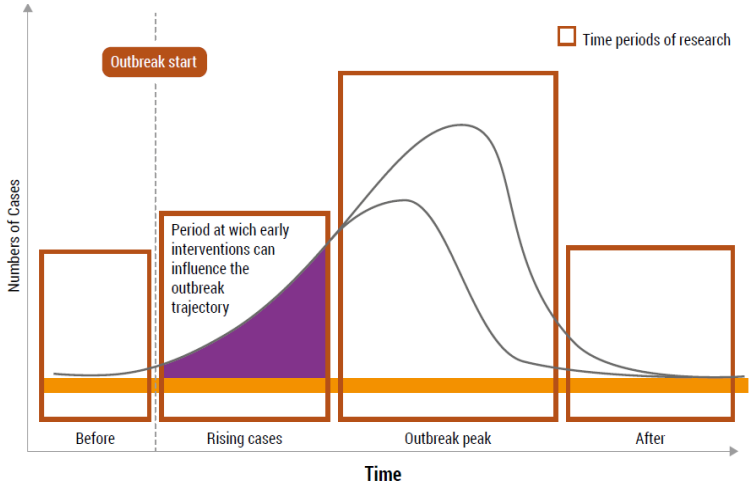
CT is the number of cycles required for a specific sample to cross the detection threshold

https://www.ncbi.nlm.nih.gov/probe/docs/techqpcr/

To monitor and model a novel pandemic, testing needs to be developed fast

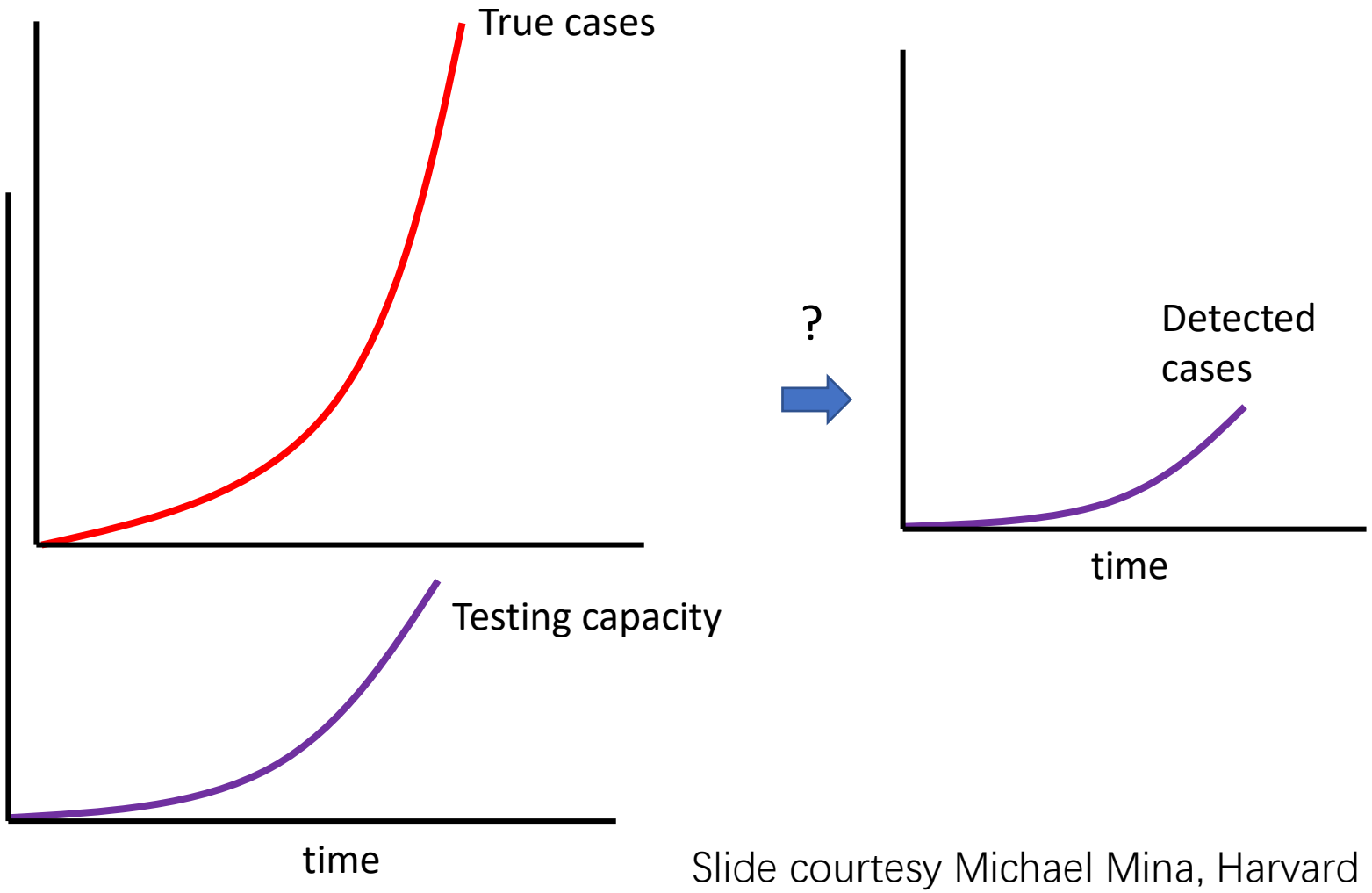
PCR is a good tool to do this





Doubling time (R0 estimation) when testing is being introduced *simultaneous to* epidemic escalation will obscure true epidemic growth

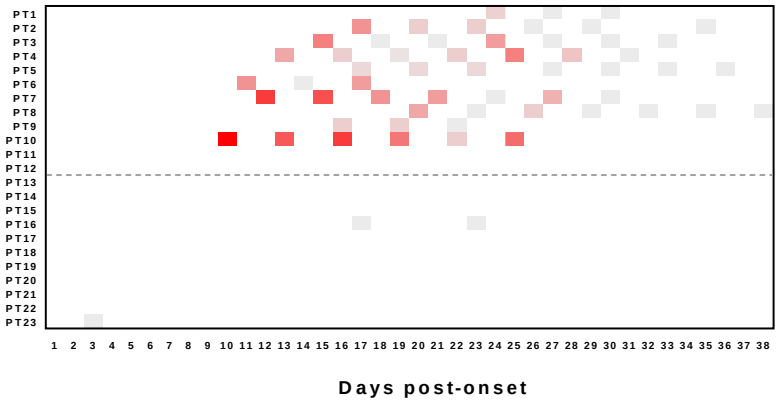
Doubling time estimates can be off and estimated cases off by orders of magnitude



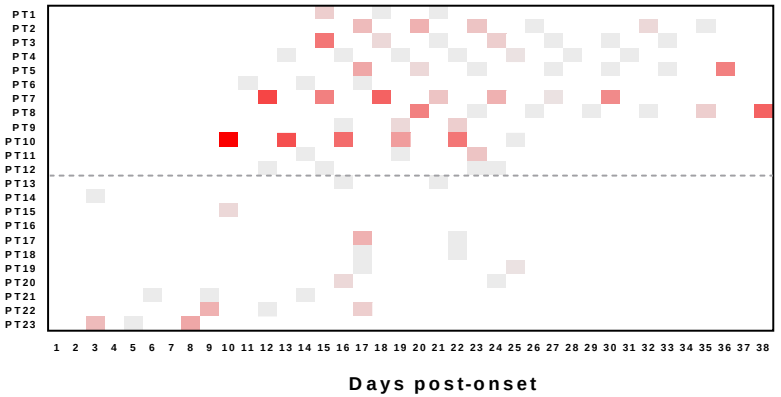
Slide courtesy Michael Mina, Harvard

Prolonged viral shedding from multiple sites in severely ill patients

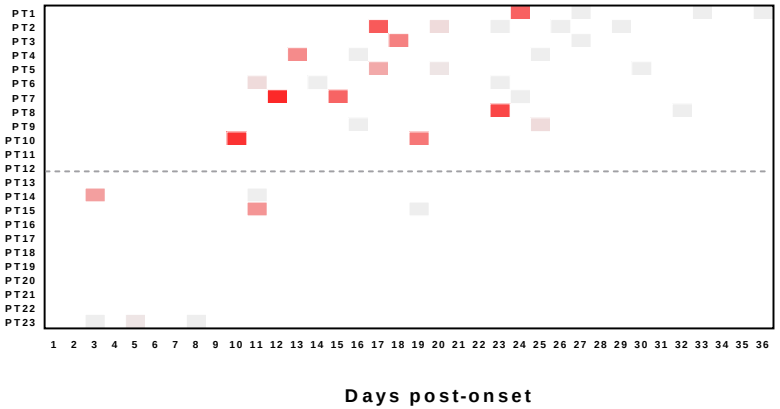
A: Nasal swab



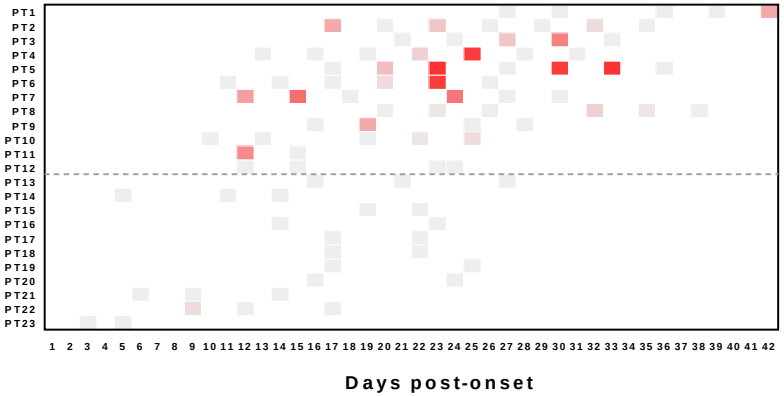
B: Pharyngeal swab



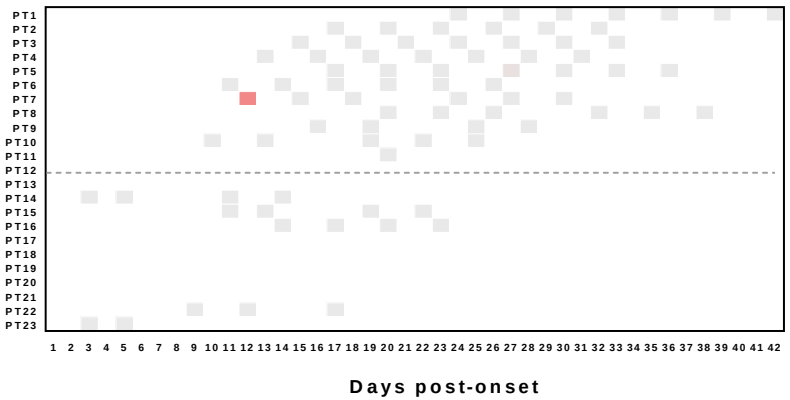
C: Sputum



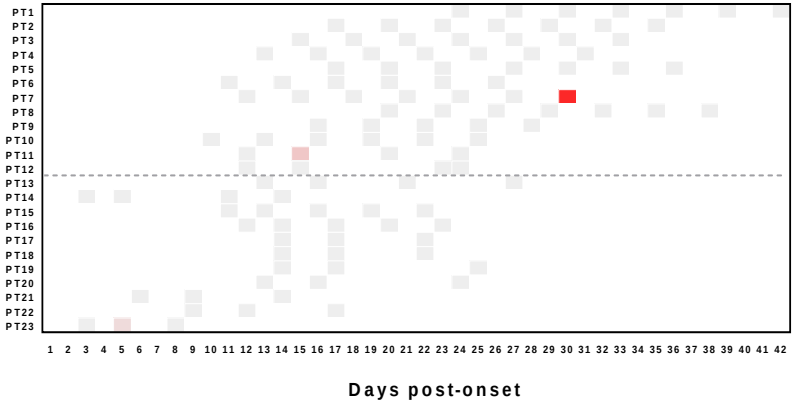
D: Feces



E: Urine



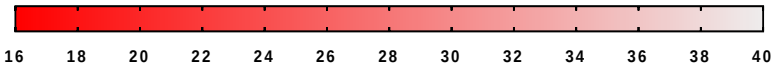
F: Blood



Red = Positive Gray =Negative

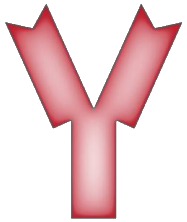
PCR cycles to reach positive threshold

Ct value



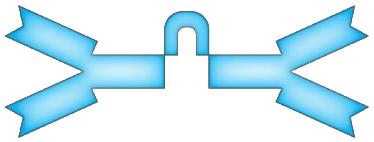
Jincun Zhao, Guangzhou Institute of Respiratory Health

Antibody isotypes



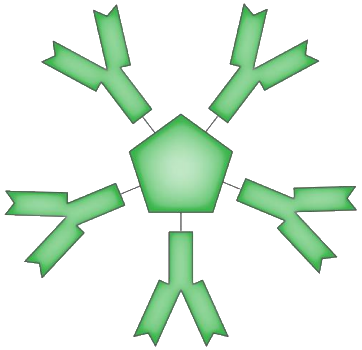
Monomer
IgD, IgE, IgG

Affinity matured antibody specific to target
Enhance phagocytosis of bound pathogens by macrophages
Can cause antibody-dependent cell-mediated cytotoxicity (ADCC)



Dimer
IgA

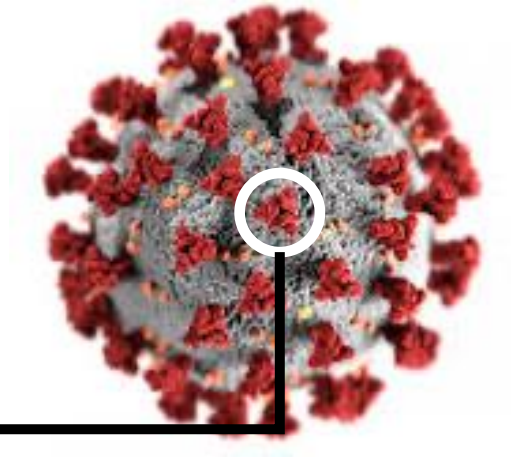
Secreted antibodies – gut, mucus, tears, saliva, milk
Can agglutinate pathogens to enhance their clearance

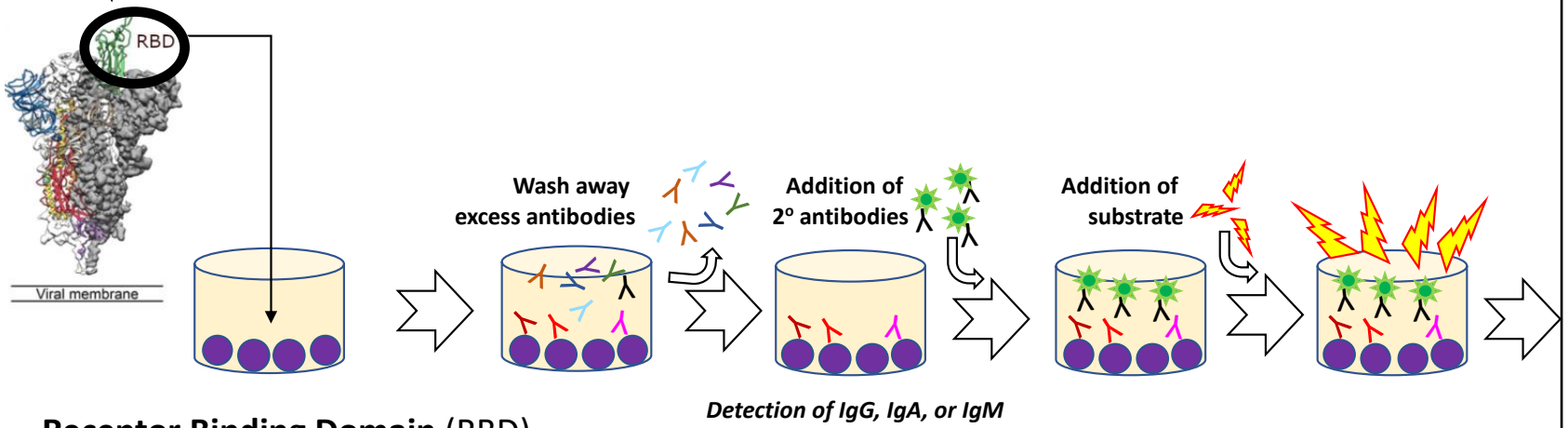


Pentamer
IgM

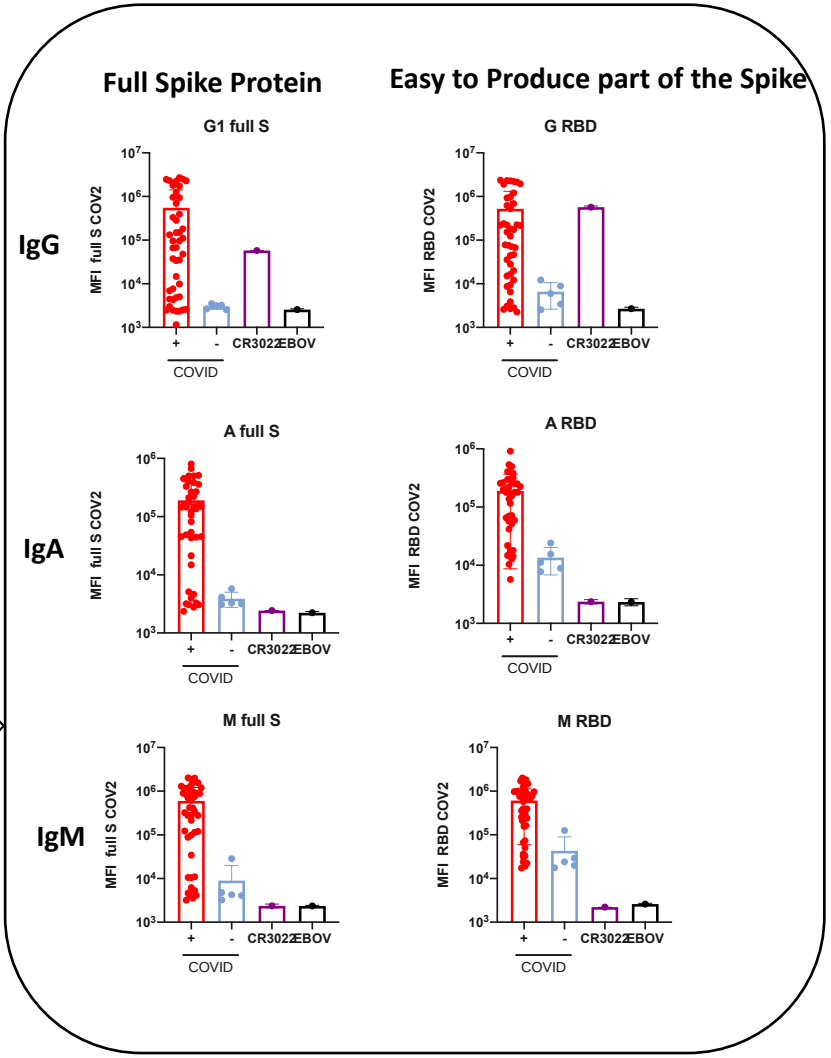
Low affinity antibodies that are expressed early
Activate the innate immune system
Can agglutinate pathogens to enhance their clearance

MGH/Ragon ELISA



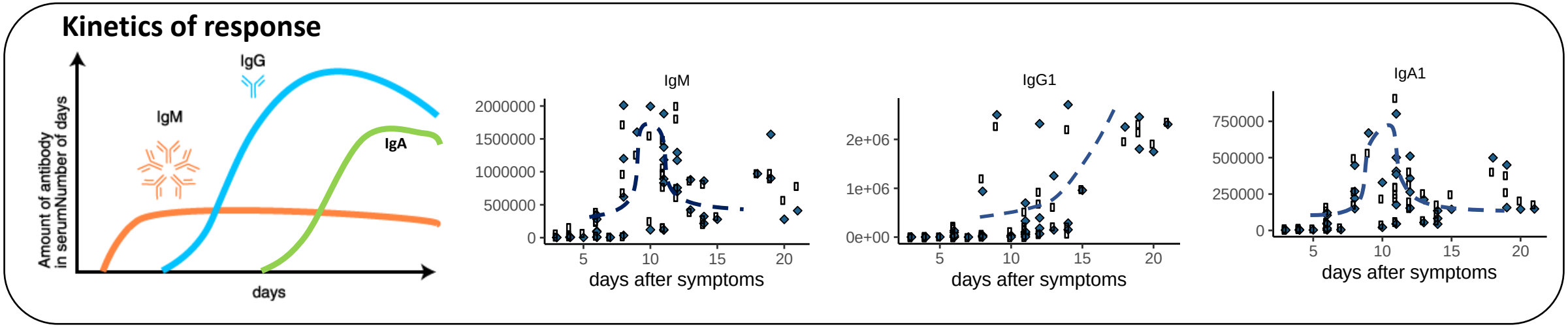


Receptor Binding Domain (RBD) of the CoV2 Spike (S)



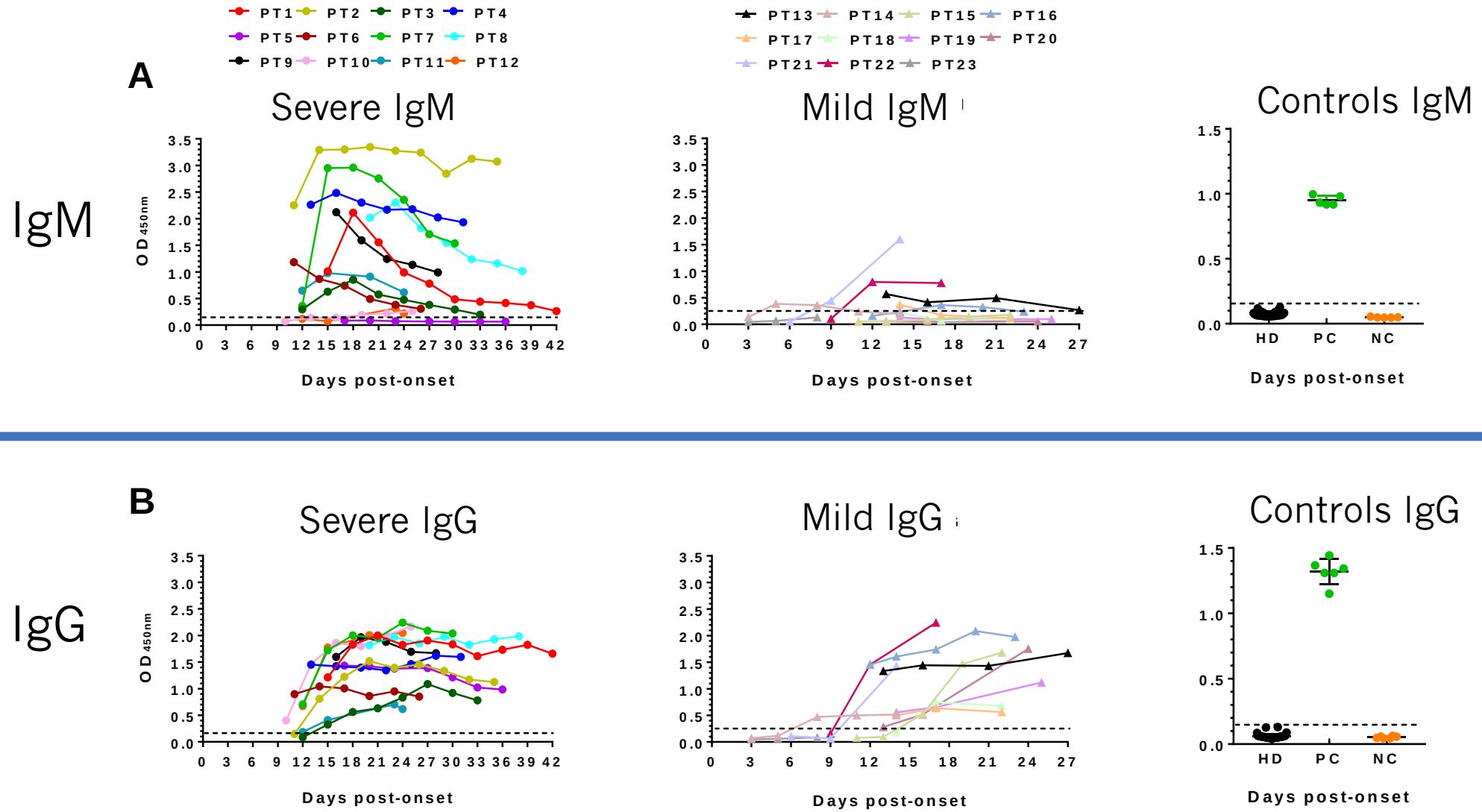
Slide courtesy Galit Alter, Ragon Institute

Sensitivity and unusual immune patterns



Slide courtesy Galit Alter, Ragon Institute

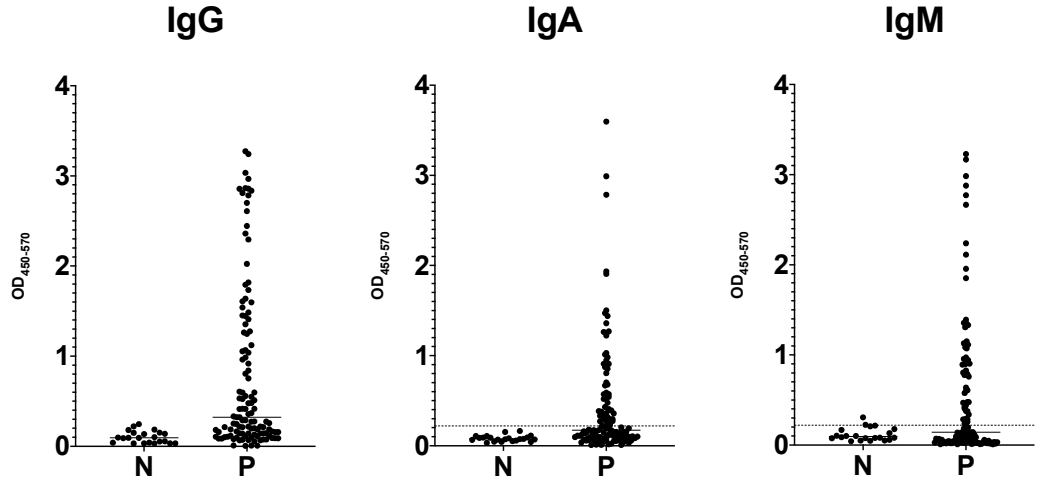
Mild patients have lower IgM responses against SARS-CoV-2



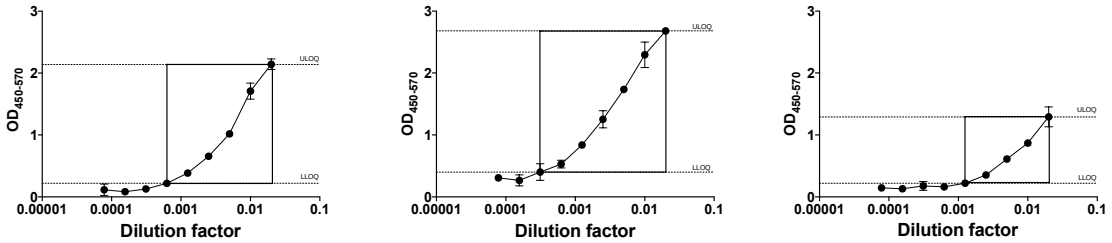
Jincun Zhao, Guangzhou Institute of Respiratory Health

Optimizing for robustness

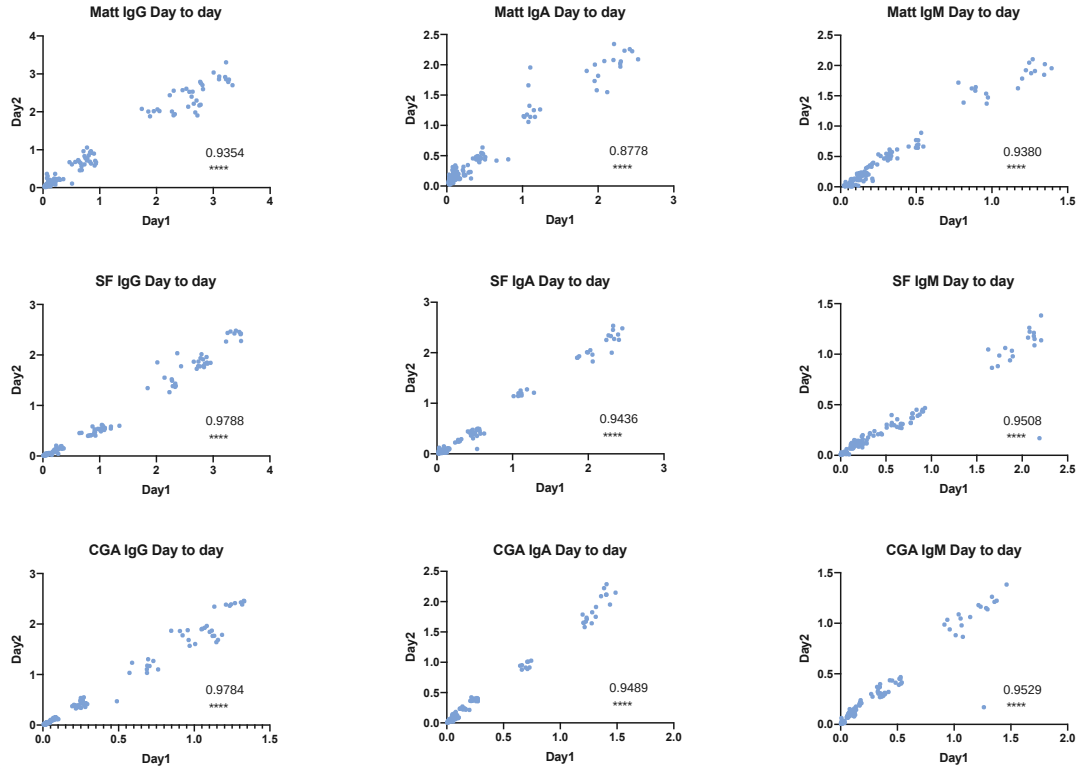
Defining background



Defining the linear range for sampling



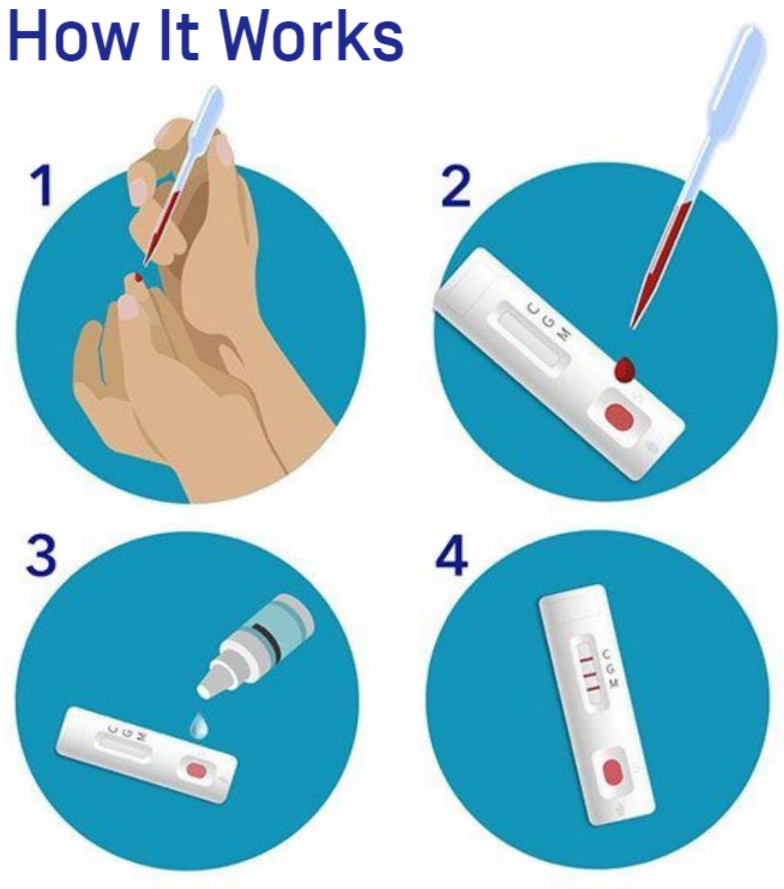
Defining precision across assays/operators



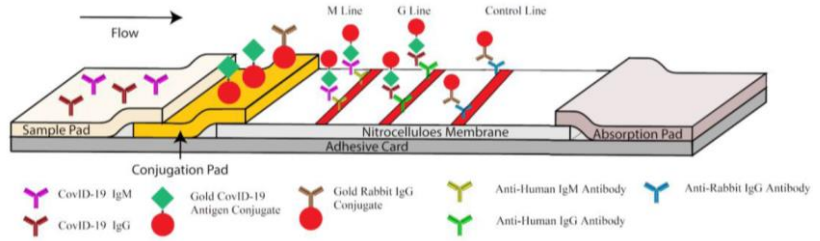
Slide courtesy Galit Alter, Ragon Institute

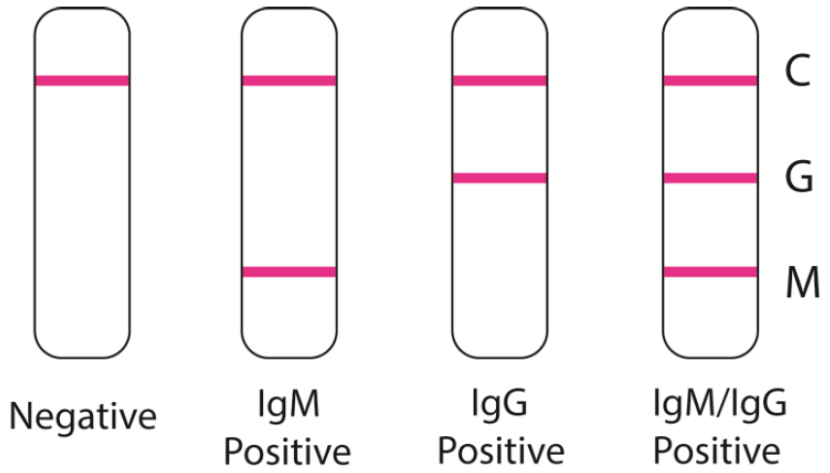
Use of the MGH-Ragon COVID-19 serologic test

Point-of-care rapid tests to detect COVID-19 IgM and IgG



Lateral flow assay





BioMedomics
TESTS FOR LIFE

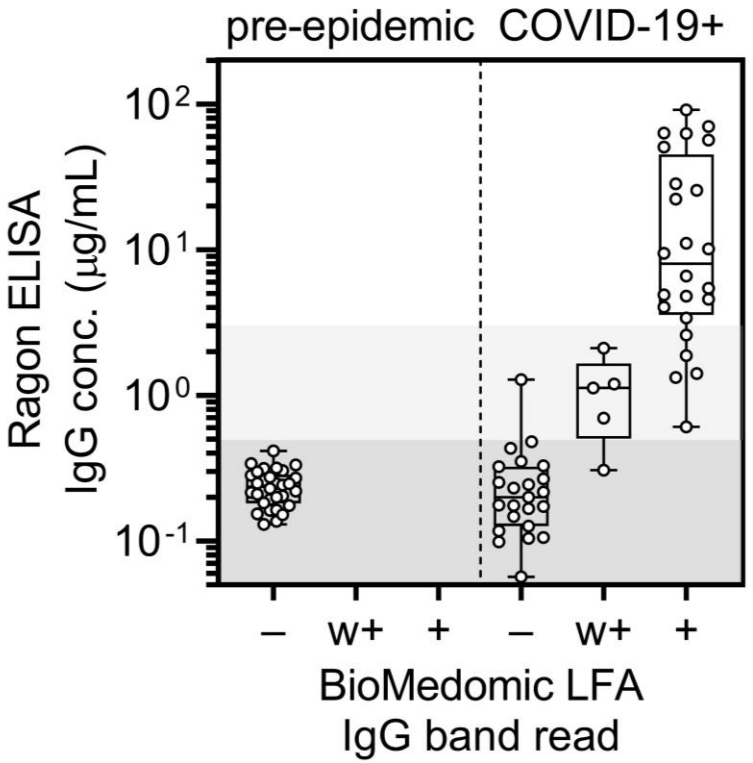
Country of Origin: China
Supplier: BioMedomics, NC;
Henry Schein/BD, USA
Cost: $8; $14 per assay

Slide courtesy Wilfredo Garcia Beltran

Use of the MGH-Ragon COVID-19 serologic test

Point-of-care rapid tests to detect COVID-19 IgM and IgG

LFA versus ELISA



Sensitivity and specificity

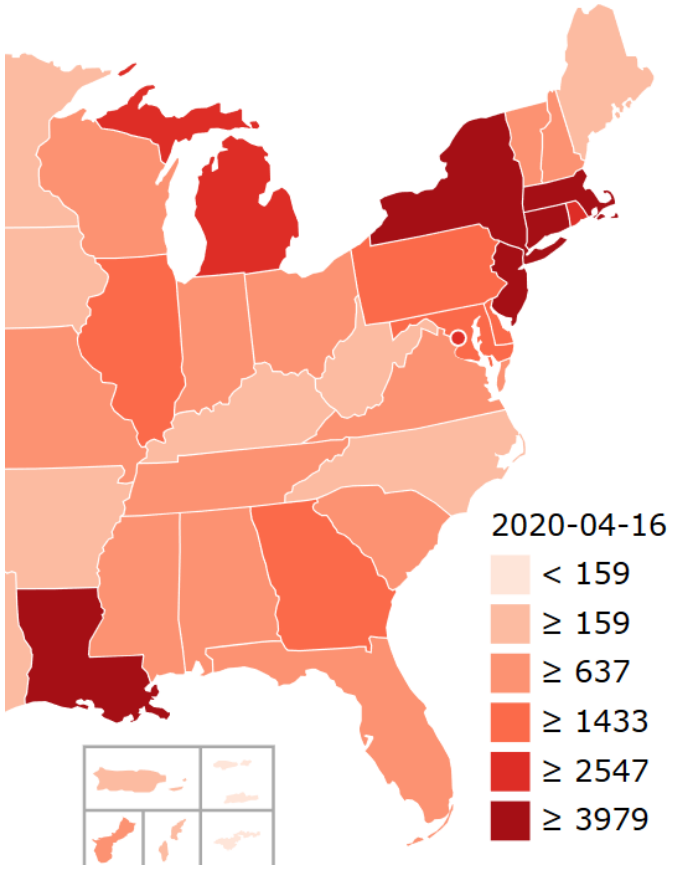
	BioMedomics			ELISA
	IgG	IgM	IgG-IgM	IgG
Specificity (*n* = 60)	100	100	100	100
Sensitivity (*n* = 57)	56	60	65	56
Sensitivity (≤7 days) (*n* =14)	7	21	21	7
Sensitivity (>7 days) (*n* = 43)	72	72	79	74
Sensitivity (>10 days) (*n* = 33)	73	76	82	79
Sensitivity (>12 days) (*n* = 20)	85	80	90	89
Sensitivity (>14 days) (*n* = 10*)	90	80	90	90

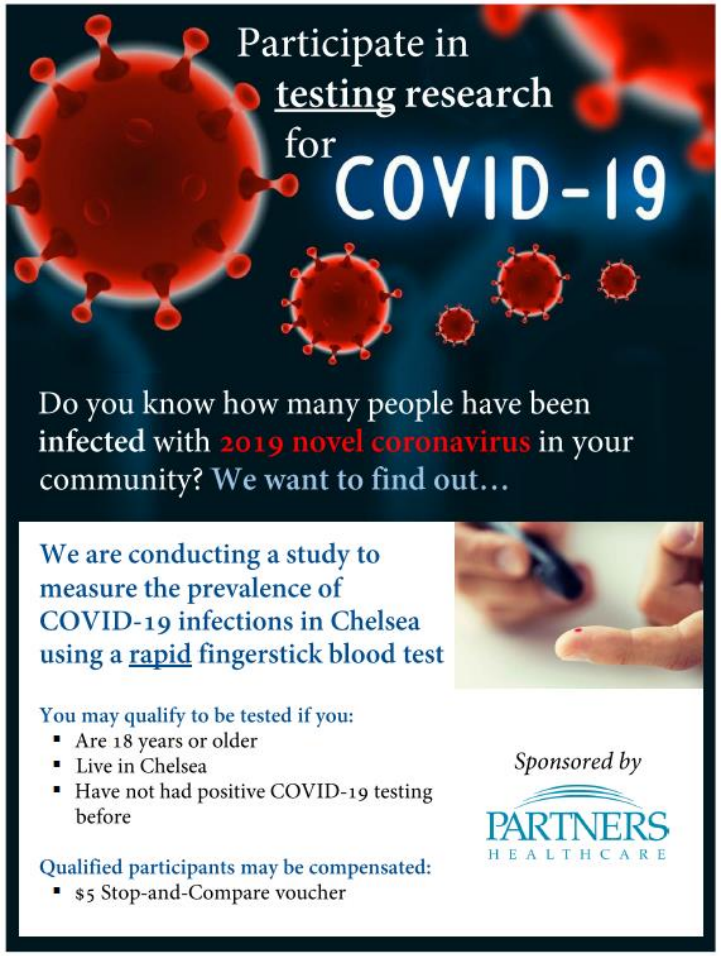
**One patient with no antibody response >14 days is immunocompromised*

Slide courtesy Wilfredo Garcia Beltran

Use of the MGH-Ragon COVID-19 serologic test

COVID-19 point-of-care rapid tests in the community



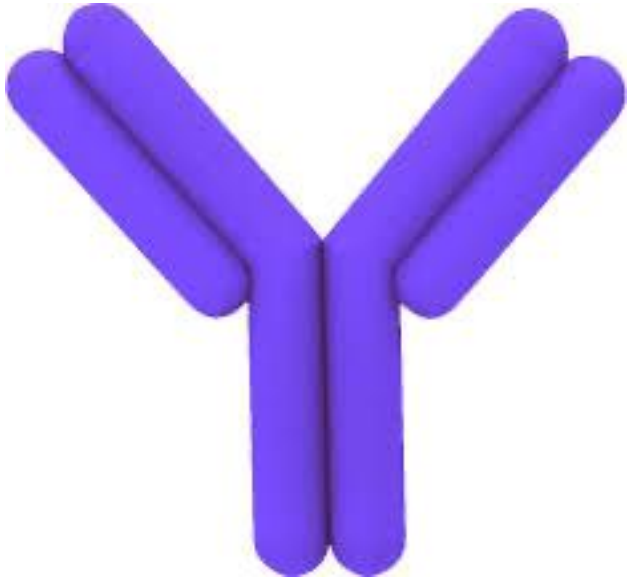


The Boston Globe

Boston Herald

Slide courtesy Wilfredo Garcia Beltran

Critical thought



≠ **Immunity**

= **Exposure**

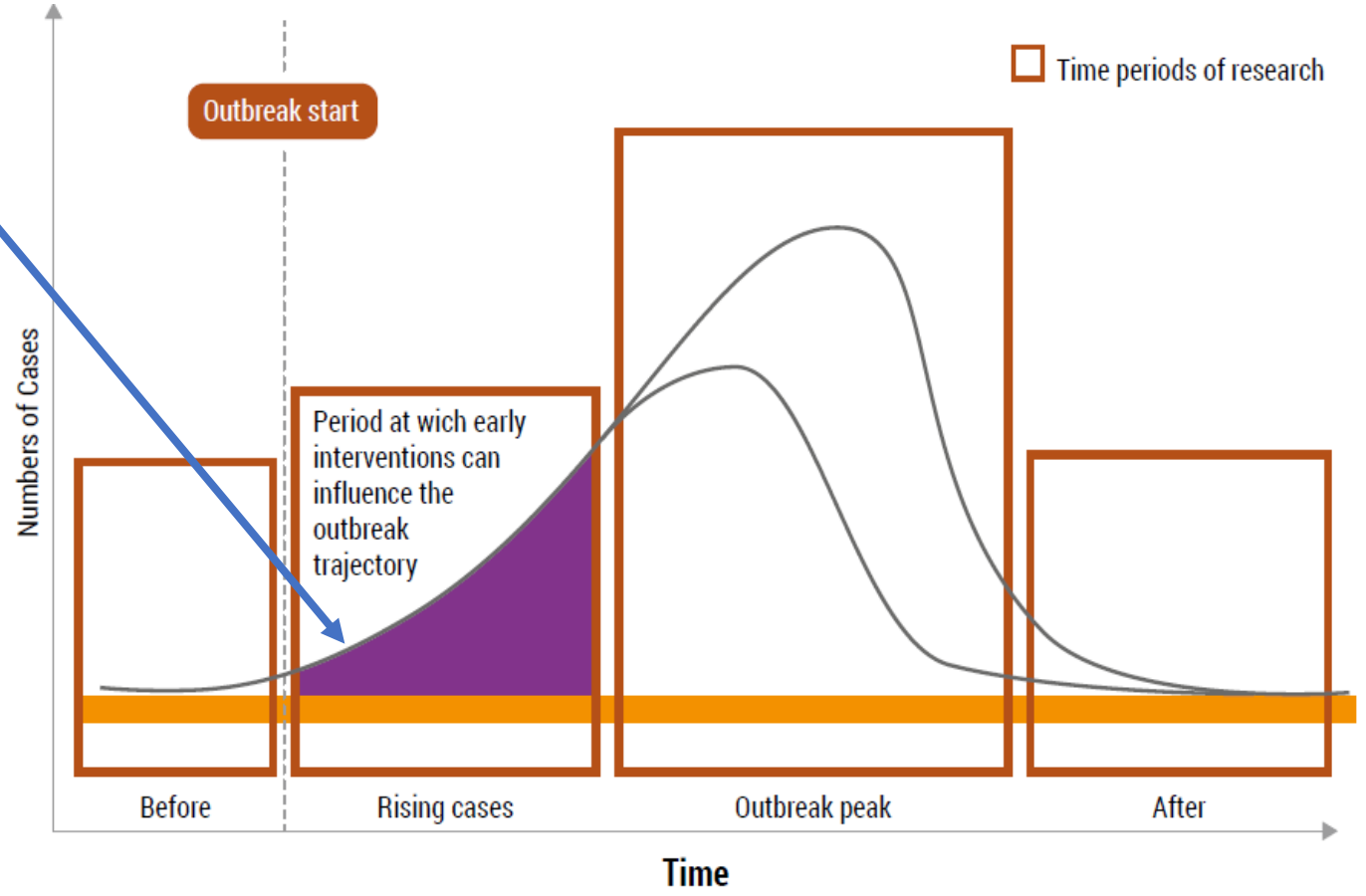
Sero-Epidemiological studies are needed to establish a threshold of immunity.

Slide courtesy Galit Alter, Ragon Institute

COVID-19 epidemiology

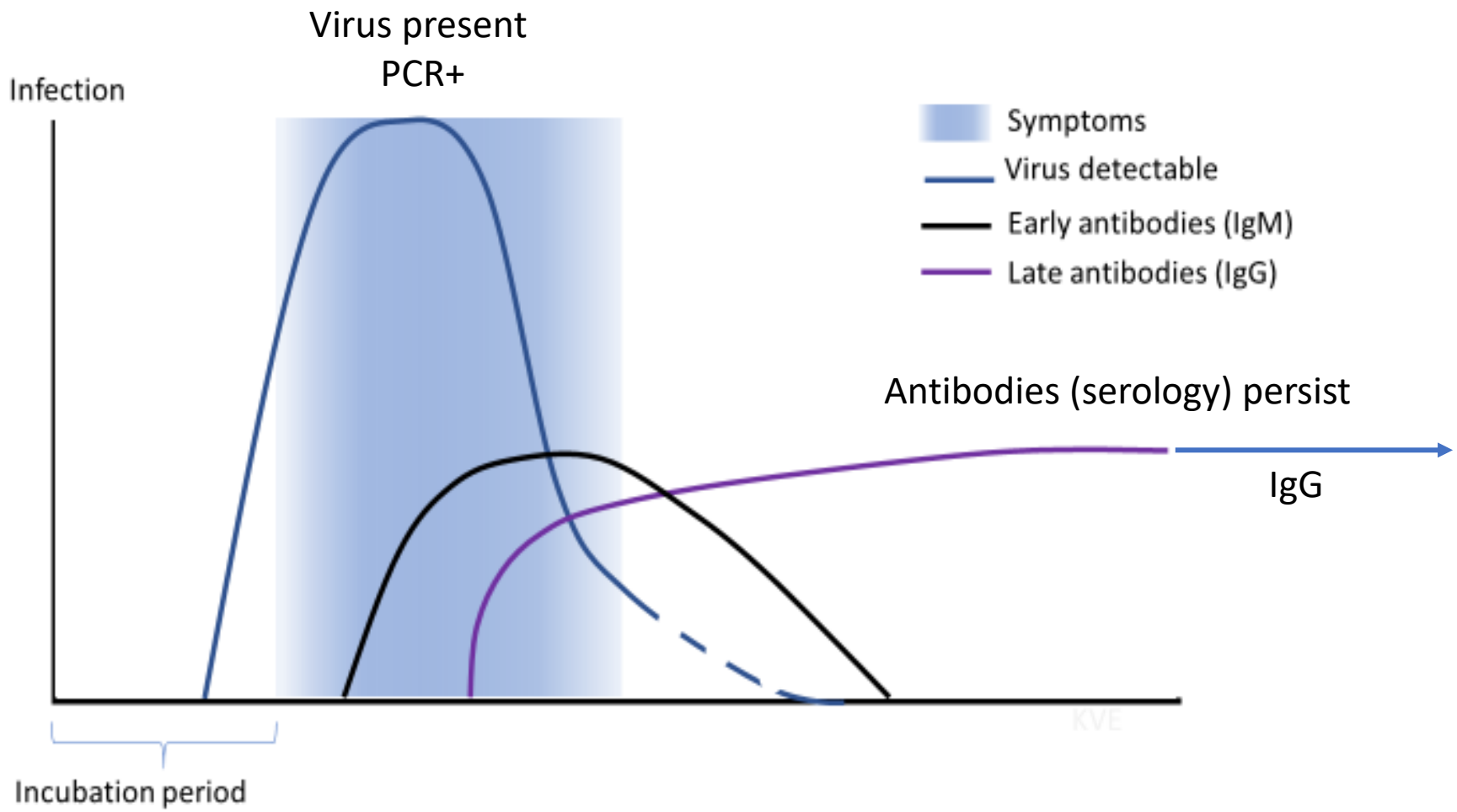
Disease Epidemiology

Monitor this slope carefully



- Helps to understand disease spread and plan strategies accordingly
- SARS-CoV-2 → immense need to estimate in real time the trajectory of an emerging epidemic…
- Monitoring the number of cases helps define **key parameters** to model the epidemic
- R0 (Basic Reproductive Number) → inferred in part through 'doubling time' of cases

Serology (antibody testing) helps fill in the gaps

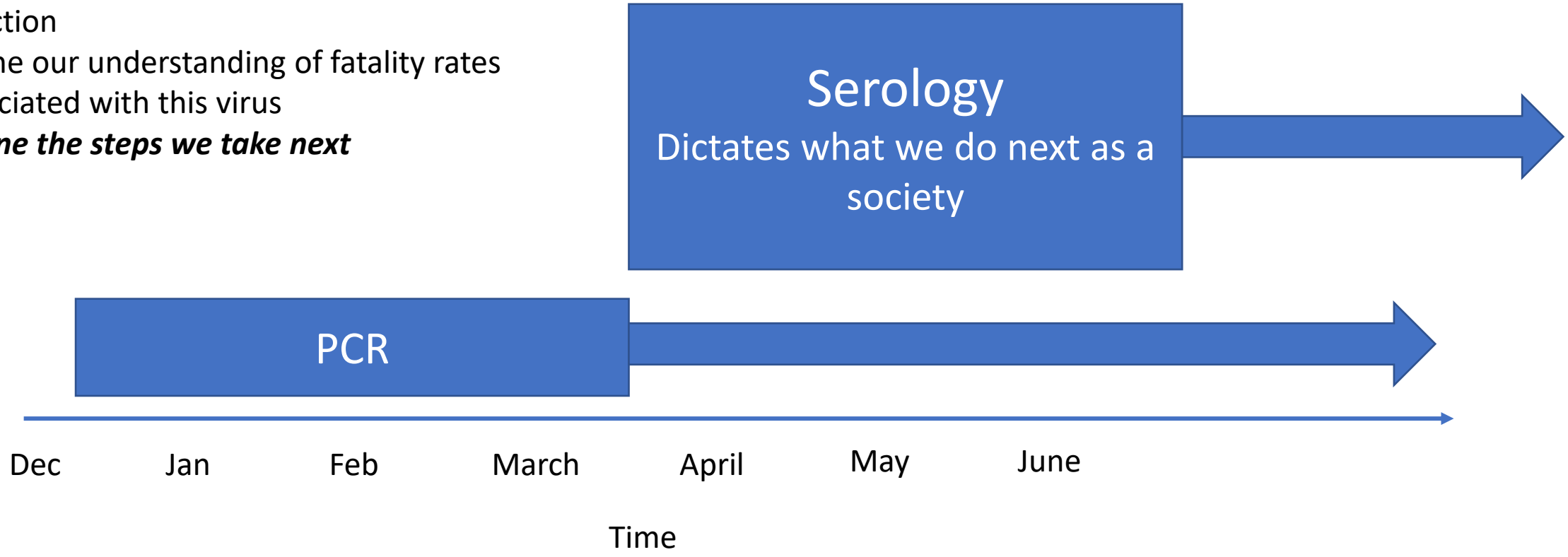


Virus exists for a short window (diagnostics and prevalence)
Antibodies exist for months-years (cumulative incidence)

Slide courtesy Michael Mina, Harvard

COVID 19 ‘simple’ timeline through testing lens

- **Serology will fill in the gaps:**
- Allow us to know the true incidence of of infection
- Refine our understanding of fatality rates associated with this virus
- ***Define the steps we take next***



Slide courtesy Michael Mina, Harvard

Mid-epidemic phase: achieving herd immunity

Rule of thumb:

- $1 - 1/R_0$ = proportion of entire population that needs to be immune to control spread.

- How do we know the proportion immune?
  - Sampling strategy needs to be stratified on age, spatial areas, gender
  - Builds on modeling structure to estimate age-specific force of infection from seroprevalence data.

Later phase: Targeting vaccination campaigns

Slide courtesy Megan Murray, Harvard

Can we use serological testing to determine when sub-groups can return to work?

- In principal, yes, with some caveats:

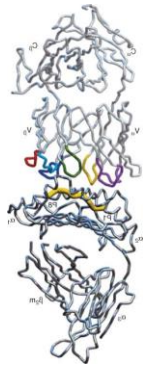
  - Serological tests perform best in high prevalence settings.
  - Unclear if serological tests correlate with immune protection.
  - Not clear if/when antibody-mediated protection wanes.
  - Challenging to maintain “closed” community if susceptibles return with people who are immune.

Slide courtesy Megan Murray, Harvard

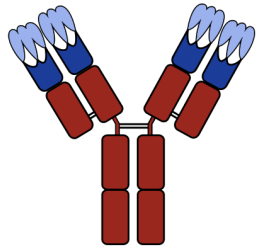
Vaccines for SARS-CoV-2

Today’s deep learning methods

- Vaccine design



- Antibody discovery and improvement

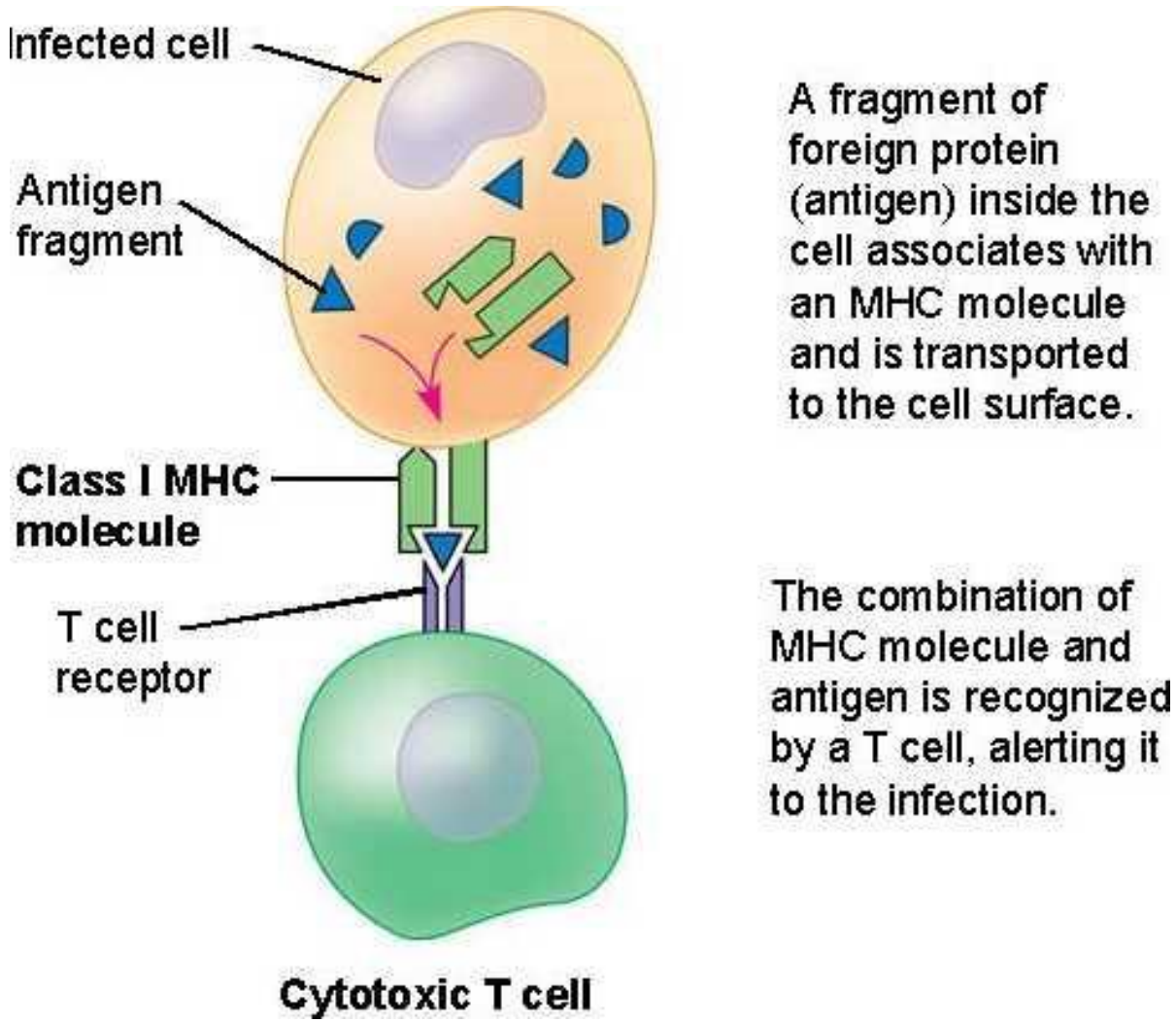


Vaccines educate the adaptive immune system to prepare it to defend against viral infection

Vaccine strategy	Advantages	Disadvantages
Inactivated virus vaccines	Easy to prepare; safety; high-titer neutralizing antibodies	Potential inappropriate for highly immunosuppressed individuals
Attenuated virus vaccines	Rapid development; induce high immune responses	Phenotypic or genotypic reversion possible; can still cause some disease
Subunit vaccines	High safety; consistent production; can induce cellular and humoral immune responses; high-titer neutralizing antibodies	High cost; lower immunogenicity; require repeated doses and adjuvants
Viral vector vaccines	Safety; induces high cellular and humoral immune responses	Possibly present pre-existing immunity
DNA vaccines	Easier to design; high safety; high-titer neutralizing antibodies	Lower immune responses in humans; repeated doses may cause toxicity
mRNA vaccines	Easier to design; high degree of adaptability; induce strong immune responses	Highly unstable under physiological conditions

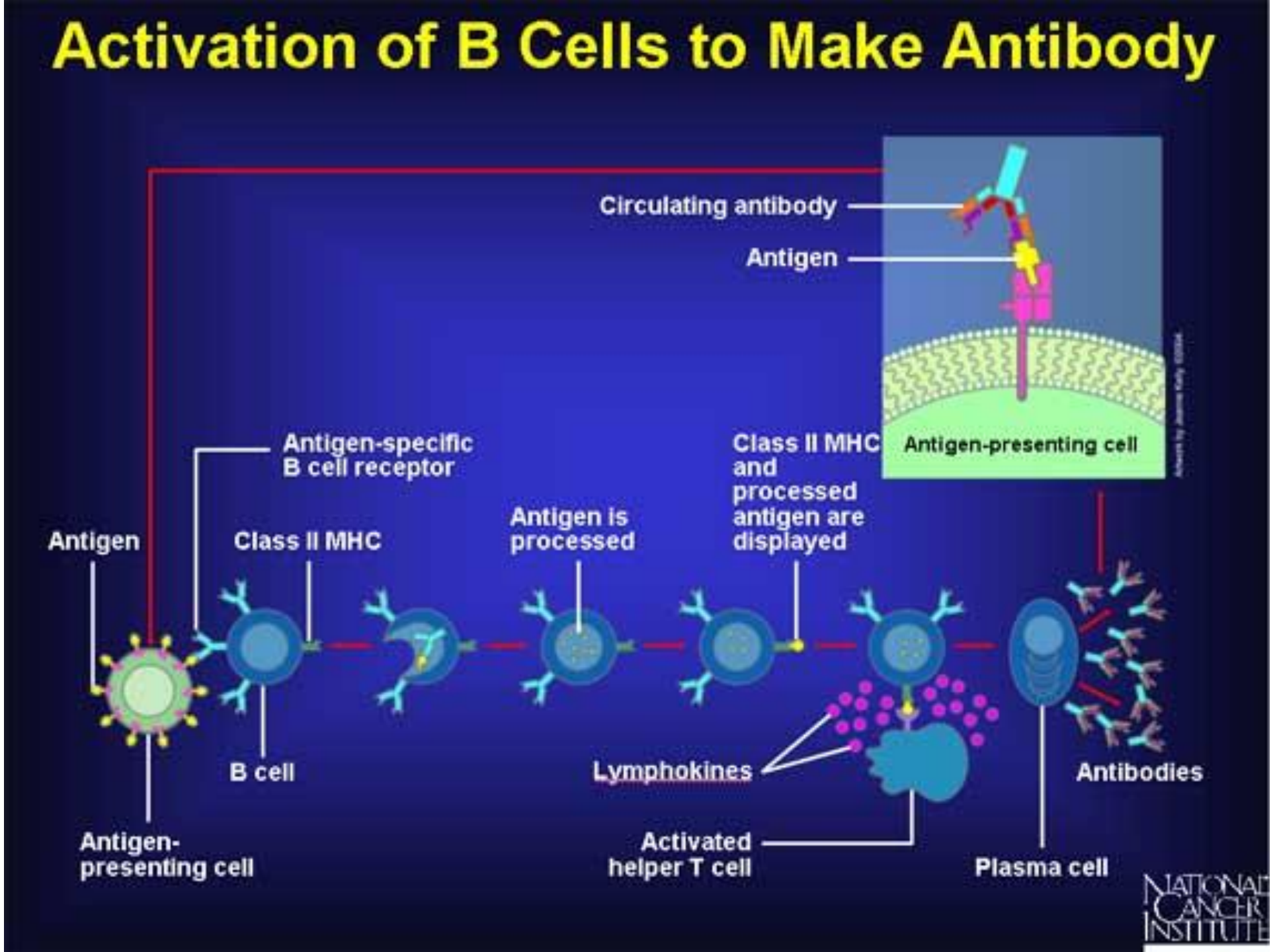
https://doi.org/10.1038/s41541-020-0170-0

Cytotoxic T lymphocytes (CTLs) recognize non-self peptides displayed by infected cells



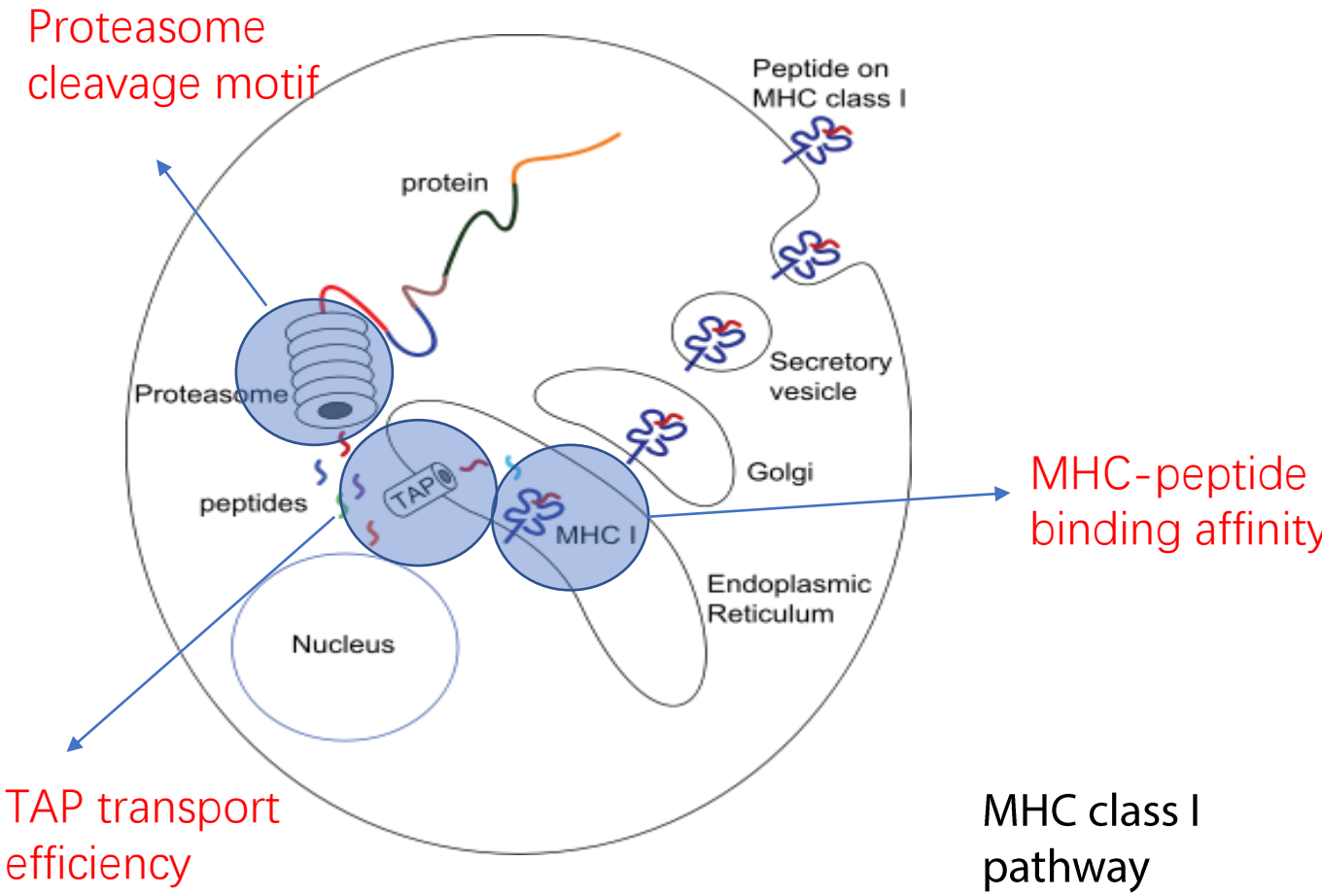
http://bio1151b.nicerweb.net/Locked/media/ch43/t_cells.html

Activation of B Cells to Make Antibody

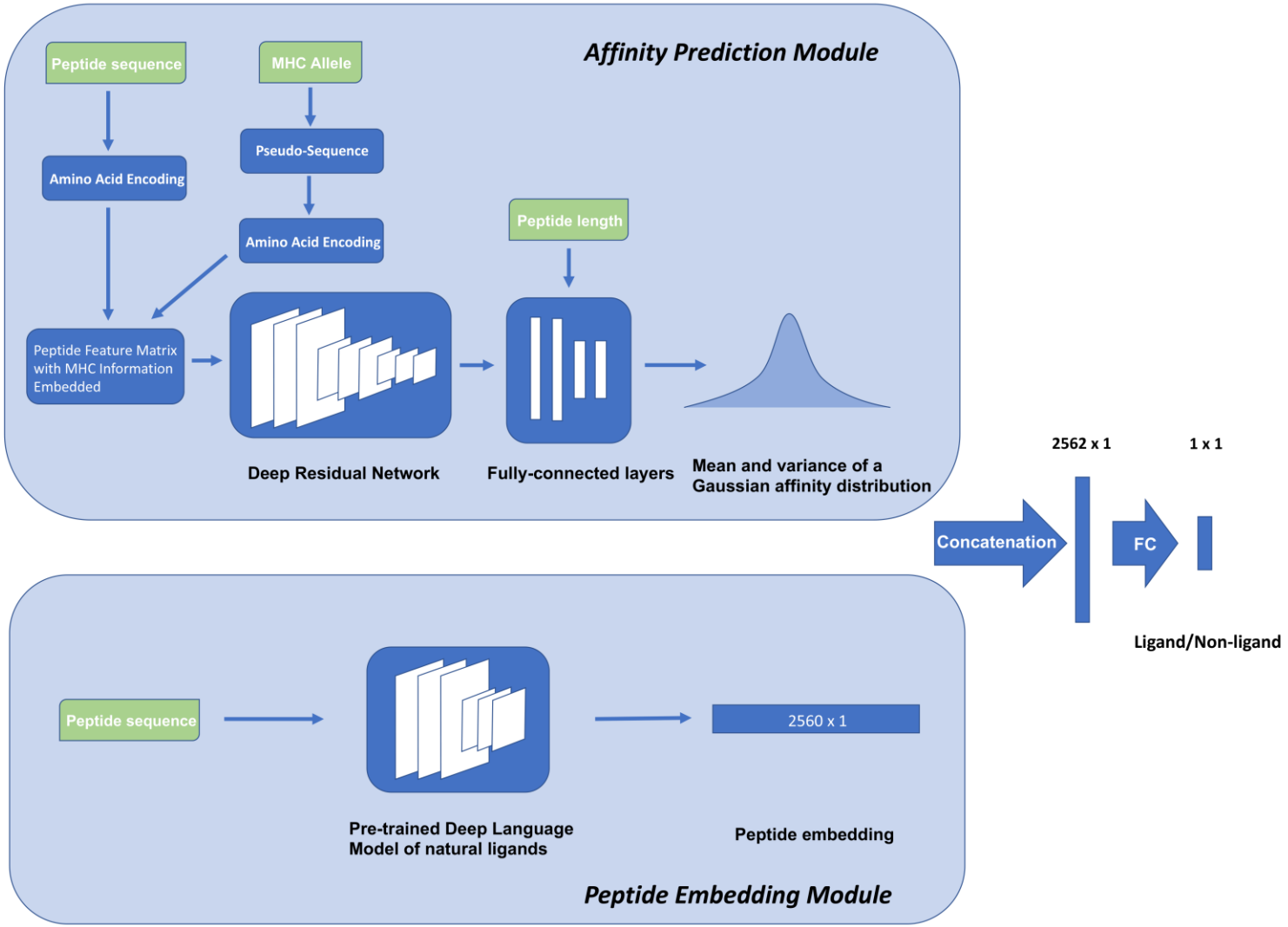


Most existing methods for peptide display focus on modeling MHC binding affinity

- Immune Epitope Database (IEDB) contains a large collection of binding affinity datasets curated from literature
- However, models trained on affinity data are not able to consider other factors in MHC ligand selection



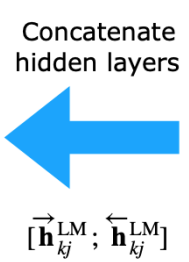
DeepLigand predicts MHC class I peptide presentation (552,252 positive examples, 2.5M negative, 192 MHC alleles)

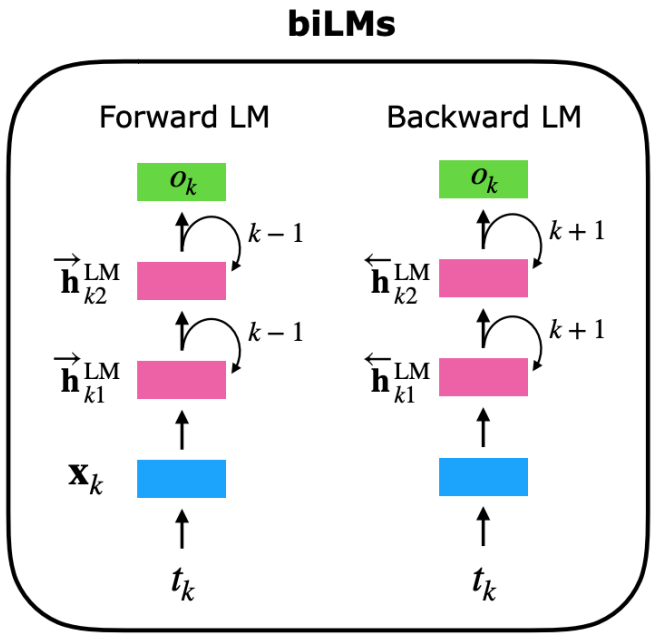


ELMo for learning contextualized word embedding

ELMo represents a word t_k as a linear combination of corresponding hidden layers (inc. its embedding)

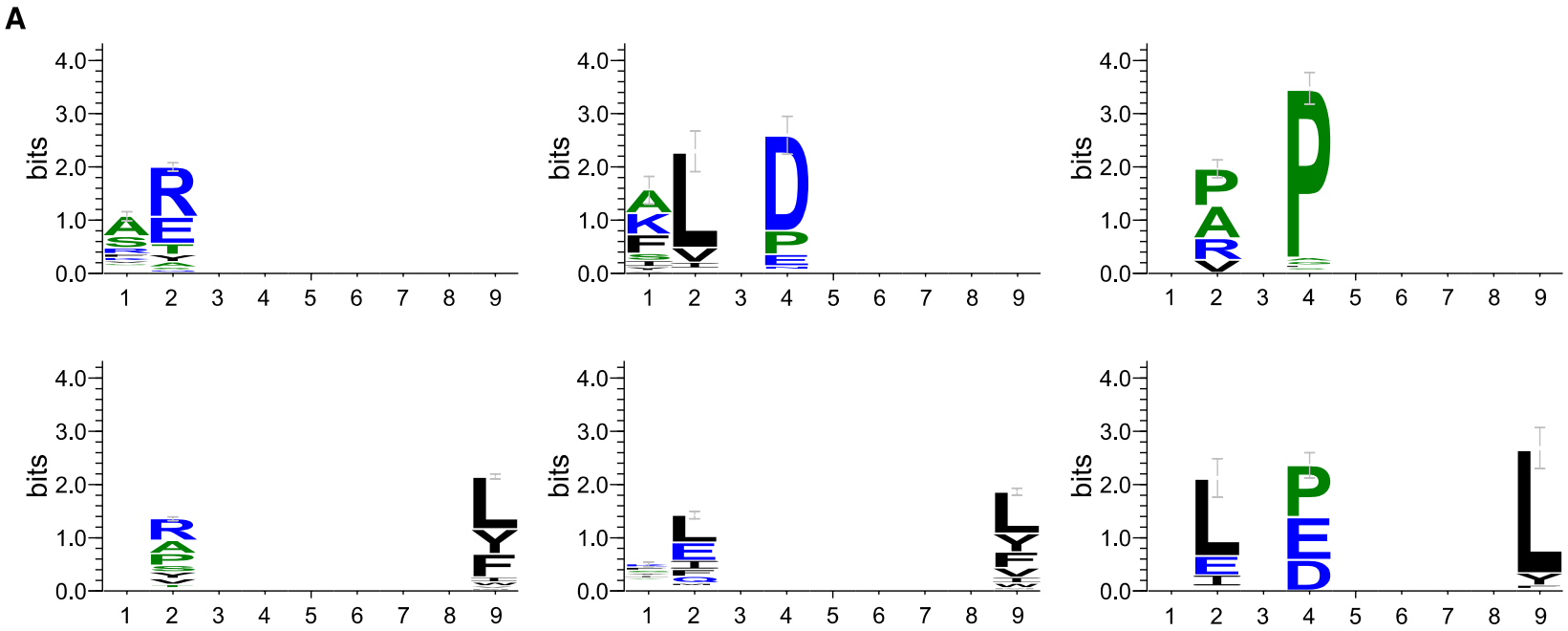
$$\mathbf{ELMo}_k^{\text{task}} = \gamma^{\text{task}} \times \sum \left\{ \begin{array}{l} s_2^{\text{task}} \times \mathbf{h}_{k2}^{\text{LM}} \\ s_1^{\text{task}} \times \mathbf{h}_{k1}^{\text{LM}} \\ s_0^{\text{task}} \times \mathbf{h}_{k0}^{\text{LM}} \quad ([\mathbf{x}_k; \mathbf{x}_k]) \end{array} \right.$$



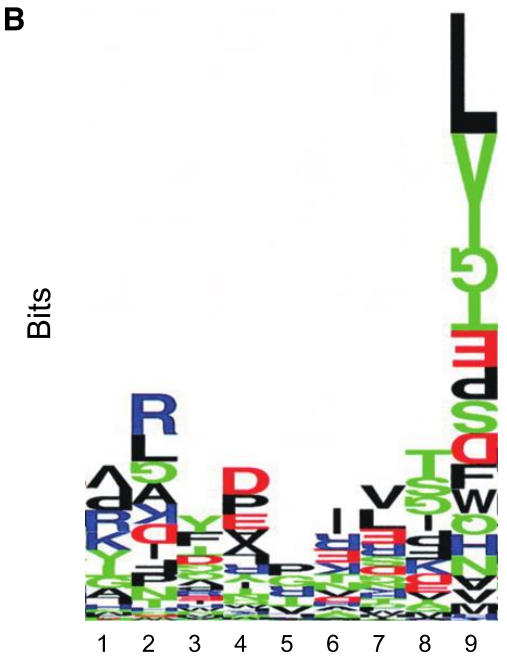


https://arxiv.org/abs/1802.05365

Class I learned language model is consistent with the known proteasome cleavage motif

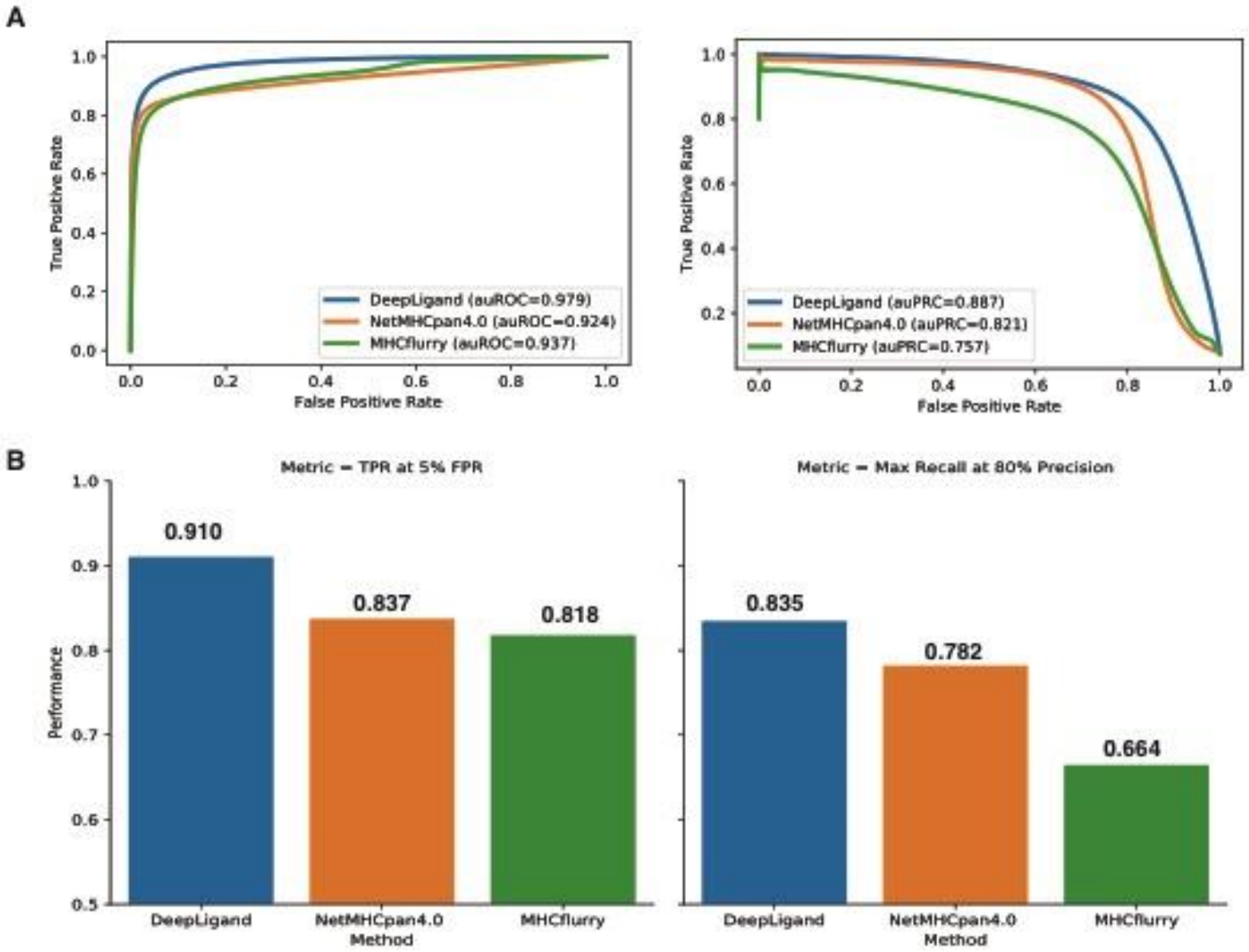


Learned Language Model



Proteasome motif

DeepLigand outperforms existing methods (Class I)

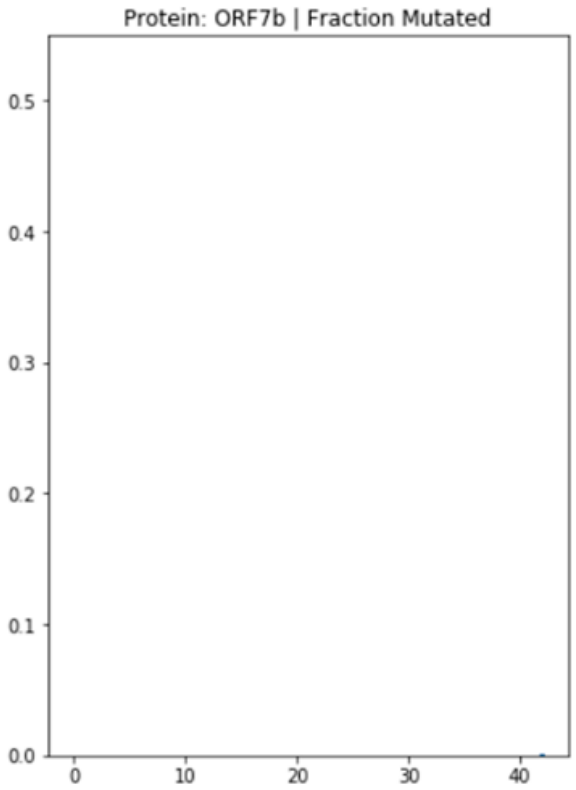


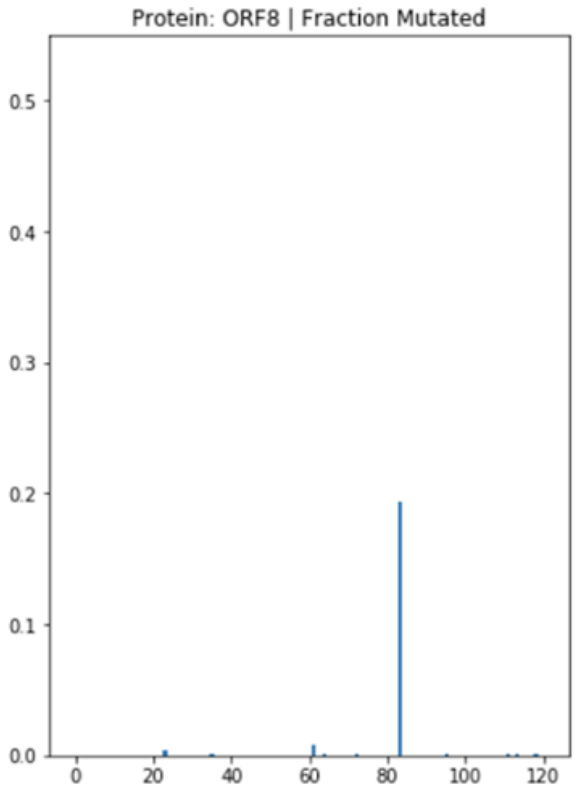
Overview of COVID-19 and MHC Allele Data

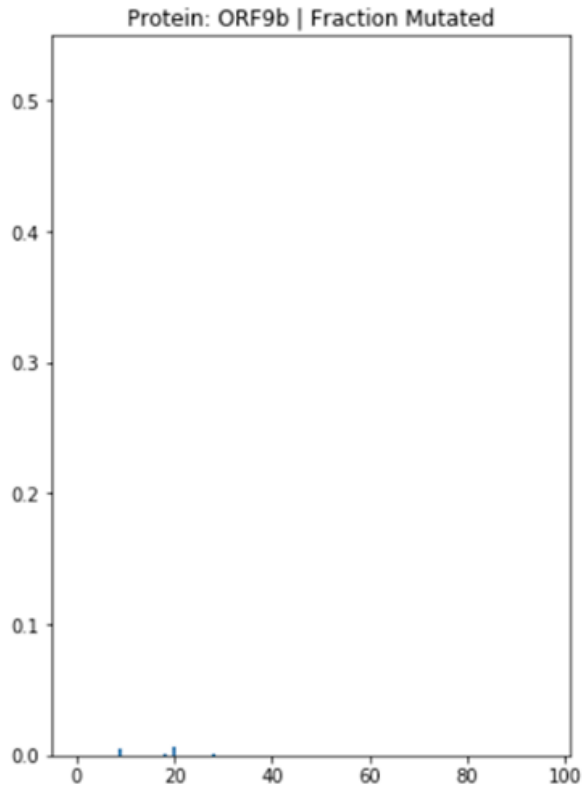
- **14 proteins** in SARS-CoV-2 proteome (length ~10k amino acids)
- Create potential epitopes using **sliding windows of size 8-11** (inclusive) across entire SARS-CoV-2 proteome
- Considered **102 MHC-I HLAs**: 42 HLA-A, 50 HLA-B, 10 HLA-C
- Considered **72 MHC-II HLAs**: 36 HLA-DR, 27 HLA-DQ, 9 HLA-DP
- Predict peptide binding affinity for each (peptide, MHC allele) pair using computational models:
  - DeepLigand (Zeng and Gifford, 2019)
  - PUFFIN (Zeng and Gifford, 2019)
  - NetMHC (Jurtz et al., 2017; Jensen et al., 2018)
  - MHCflurry (O'Donnell et al., 2018)

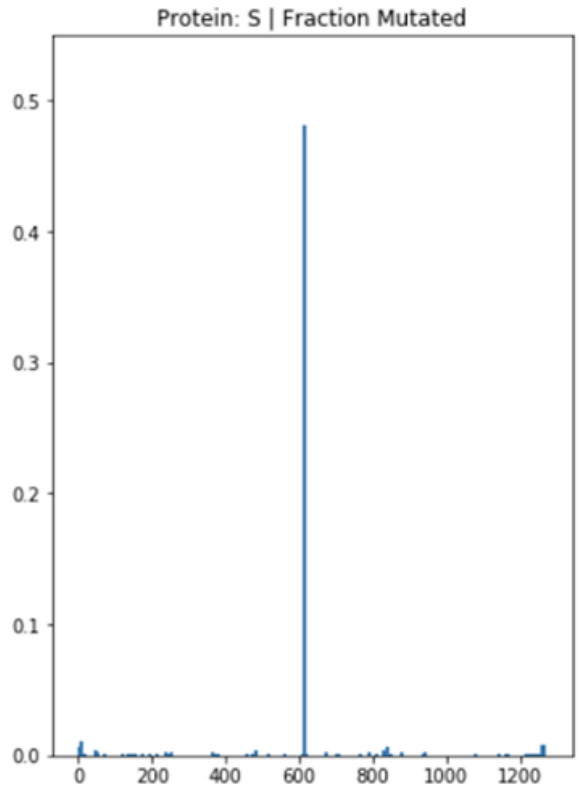
SARS-CoV-2 conservation across proteome

2,847 proteins preprocessed, translated and aligned using the NextStrain processing pipeline. Sequences downloaded from GISAID on April 3rd. Uses one of the first genome sequences 'Wuhan-Hu-1/2019' as reference. X-axis, residue position. Y-axis, fraction changed.

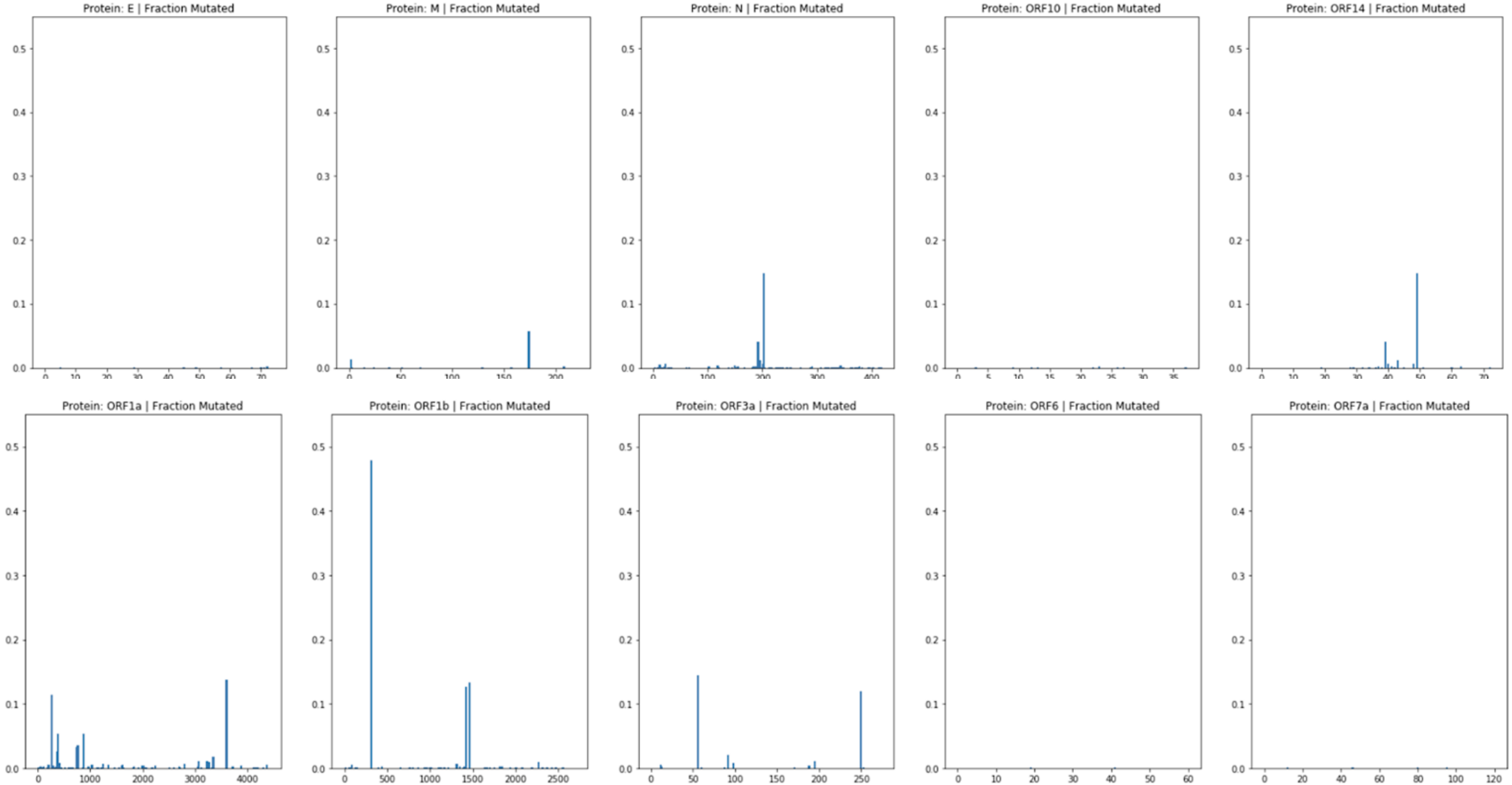






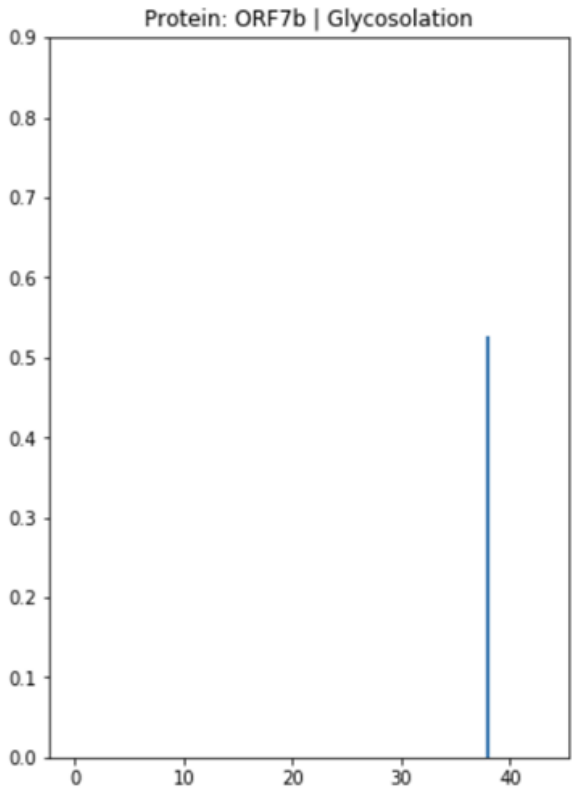


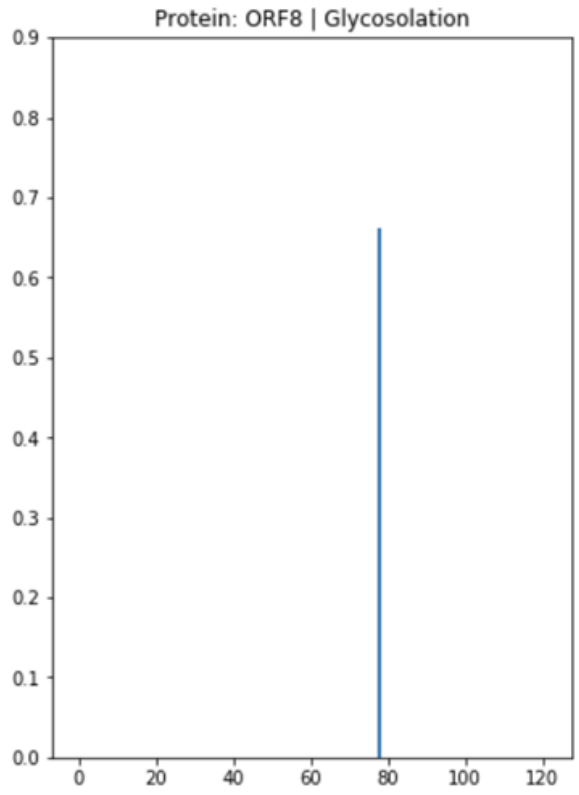
SARS-CoV-2 conservation across proteome

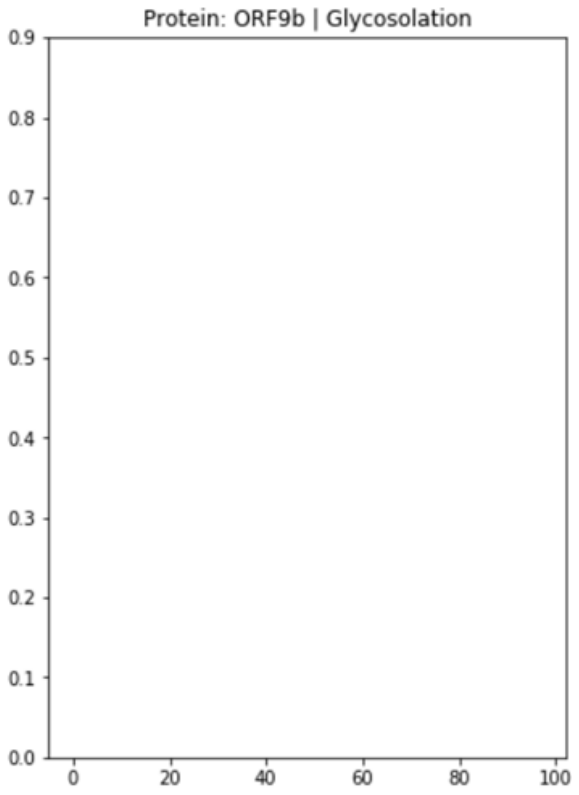


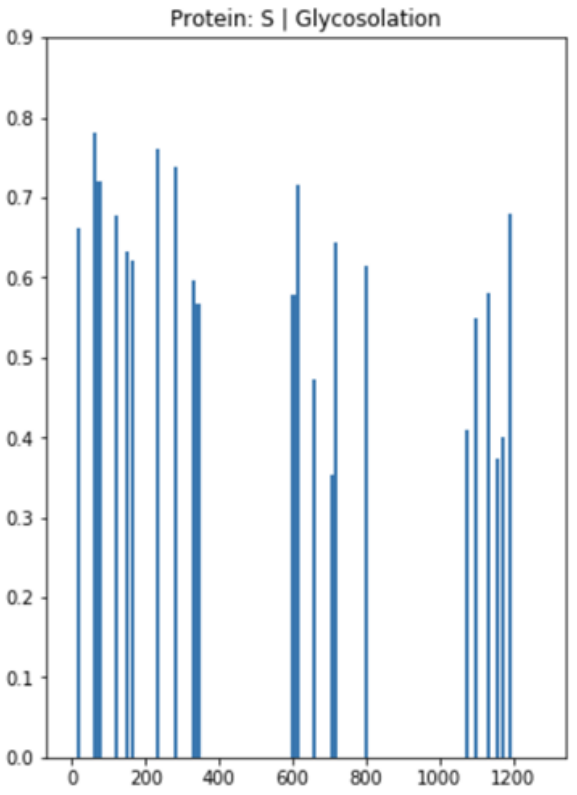
SARS-CoV-2 predicted glycosylation

Mass-Spec data from Zhang et al. (https://www.biorxiv.org/content/10.1101/2020.03.28.013276v1) on the Spike protein glycosolation sites shows that the predictions made here are 100% accurate with no false positives for any positive probability.

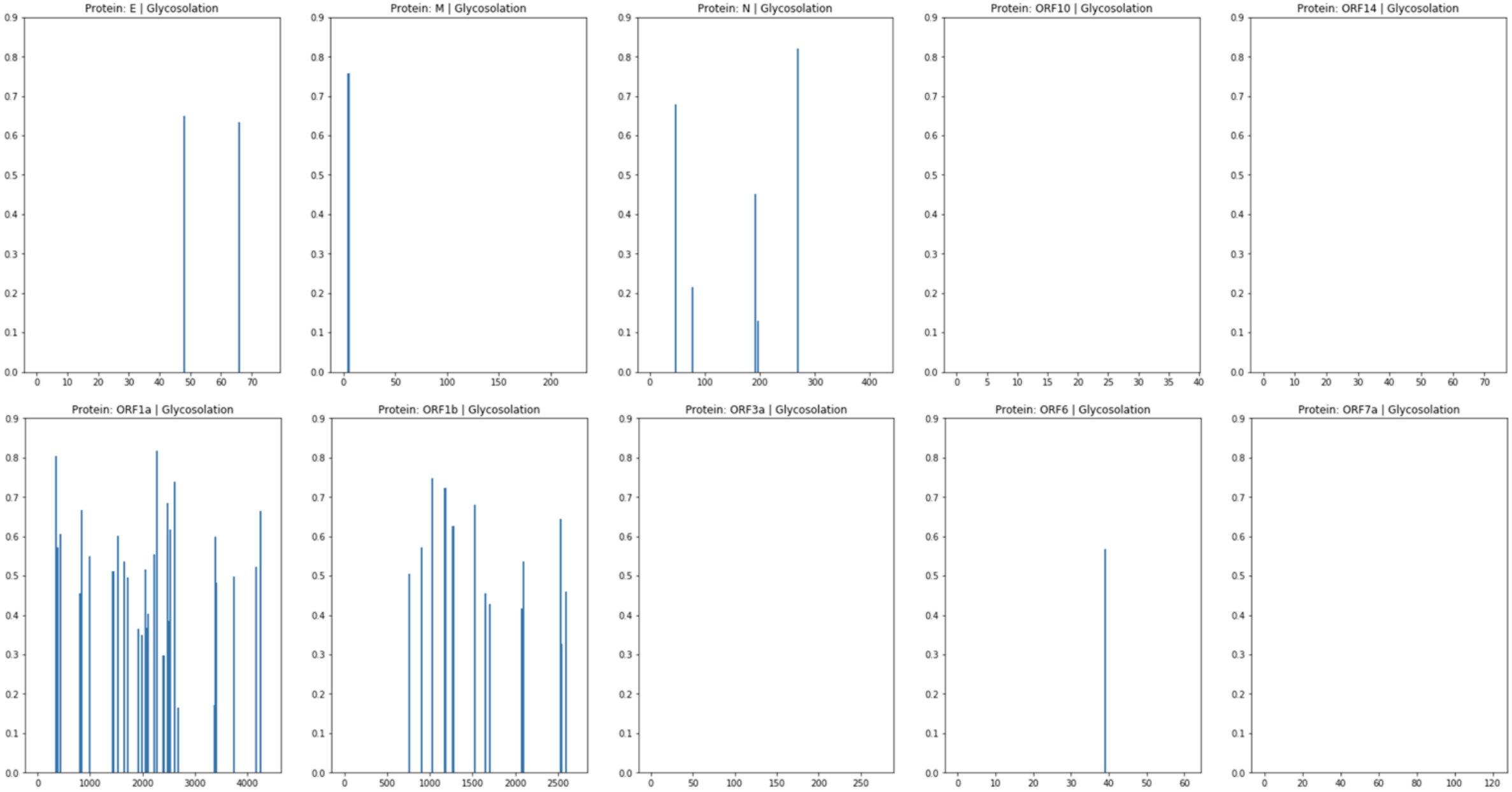








SARS-CoV-2 predicted glycosylation



OptiVax Population Coverage Optimization Pipeline

Allele-specific binding prediction for candidate peptide pool

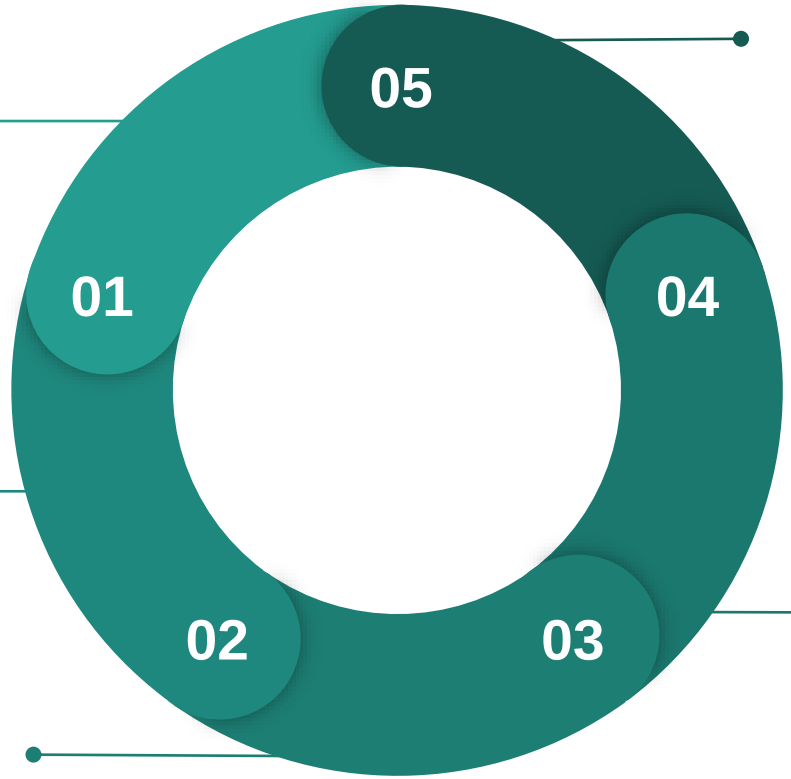
Deep learning models that predict binding affinity/likelihood for each peptide over **102** selected MHC I alleles in 3 loci (HLA-A/B/C) and **72** MHC II alleles in 3 loci (HLA-DR/DQ/DP)

Modelling other MHC-binding related characteristics of the peptides

Predict protein expression level, glycosylation probability, structural and sequence mutation entropies, etc.

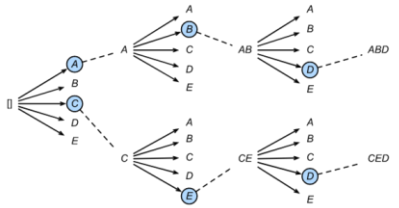
Post-process binding predictions and combine with related characteristics

Truncating predicted binding metrics to focus on high-affinity candidates, factor in other related characteristics to produce final allele-level binding estimation for downstream optimizations.



Population coverage optimization

Through iterative optimization algorithm (greedy or beam search) we select a minimal set of peptides that achieve population level binding above a given cutoff (99.5%).



Population level binding estimation

Our population coverage probabilistic model considers allele frequencies in a given population, and models the overall probability of peptide presentation across different diploid locus combinations, given a set of peptides and their allele-level binding estimations.

$$F_k(A_i, A_j) = G_k(A_i)G_k(A_j)$$

$$e_k(A_i) = 1 - \prod_{n=1}^{N}(1 - e_k^n(A_i))$$

$$B_k(A_i, A_j) = \begin{cases} 1 - (1 - e_k(A_i))(1 - e_k(A_j)), & \text{if } i \neq j \\ e_k(A_i), & \text{if } i = j \end{cases}$$

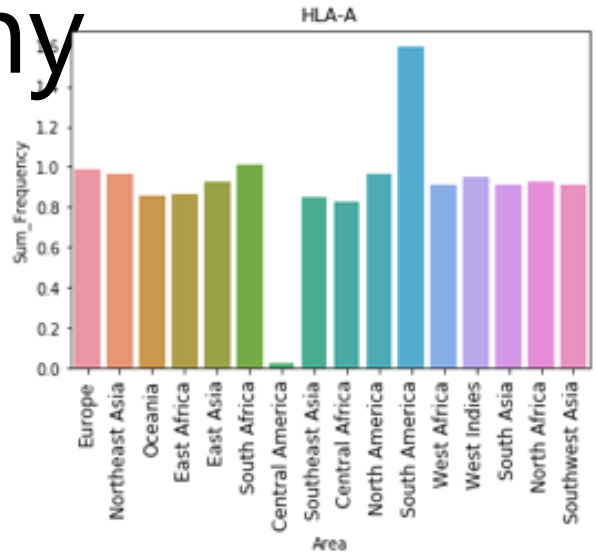
$$F_k(P) = \sum_{i=1}^{M_k}\sum_{i=1}^{M_k} F_k(A_i, A_j)B_k(A_i, A_j)$$

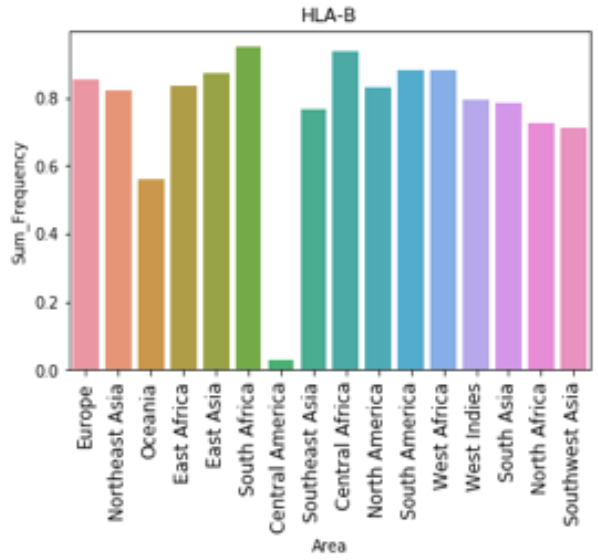
$$P(P) = 1 - \prod_{k=1}^{K}(1 - F_k(P))$$

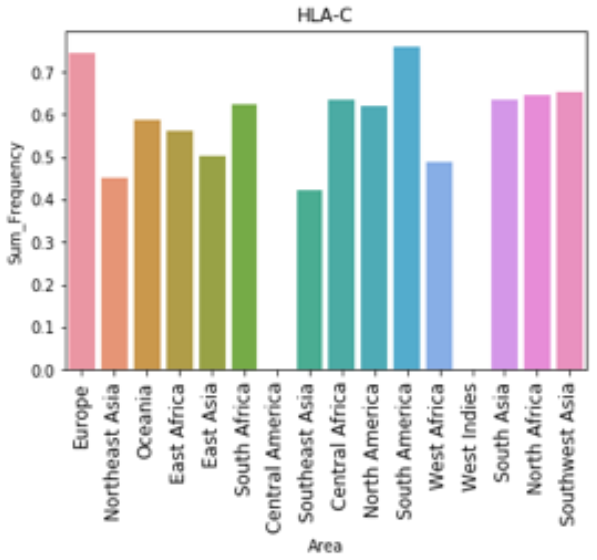
Coverage of 102 MHC Class I alleles by geography

dbMHC
(102 alleles covered, 16 areas, 86 countries)

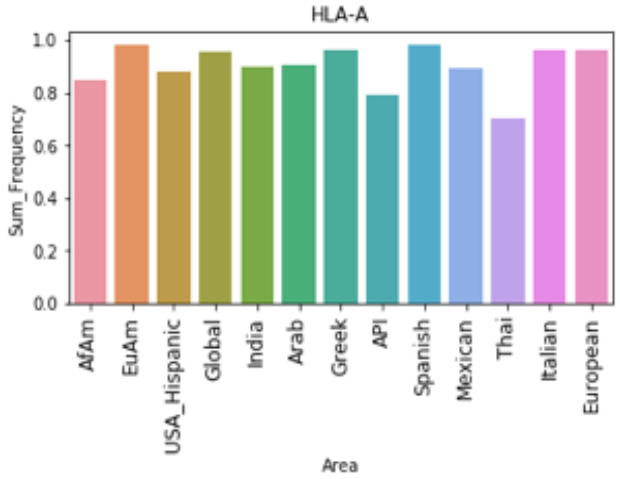
Used for present vaccine optimization

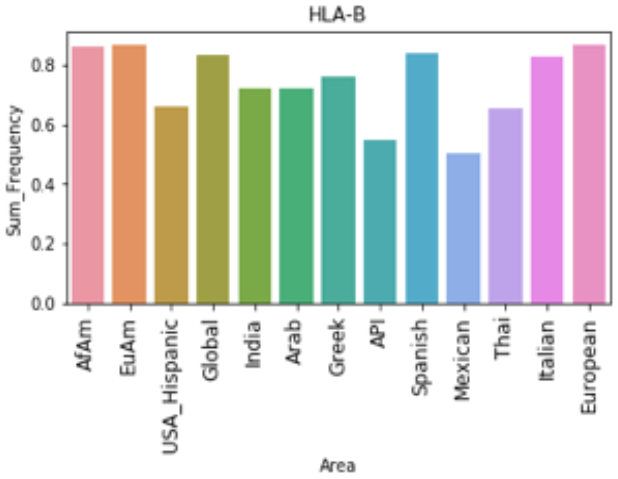


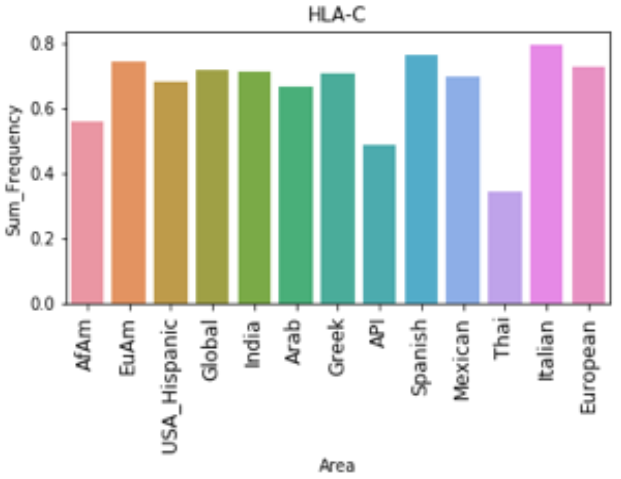




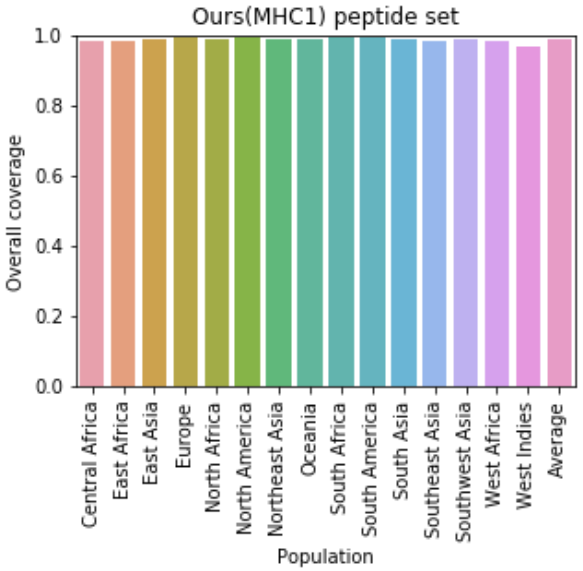
17th IHIW NGS HLA Data
(83 alleles covered, 12 groups)

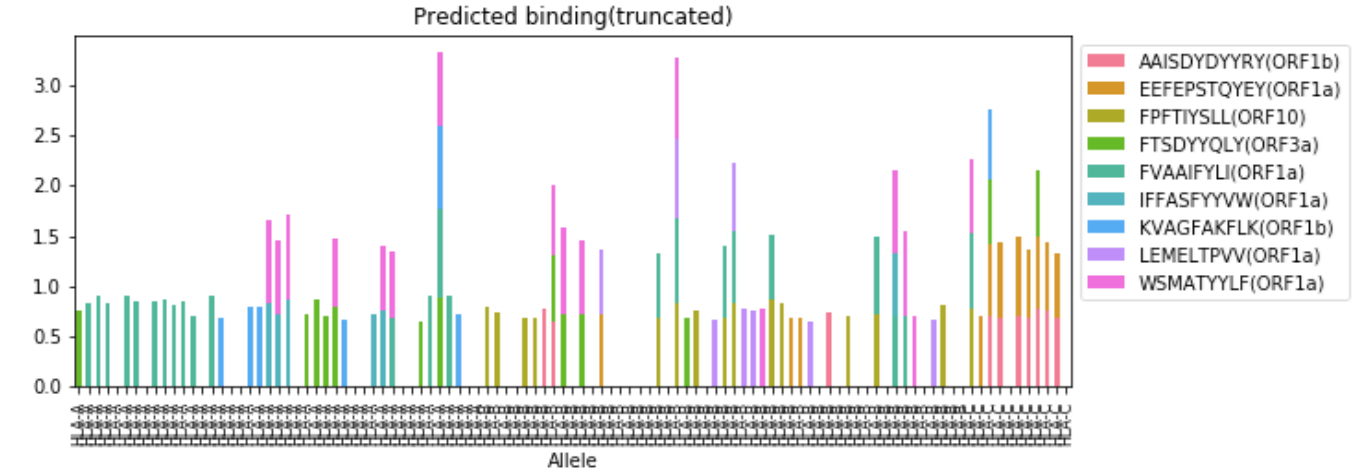


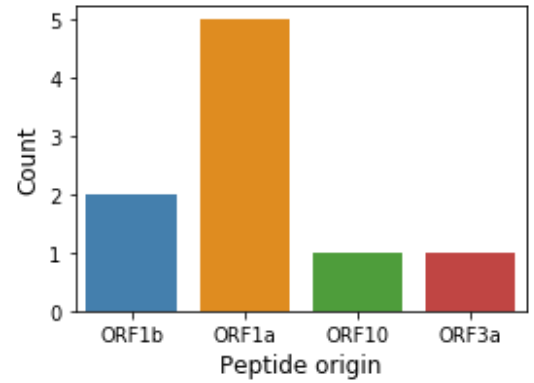




OptiVax Class I MHC results for SARS-CoV-2

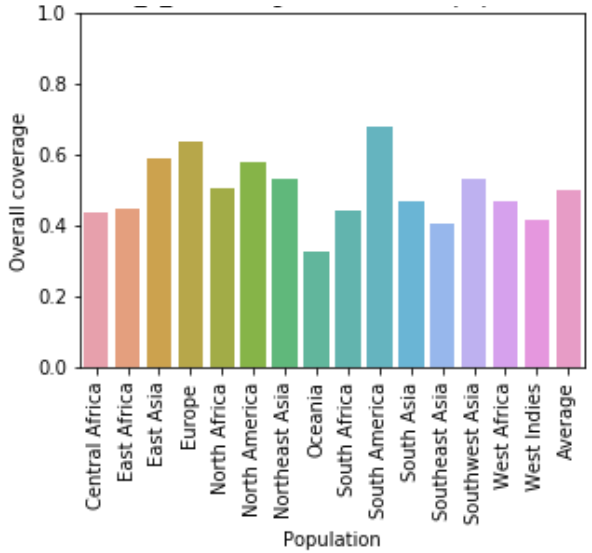


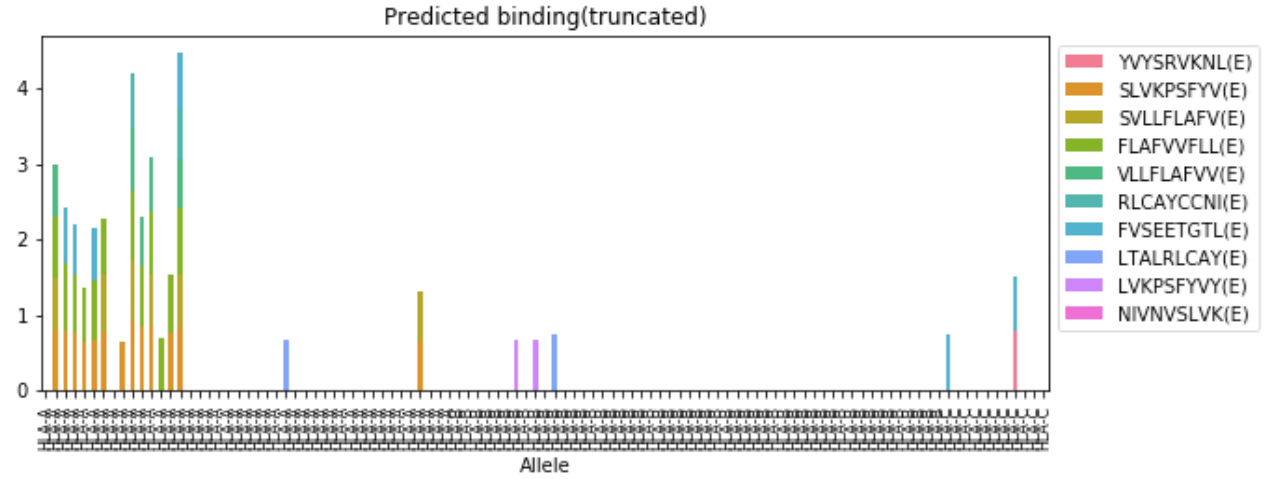




sequence	protein	entropy_sum	epi_start_pos	epi_len	glyco_probs	crosses_cleavage	In_SARS_Cov1
AAISDYDYYRY	ORF1b	0.051702	438	11	0.0	0	True
EEFEPSTQYEY	ORF1a	0.020229	938	11	0.0	0	False
FPFTIYSLL	ORF10	0.022220	8	9	0.0	0	False
FTSDYYQLY	ORF3a	0.004446	206	9	0.0	0	False
FVAAIFYLI	ORF1a	0.000000	2781	9	0.0	0	False
IFFASFYYVW	ORF1a	0.006367	2384	10	0.0	0	False
KVAGFAKFLK	ORF1b	0.009235	31	10	0.0	0	True
LEMELTPVV	ORF1a	0.002395	1011	9	0.0	0	False
WSMATYYLF	ORF1a	0.004789	899	9	0.0	0	False

OptiVax optimization results outperform baselines in literature



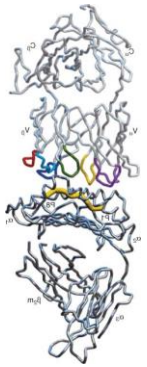


sequence	protein	entropy_sum	epi_start_pos	epi_len	glyco_probs	crosses_cleavage	In_SARS_Cov1
YVYSRVKNL	E	0.004789	56	9	0.0	0	True
SLVKPSFYV	E	0.006840	49	9	0.0	0	False
SVLLFLAFV	E	0.007184	15	9	0.0	0	True
FLAFVVFLL	E	0.002395	19	9	0.0	0	True
VLLFLAFVV	E	0.007184	16	9	0.0	0	True
RLCAYCCNI	E	0.004789	37	9	0.0	0	True
FVSEETGTL	E	0.002395	3	9	0.0	0	True
LTALRLCAY	E	0.004789	33	9	0.0	0	True
LVKPSFYVY	E	0.006840	50	9	0.0	0	False
NIVNVSLVK	E	0.004789	44	9	1.0	0	True

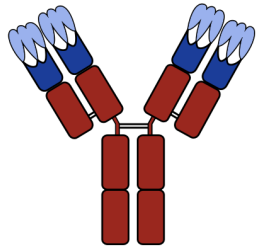
High probability of glycosylation

Today’s deep learning methods

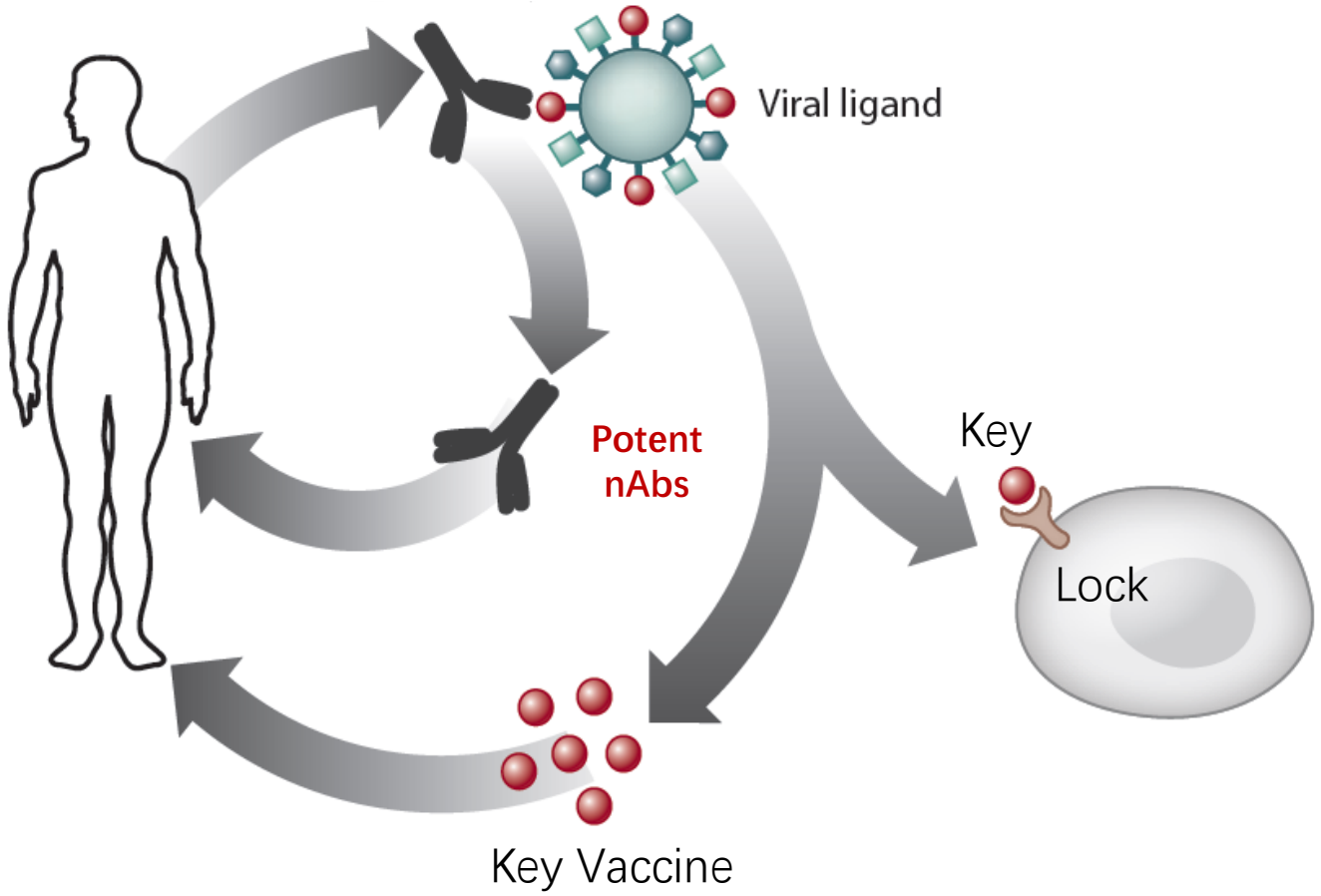
- Vaccine design

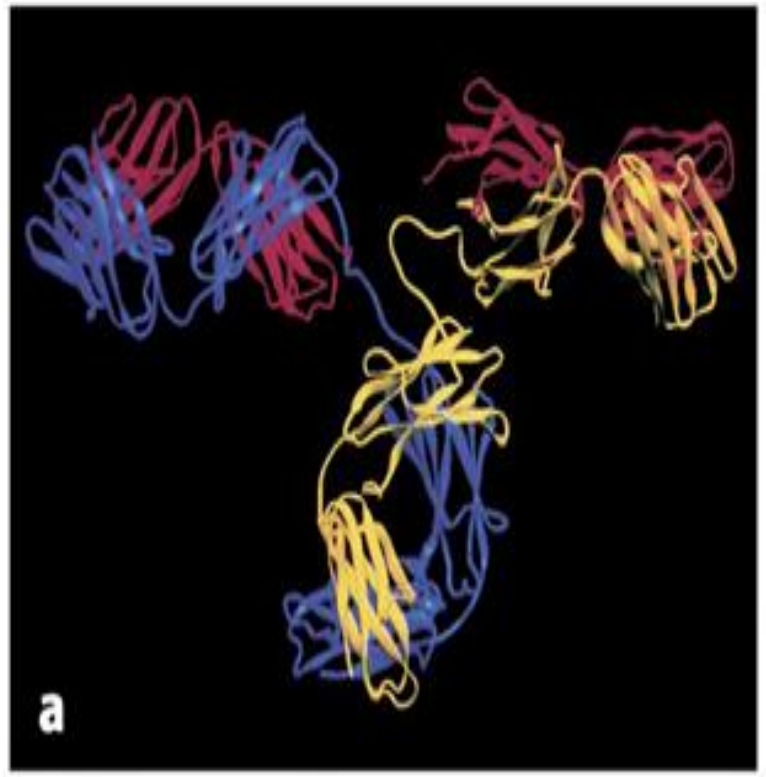


- Antibody discovery and improvement



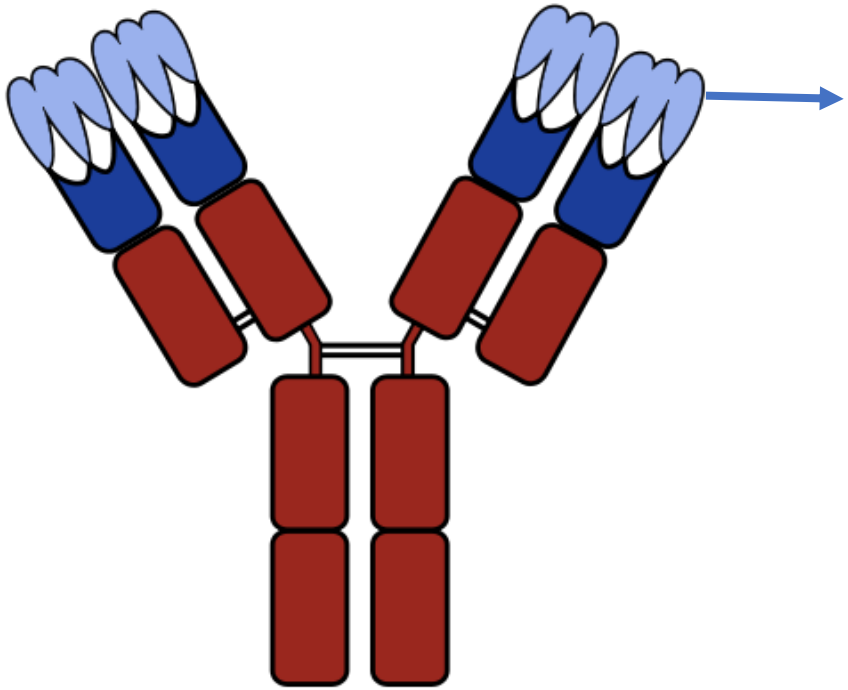
Antibody response in viral infection



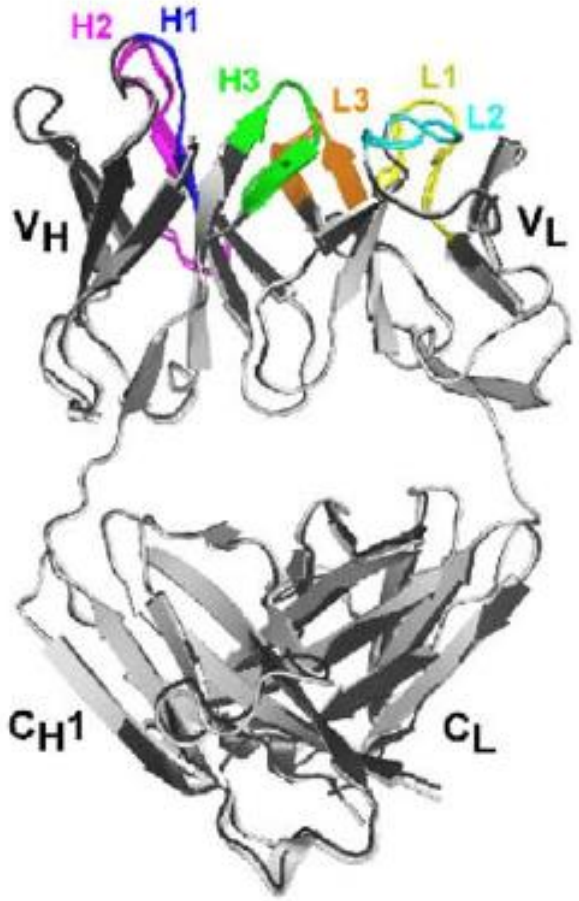


~20% of plasma protein

Complementarity-determining regions (CDRs) largely determine target affinity



Six CDRs in total for each Fab



https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101210/

Design flow

Naïve library 10^{10} CDRH3 sequences

↓

Multiplexed Affinity Assay

↓ Training Data

ML Design Methods

↓

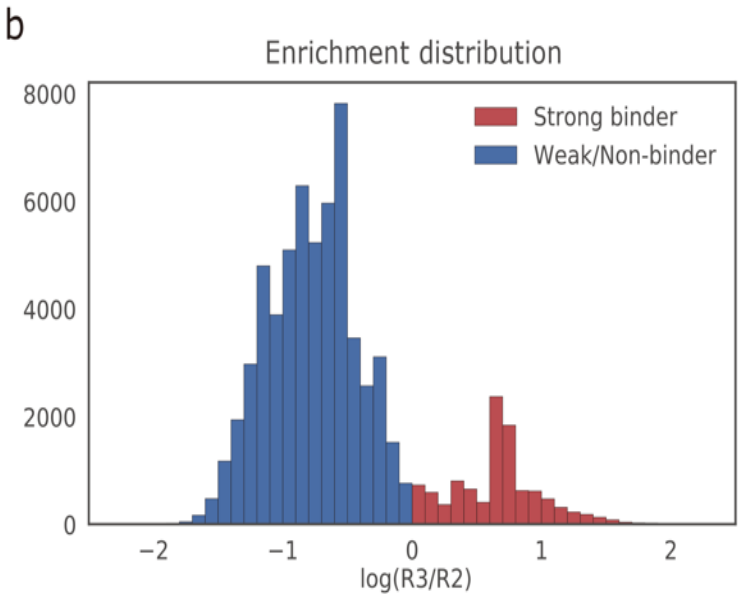
Naïve library → ML library fabrication

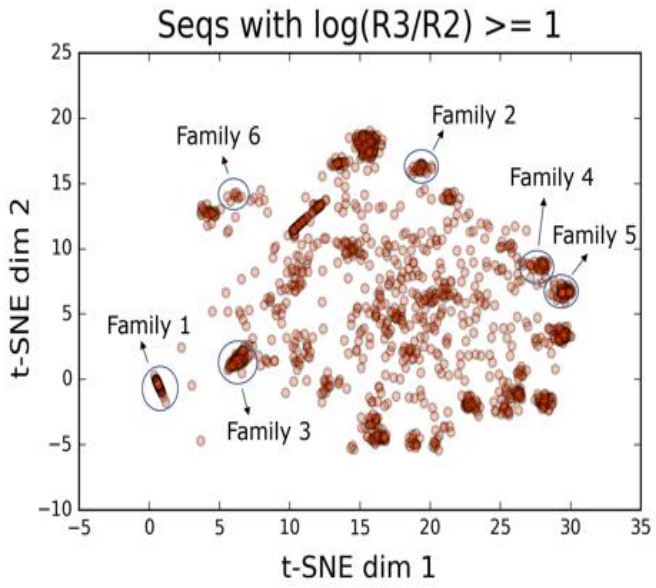
↓

Multiplexed Affinity Assay

↓

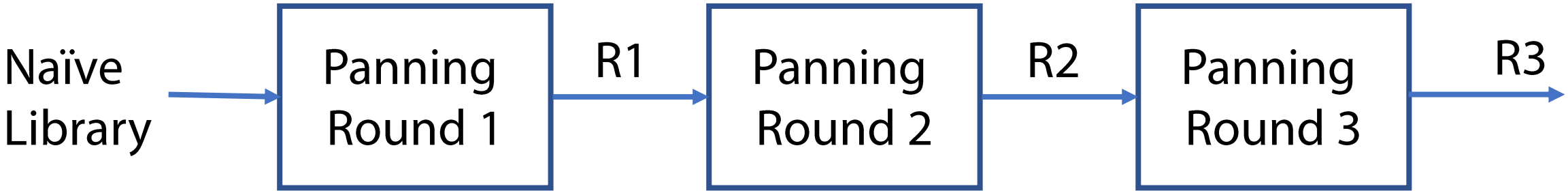
Full length antibody characterization





Training Data

Enrichment is defined by the output of three panning rounds



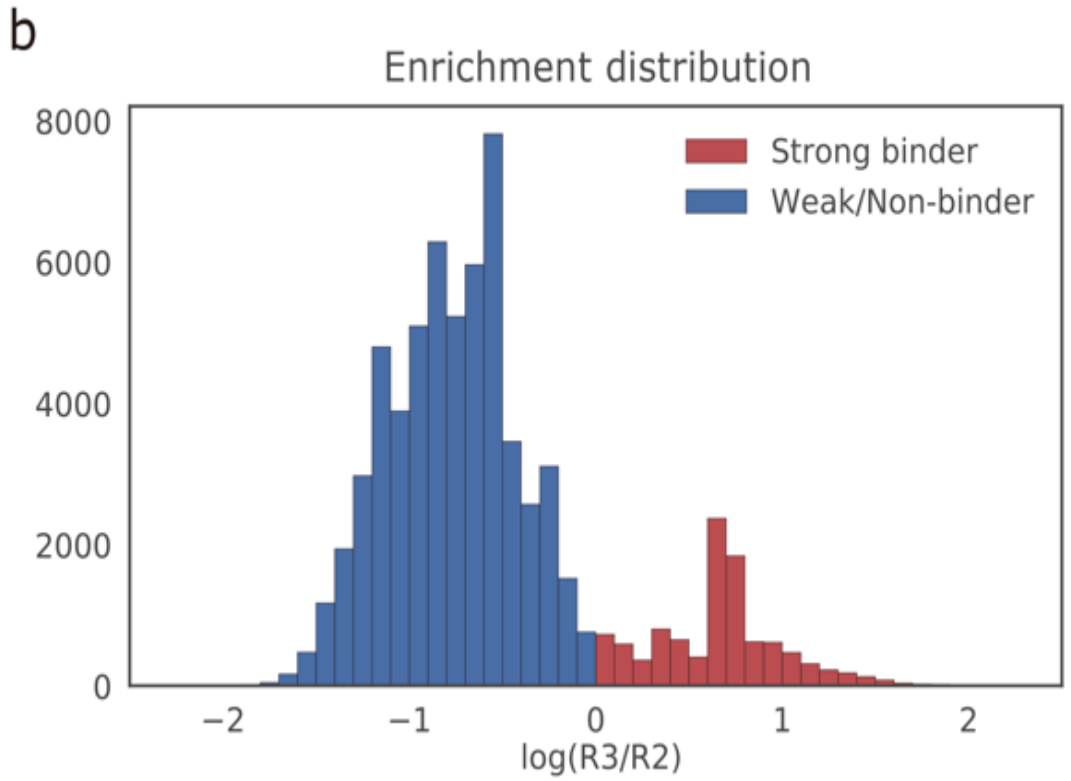
$$\text{Enrichment}_{R3/R1} = \text{Log}_{10}\,(\text{Frequency R3} / \text{Frequency R1})$$

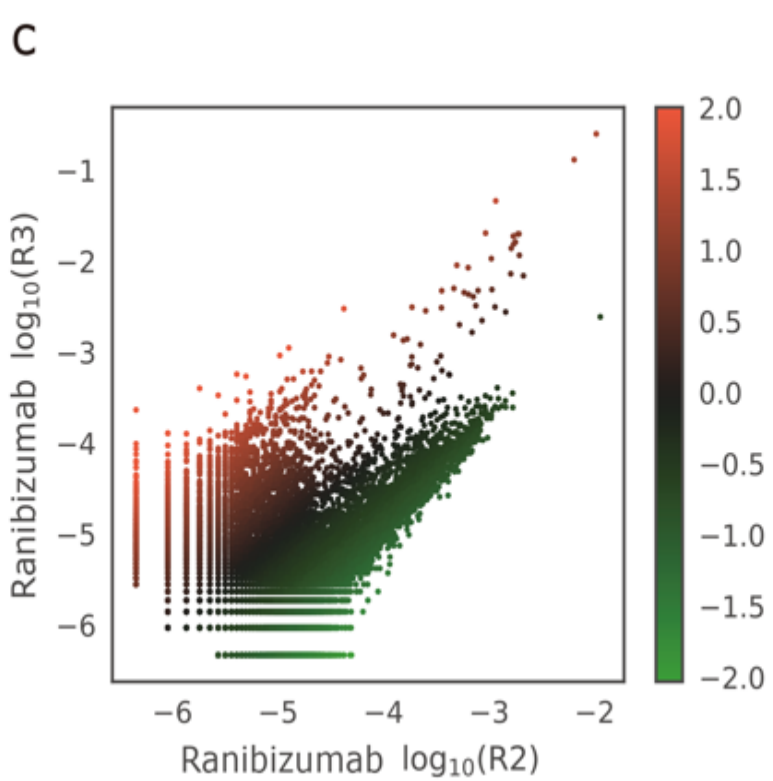
$$\text{Enrichment}_{R3/R2} = \text{Log}_{10}\,(\text{Frequency R3} / \text{Frequency R2})$$

Train on CDR-H3 sequence and enrichment

Sequence	Log(R3/R2)	Enbrel_a	Avastin_a	Herceptin_a	
ADGAFDAYMDY	-0.9561	-0.5989	-0.9730	-1.2414	
ADGYRVYYYAMDY	1.2253	1.4830	0.9872	1.1175	←
ADRRPPLIFFDY	0.8519	0.8458	1.9072	1.9057	
ADWLSLLYRFDY	-0.4779	-0.9202	-0.7767	-0.8339	
AEHVAYHPRYSFDY	-0.9474	-0.7291	-0.9730	-0.8649	
AGRYWWLLDY	0.3242	0.3843	1.7872	0.6588	
AGYHQTWPYGLDY	1.0482	0.8792	0.9135	-0.2221	
AKRRRQYVYHPIYFDY	1.6727	1.4852	1.9769	2.0698	
AKYADTYGLDY	0.4839	0.2024	-0.2996	0.9655	
AKYGSYYGFDY	0.5650	0.3526	0.3929	0.5801	
DAYPGWDLWPDYPFDY	0.2757	-0.0151	-0.0842	0.4879	←
DDIHHLLYYFDY	0.9610	1.1010	0.9135	1.5183	
DDQYVGYFYGEGGLDY	-0.2620	0.0897	0.3372	0.1532	
DDVKGHSKQDLRVFDY	0.7702	-0.0341	1.7246	0.1893	
DDVYWIAAFDY	-0.5247	0.8792	-0.8859	-0.4439	←
DDWYGGLERGLIQFDY	0.2621	-0.0544	1.3027	0.3120	

High enrichment suggests high affinity sequences (Ranibizumab, 67769 sequences)





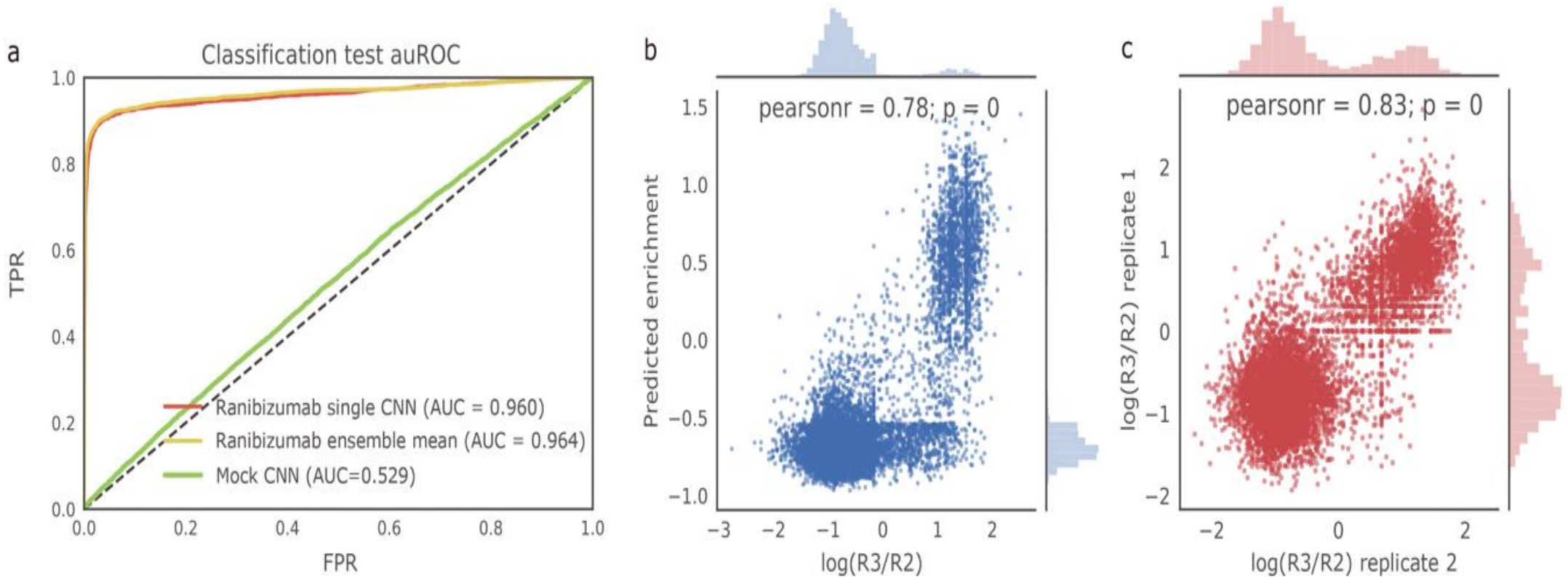
We used six different model architectures

	# of Convolutional layers	# of Convolutional filters	Convolutional filter size	# of Fully connected layer	# of Fully connected neurons	# of parameters in total
2fc	0	0	0	2	32	13954
1conv(32*5)+1fc	1	32	5	1	16	8402
2conv(32*5_64*5)+1fc	2	32,64	5	1	16	18706
1conv(64*5)+1fc	1	64	5	1	16	16754
1conv(32*3)+1fc	1	32	3	1	16	7122
2conv(8*1_64*5)+1fc	2	8, 32	1, 5	1	16	13082

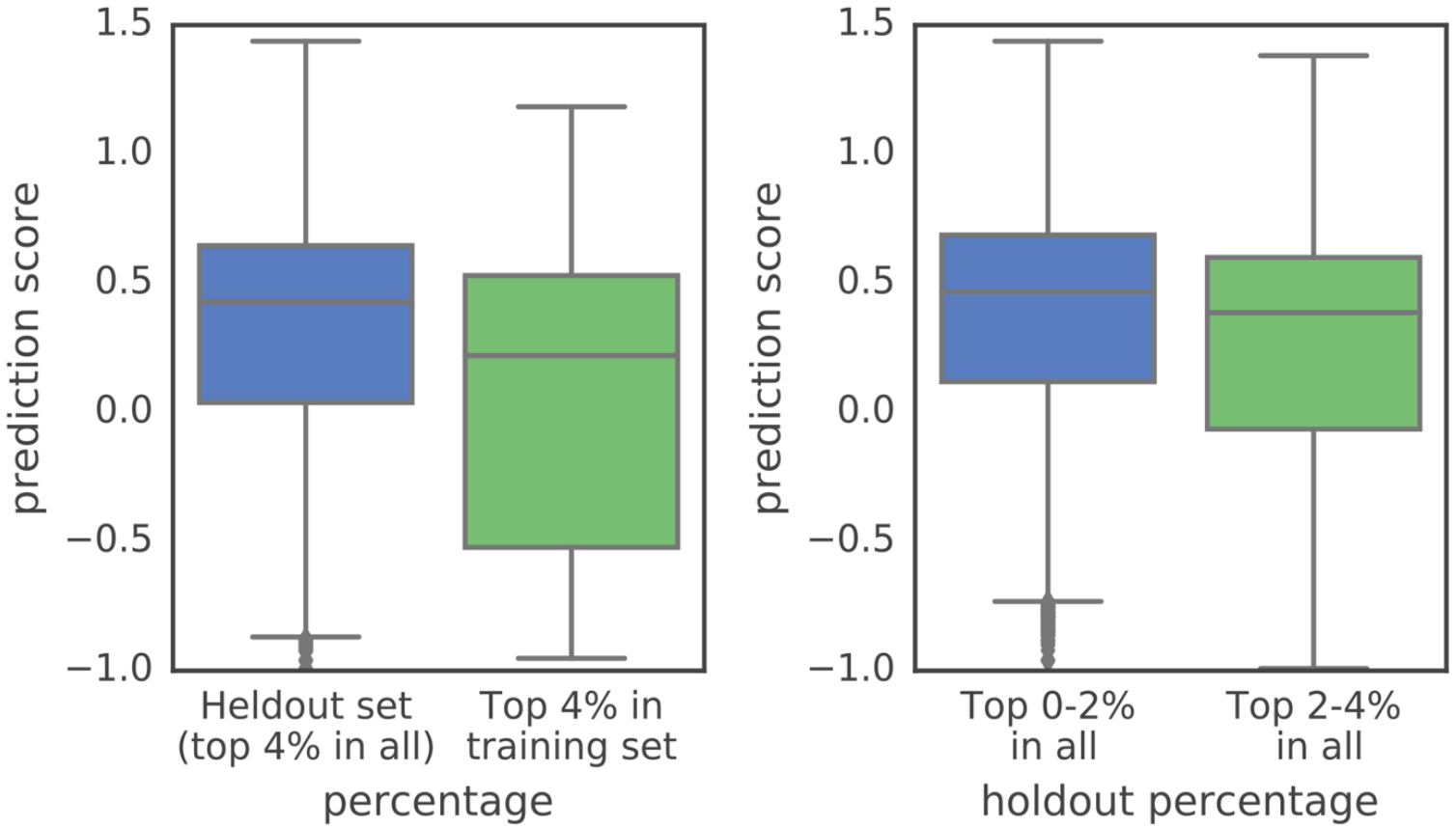
Output layer: Classification – binary cross entropy loss

Regression – mean squared error

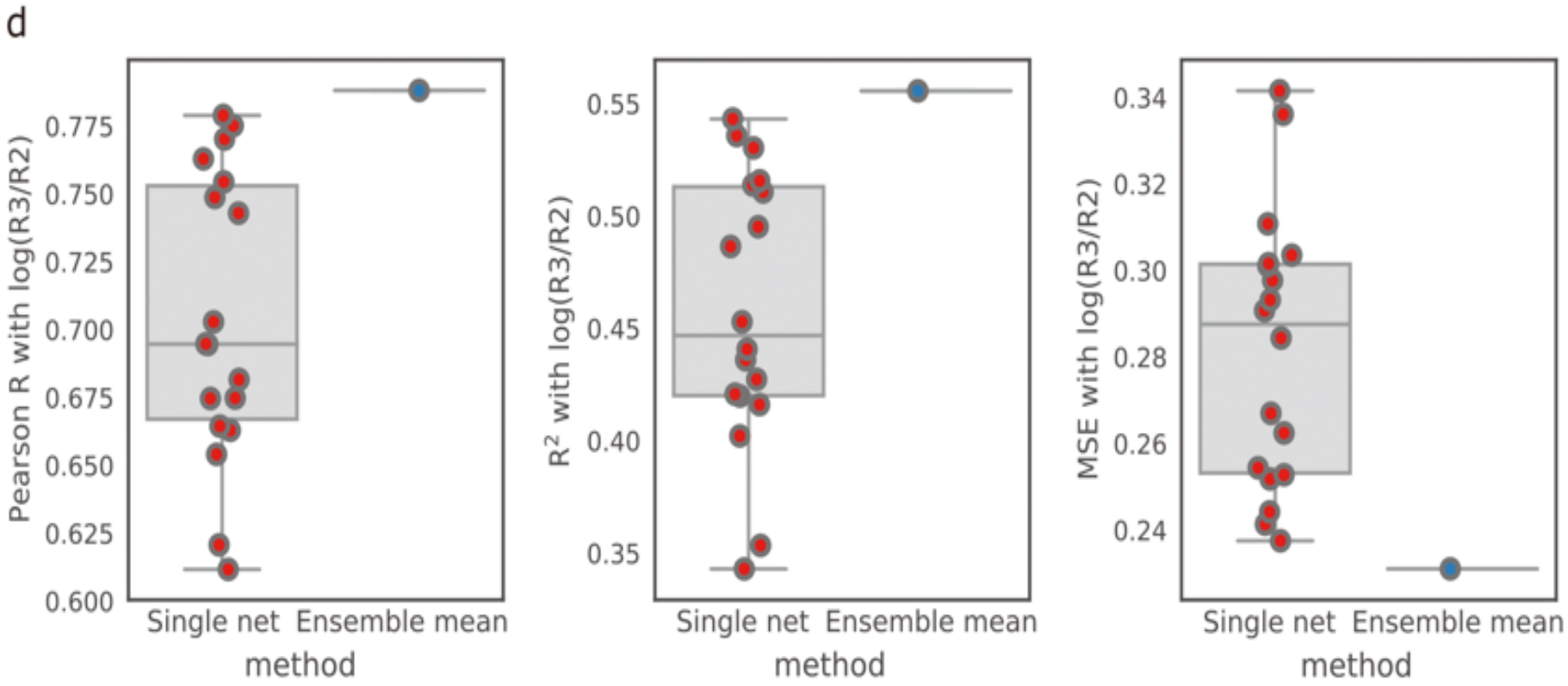
Regression performance is comparable to replicate experiment performance



CNNs produce better scores than they have seen in training for top sequences



An ensemble of 24 networks is more robust than the individual networks



How can we optimize CDRs?

Design flow

Naïve library 10^{10} CDRH3 sequences

↓

Multiplexed Affinity Assay

↓ Training Data

ML Design Methods

↓

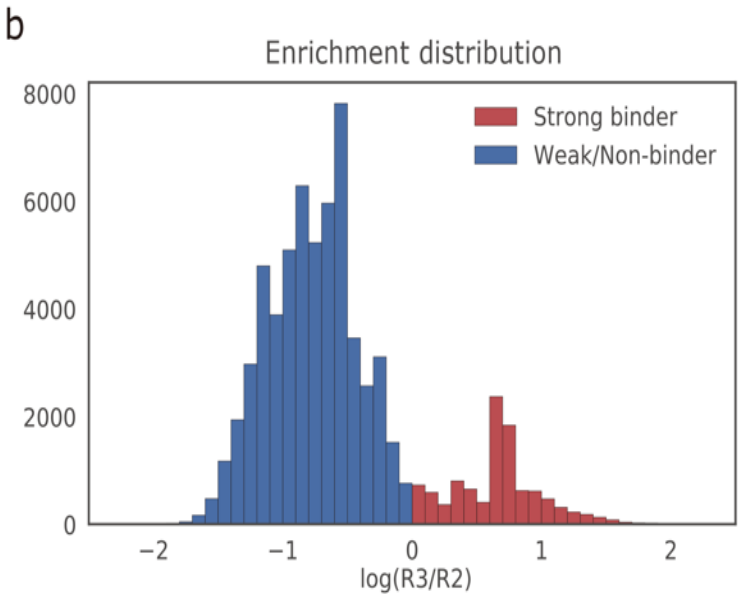
Naïve library → ML library fabrication

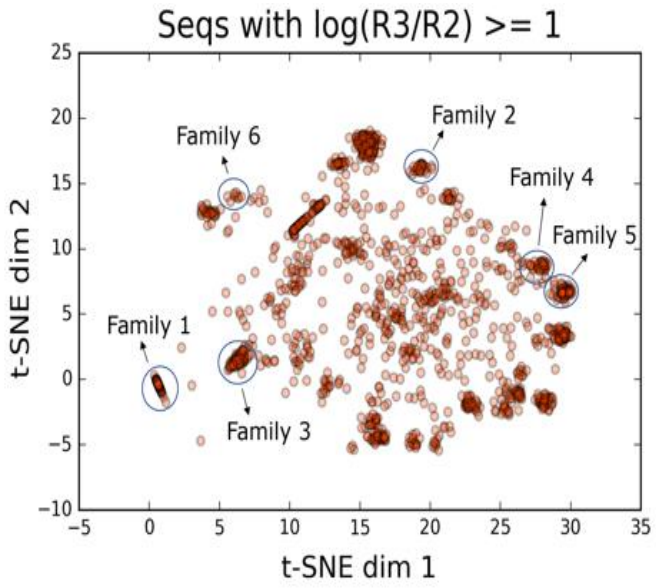
↓

Multiplexed Affinity Assay

↓

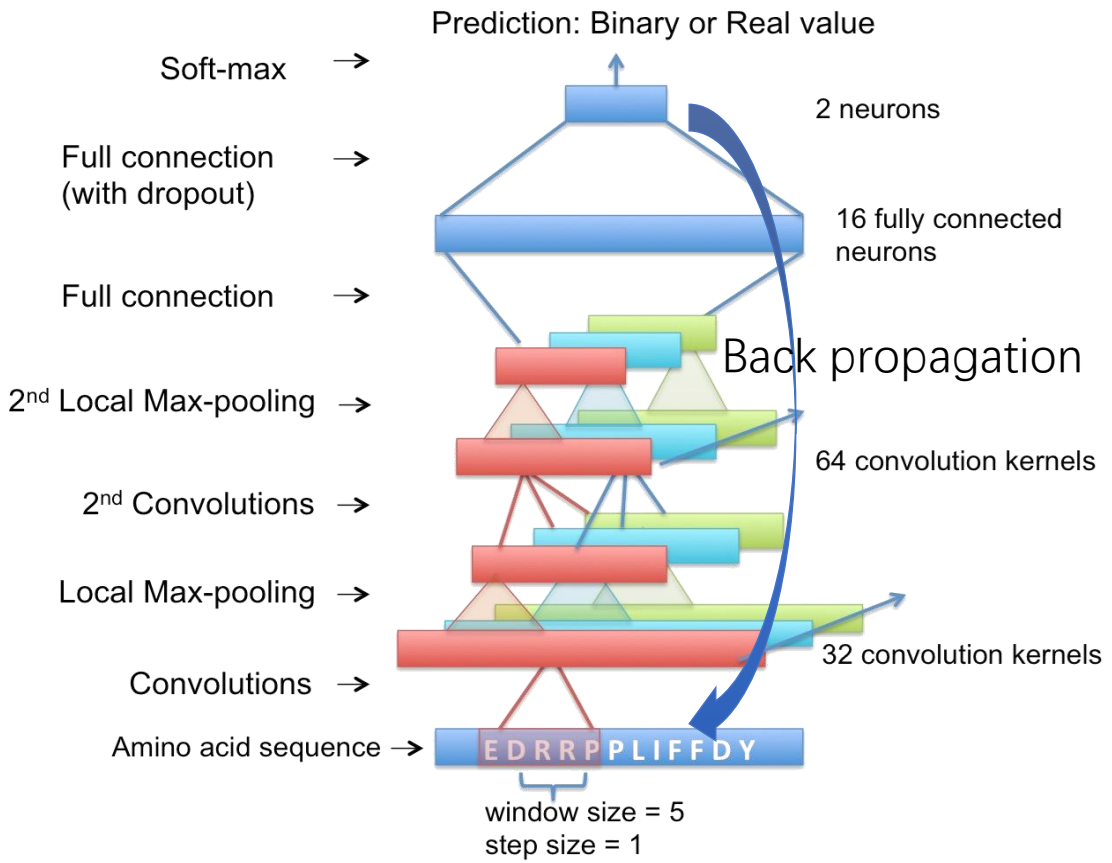
Full length antibody characterization



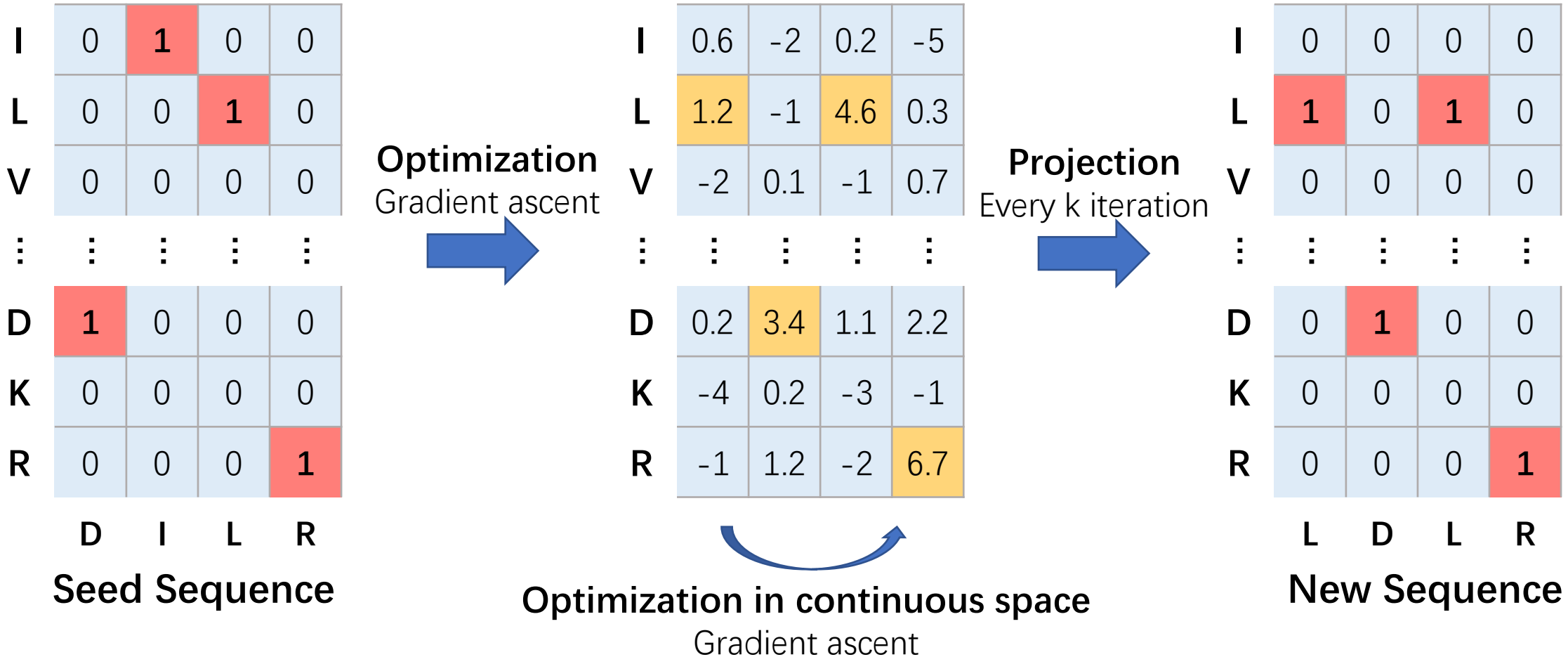


Training Data

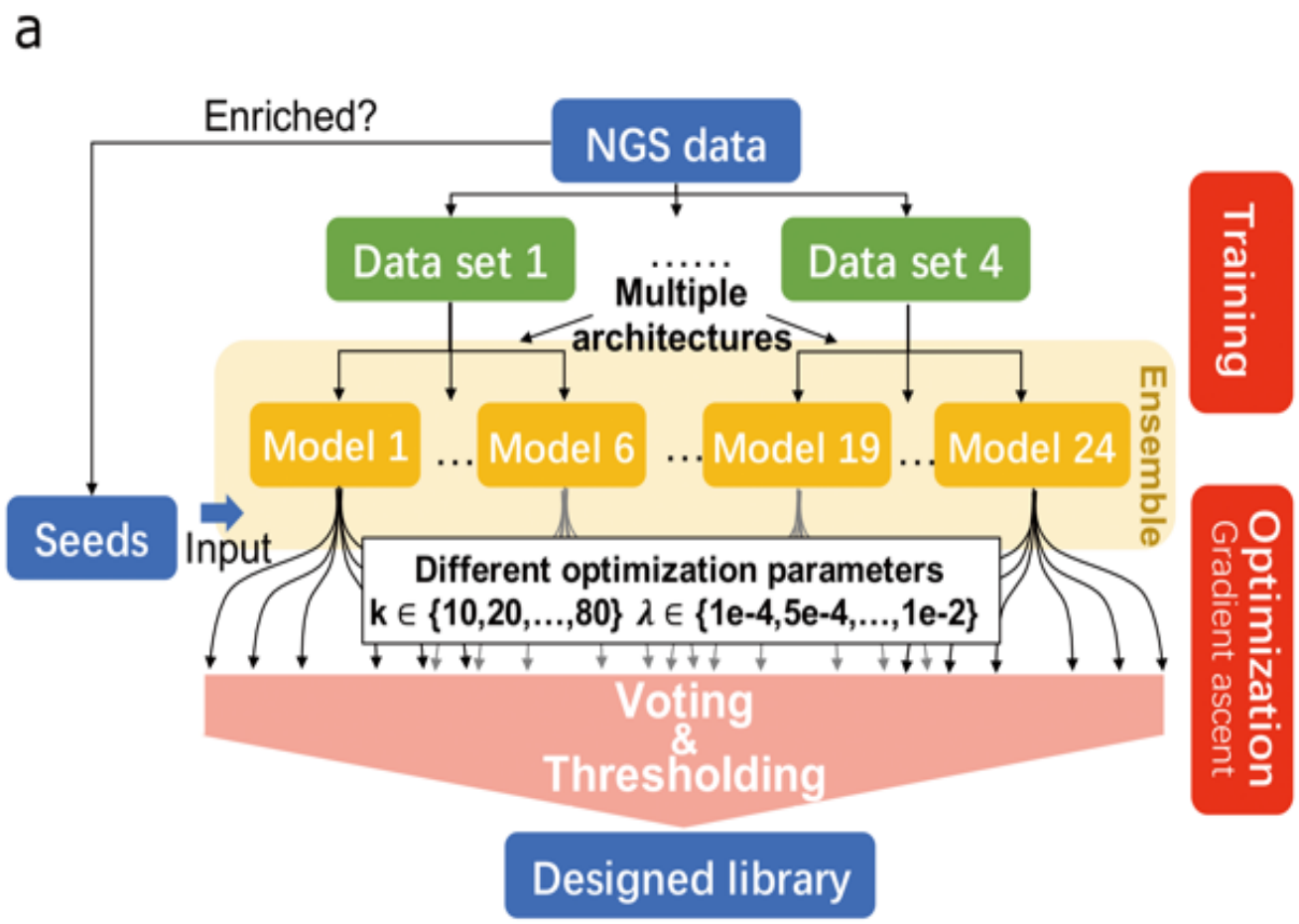
Our model from sequence to enrichment is differentiable
Method 1 - Optimization with gradients



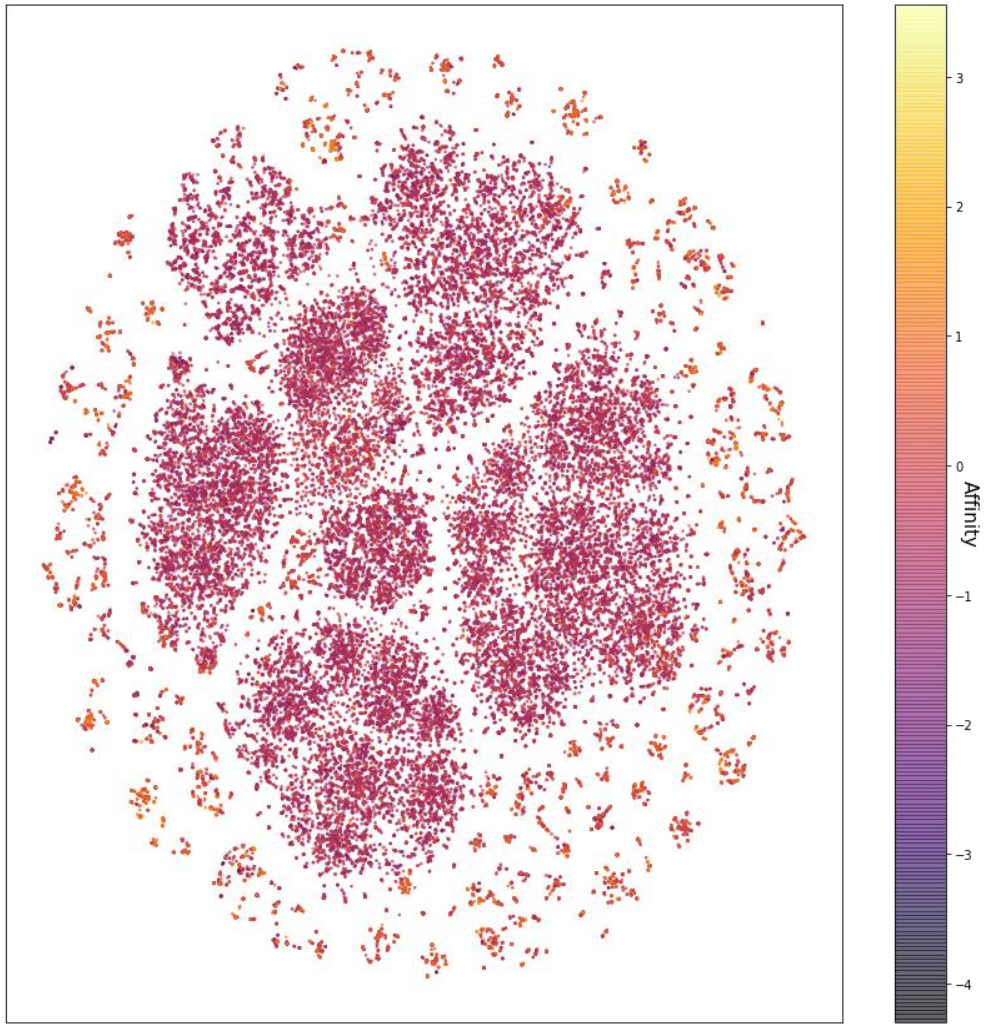
Projecting continuous representation into one-hot representation

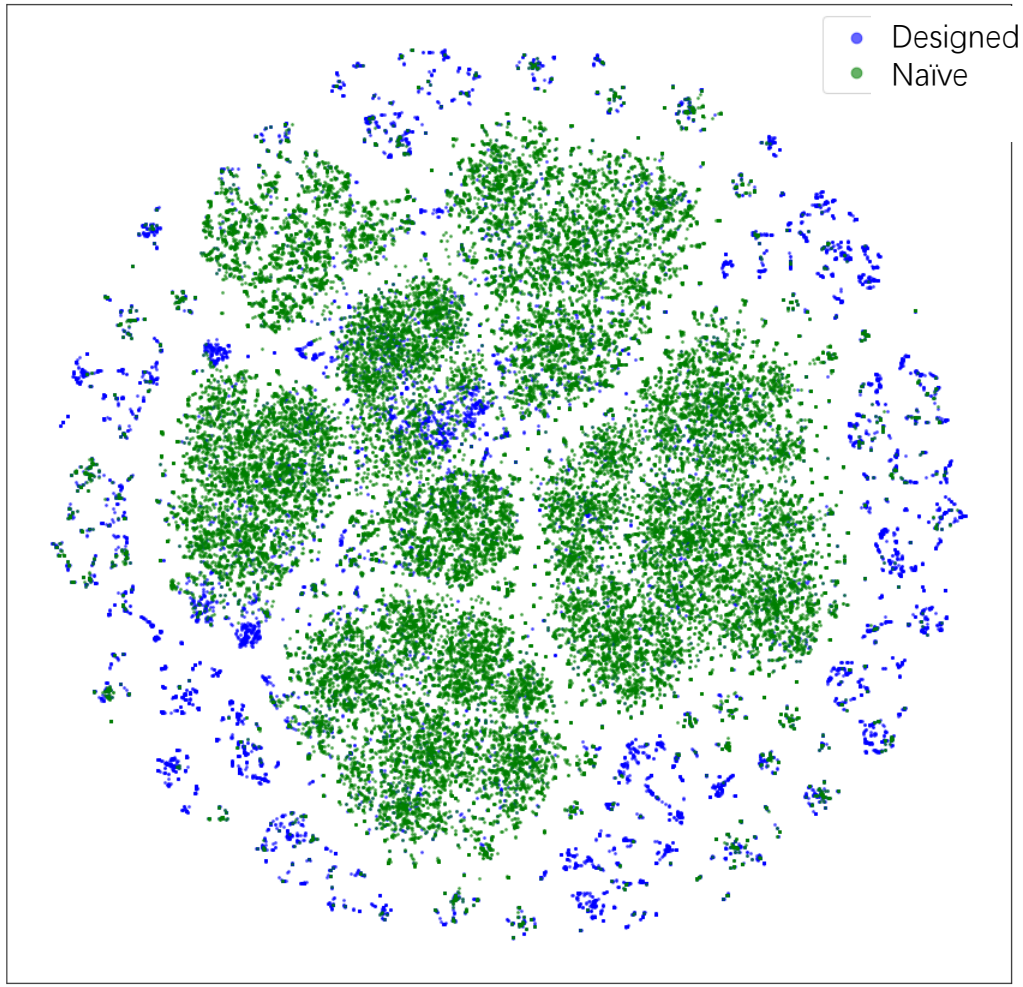


Ens-Grad uses voting across ensembles and hyper-parameters to choose sequences



Designed sequences appear in islands of enrichment

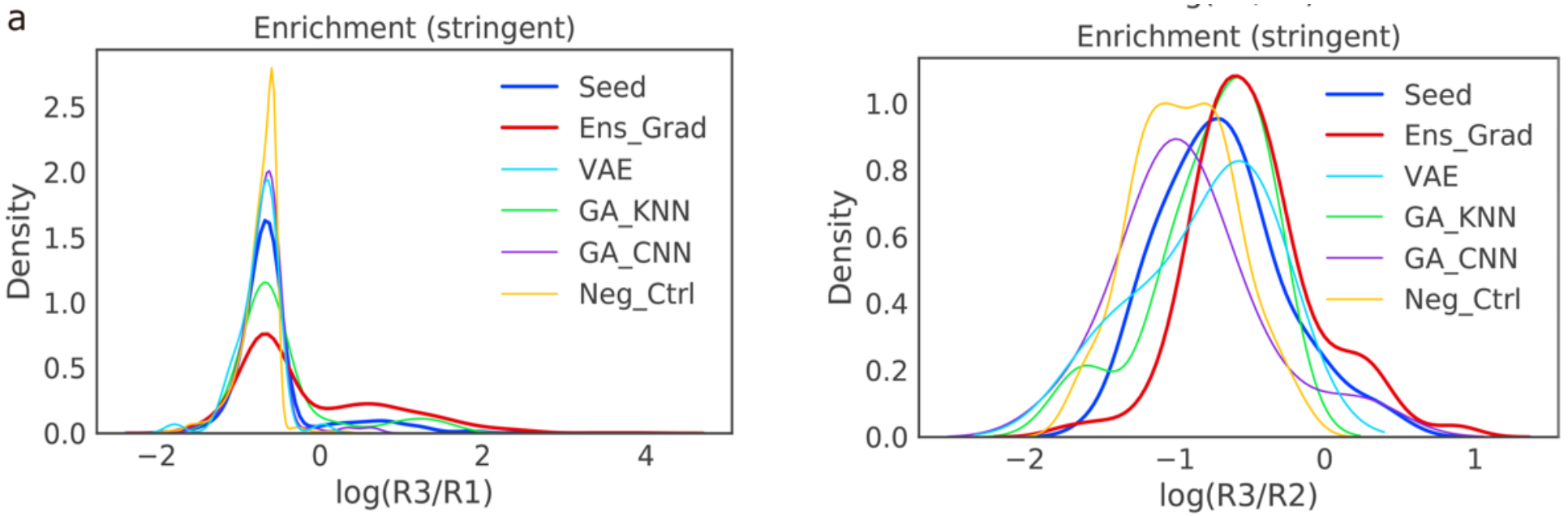




Testing of Fab sequences by direct synthesis

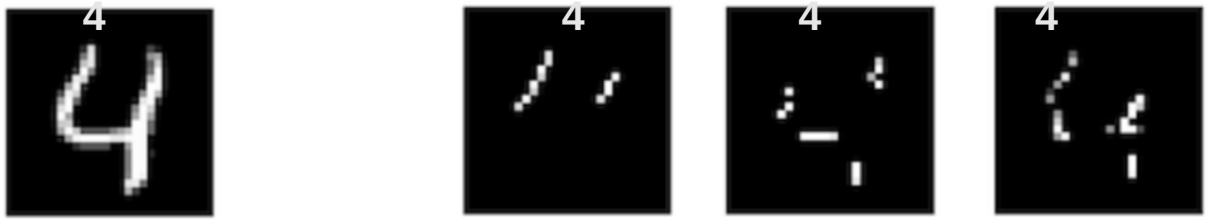
- We computed 77,596 novel machine learning proposed CDR-H3 sequences (Ens-Grad 5,467 sequences)
- We added 26,939 controls and synthesized a total of 104,525 oligonucleotides encoding CDR-H3 sequences
- The oligonuclotides were cloned into a Fab framework and expressed on phage
- Our library with complexity 10^5 was mixed 1:100 into a native library of complexity ~10^{10}
- The combined library was subject to rounds of panning

Ens-Grad sequences are on average more enriched than seeds and the synthetic results of other ML methods



Sufficient Input Subsets provide model interpretation

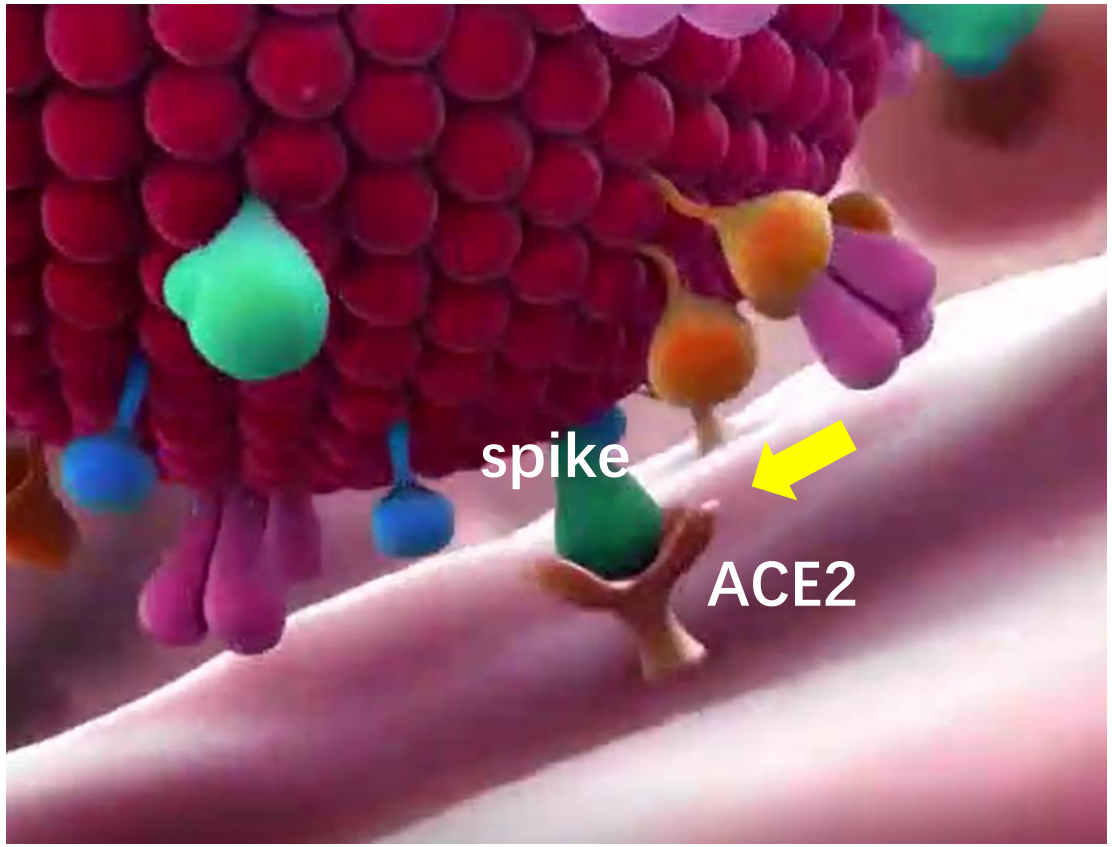
- One simple rationale for ***why*** a black-box decision is reached is a sparse subset of the input features whose values form the basis for the decision
- A ***sufficient input subset*** (SIS) is a minimal feature subset whose values alone suffice for the model to reach the same decision (even without information about the rest of the features' values)

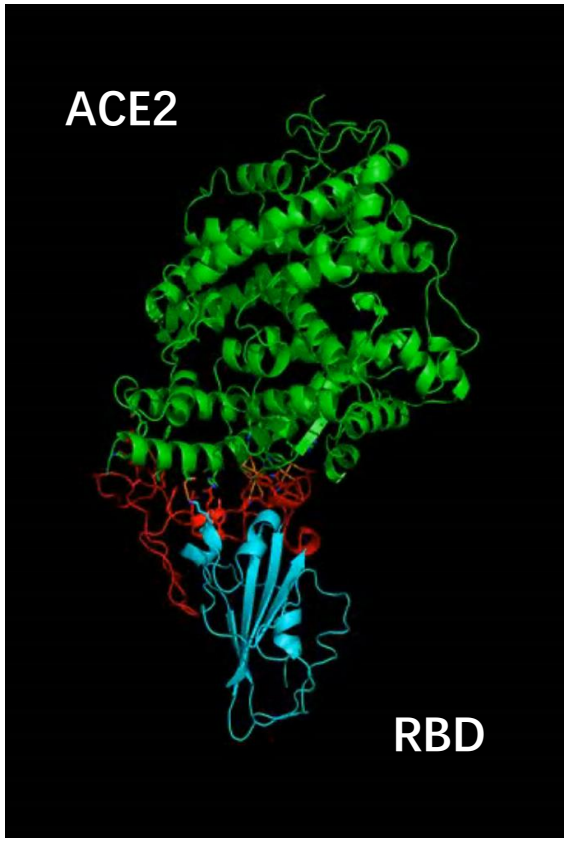


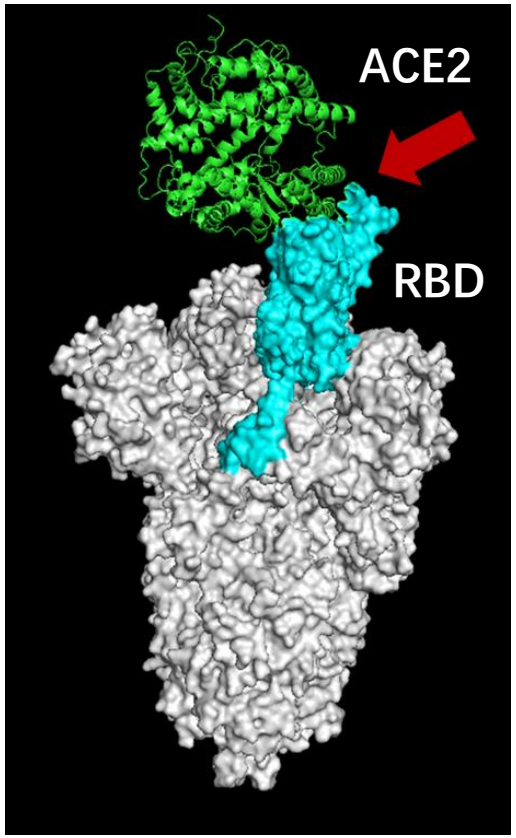
	CDR-H3 Sequence	Group	R^2	EC50(nM)	Standard log(R3/R1)	Stringent log(R3/R1)
Family 1	HKPQAKSYLPLRLLDY	Ens_Grad	0.99	0.47	3.369	2.399
	HKPQAISYLPYRLLDY	Ens_Grad	0.998	0.5	2.61	2.577
	HKPQAISYLPYRILDY	Seed	0.993	0.62	2.418	2.467
	HKPQAKSYLPMRLLDY	Ens_Grad	0.98	0.93	2.409	0.836
	HKPQAVSYLPYRILDY	Ens_Grad	0.994	0.98	2.915	2.561
	HKPQAKSYLPYRLLDY	Seed	0.996	1.48	2.693	1.128
	HKPQAKSYLPYRTLDY	Seed	0.993	2.49	2.371	1.986
	HKPQSKSYLPYRLLDY	Seed	0.995	4.78	2.634	0.445
	HKPQAKSYLPYRILDY	Seed	0.992	6.55	1.41	1.112
Family 2	YRSPHHRGGATWQFDY	Seed	0.992	5.79	-0.037	0.036
Family 3	DLFRYYYFMWPLDY	Ens_Grad	0.986	34.05	2.638	0.523
	DLFRYYYFFWPLDY	Seed	0.99	109.5	2.988	1.283
Family 4	MHYYDIGVFPWDTFDY	Ens-Grad	0.971	**0.29**	2.089	3.381
	GHYYDIGVFPWDTFDY	Seed	0.99	0.49	0.703	1.593
Family 5	WQQWAGYPRQKYSFDY	Seed	0.986	3.31	2.657	1.888
	WQQWSGYPRQKYSFDY	Seed	0.975	66.81	0.264	-0.219
Family 6	GKSLYGQETTWPHFDY	Seed	0.99	0.67	2.002	0.946

Neutralizing antibodies for COVID-19 Therapeutics

Interaction between RBD of spike and ACE2



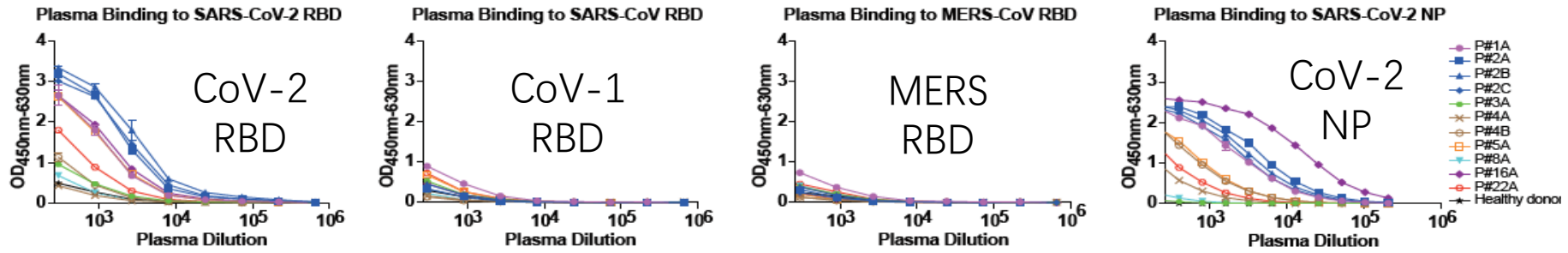




Lan et al., Nature March 30, 2020

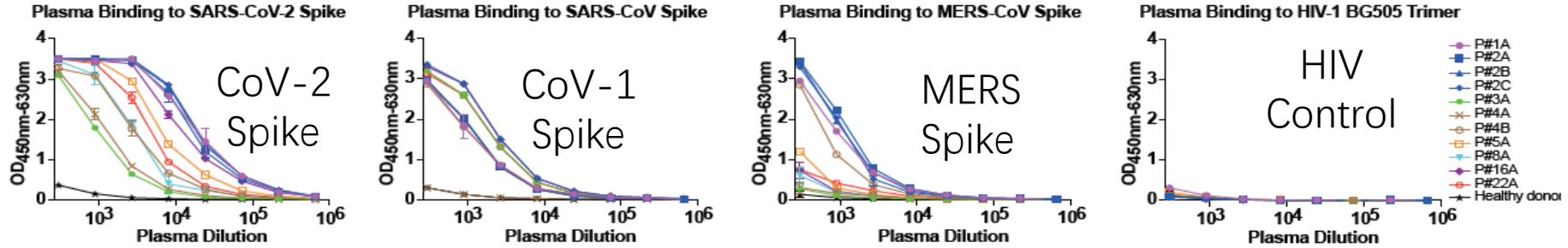
Plasma reactivity to RBD and spike

Binding to Receptor Binding Domain



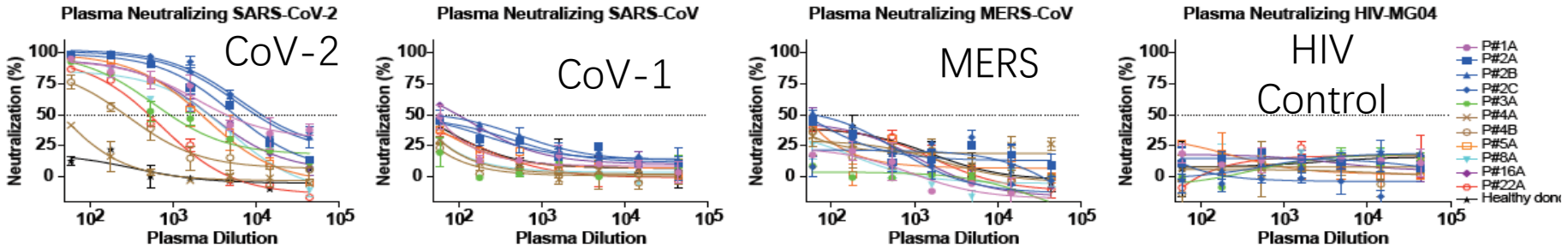
Binding to Spike

B

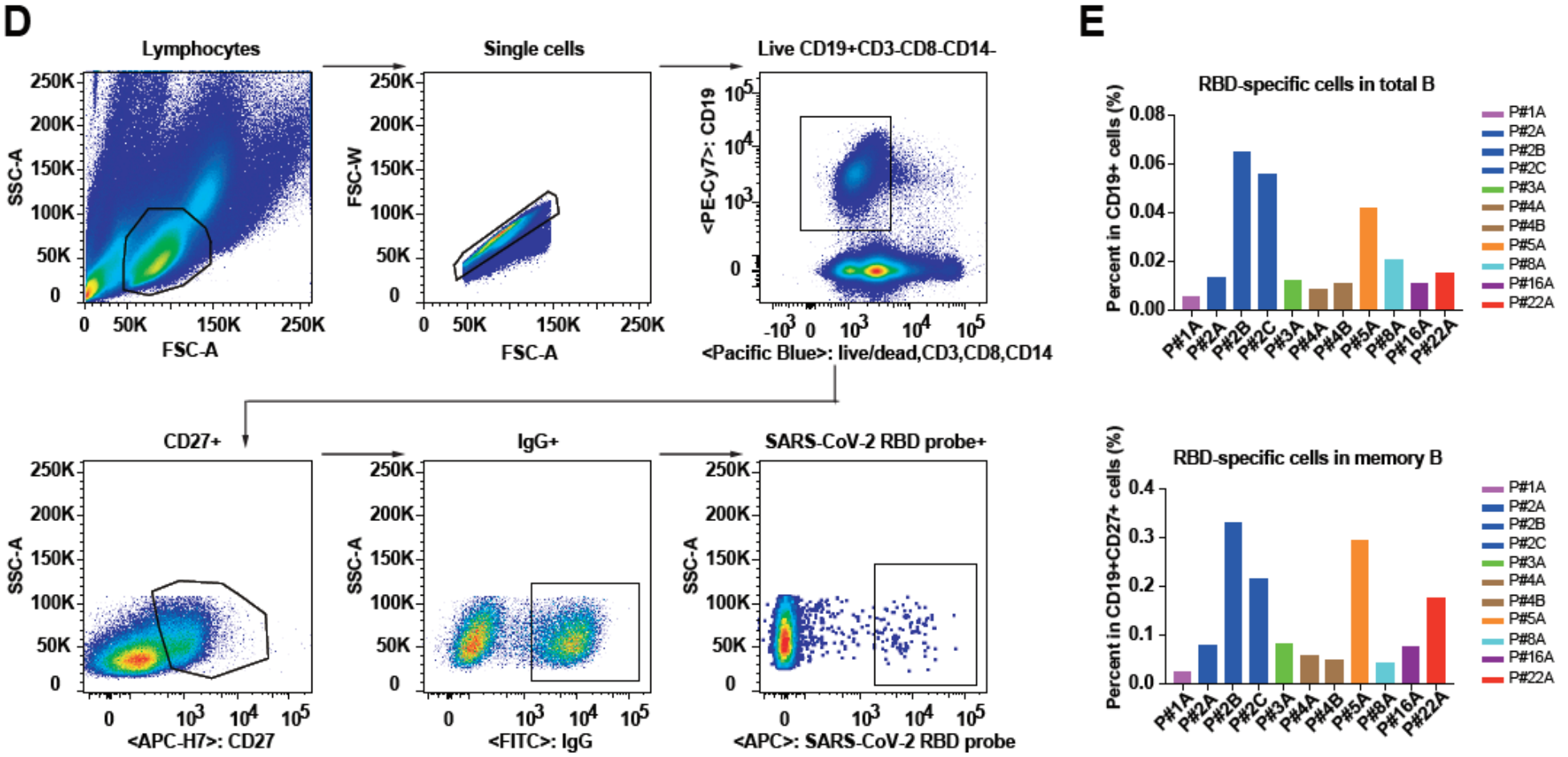


Plasma Neutralization

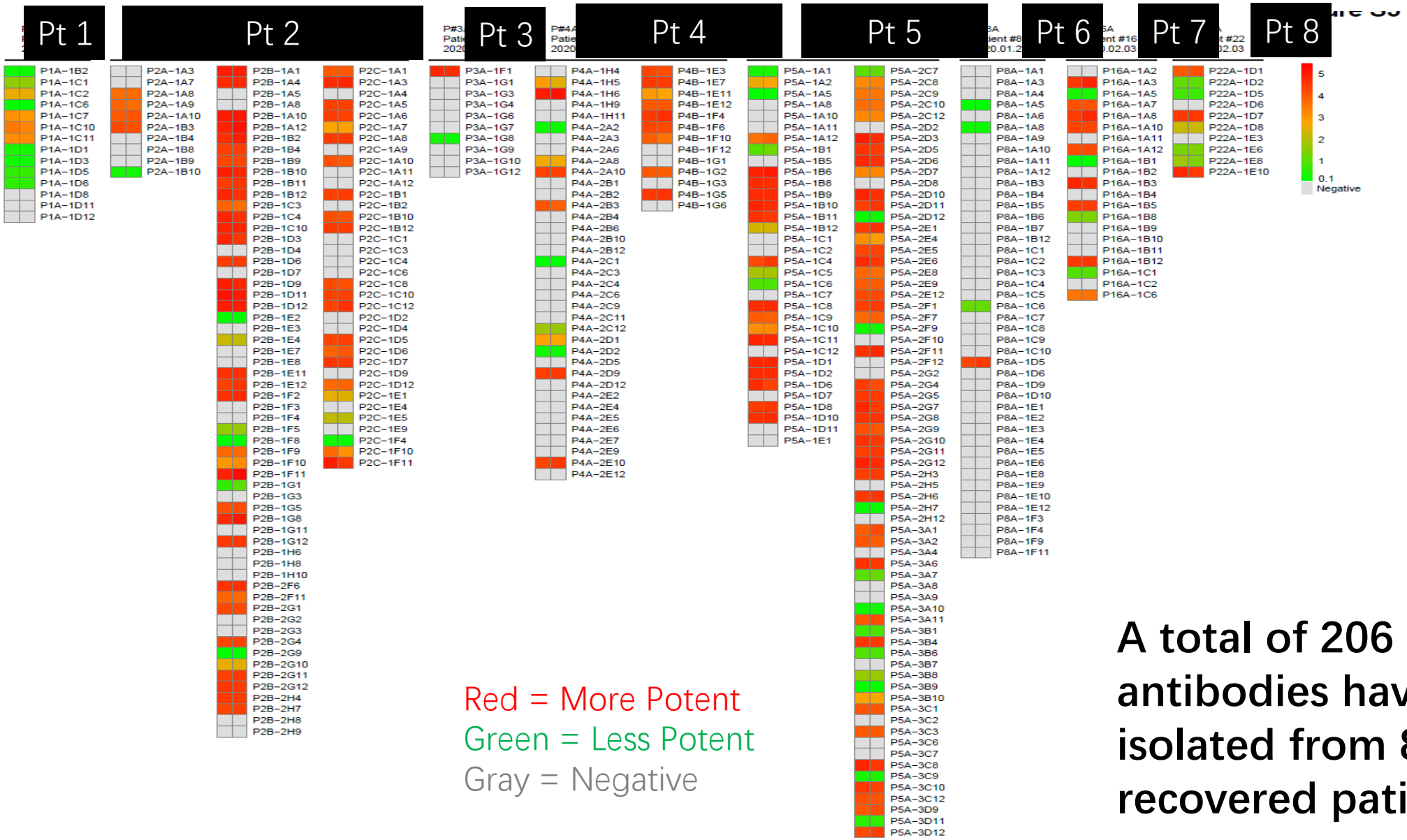
C



Isolating RBD-specific single B cells



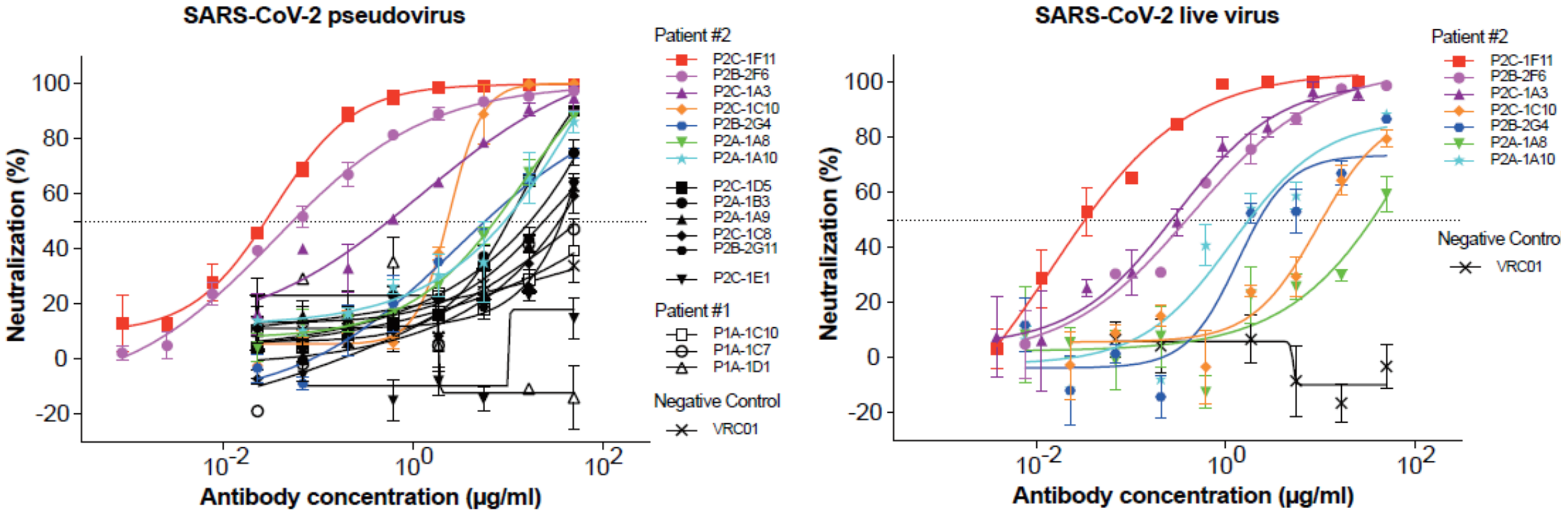
Linqi Zhang, PhD School of Medicine Tsinghua University



A total of 206 antibodies have been isolated from 8 recovered patients

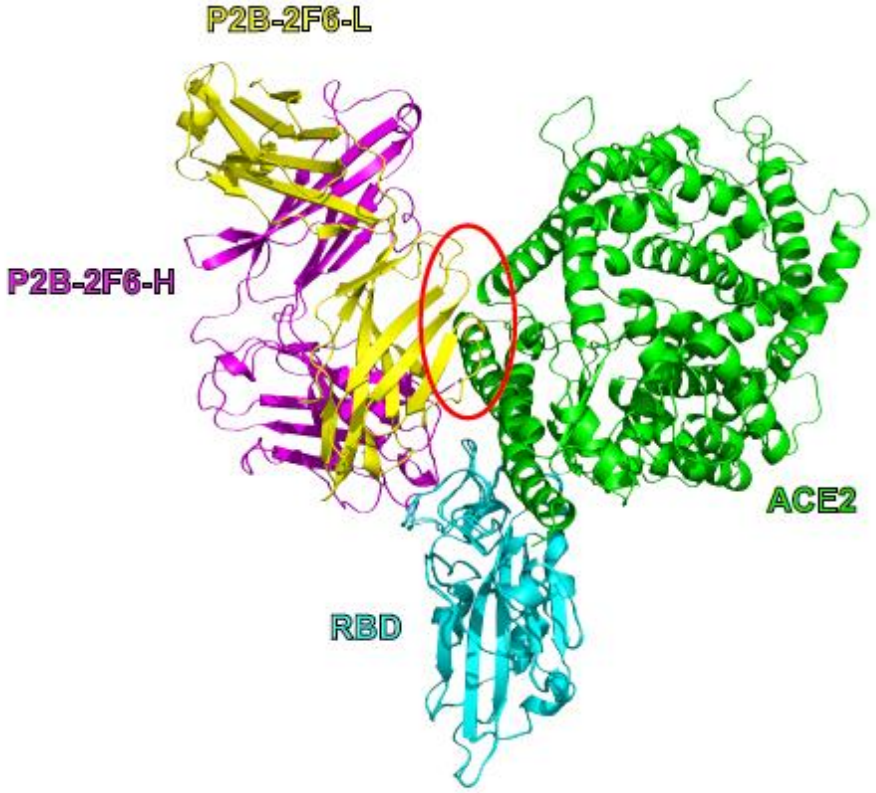
Linqi Zhang, PhD School of Medicine Tsinghua

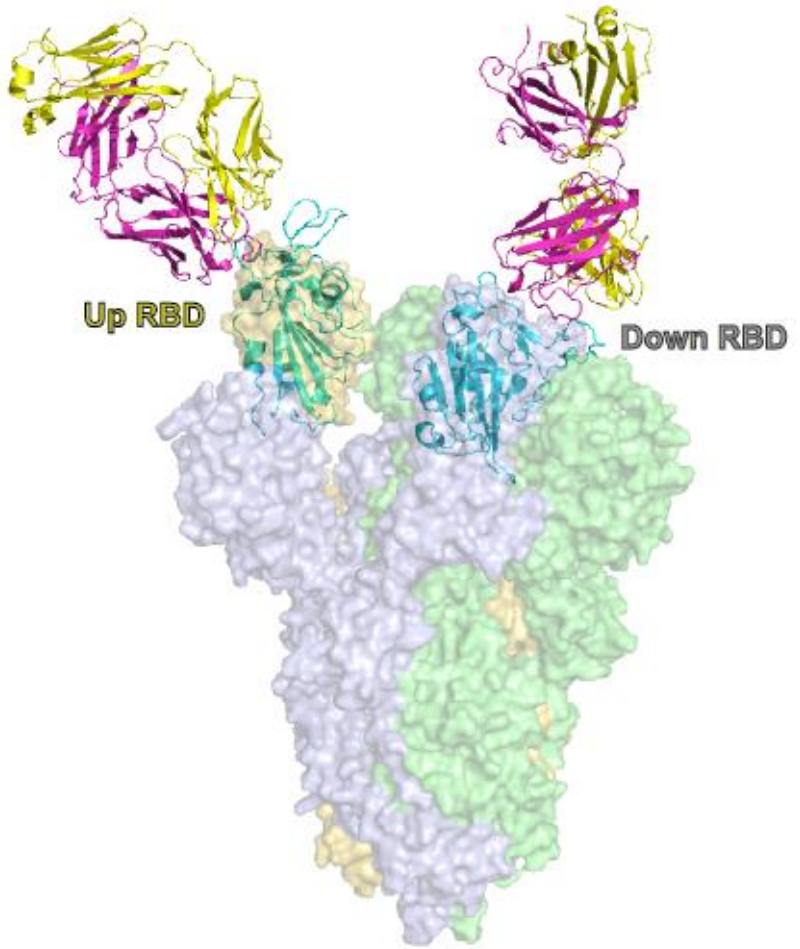
Neutralizing activity of isolated mAbs



Linqi Zhang, PhD School of Medicine Tsinghua University

Structural basis for antibody neutralization





Linqi Zhang, PhD School of Medicine Tsinghua University