Cov2GenX

An intelligent model for predicting mutation-resistant peptide candidates for Covid-19 vaccines

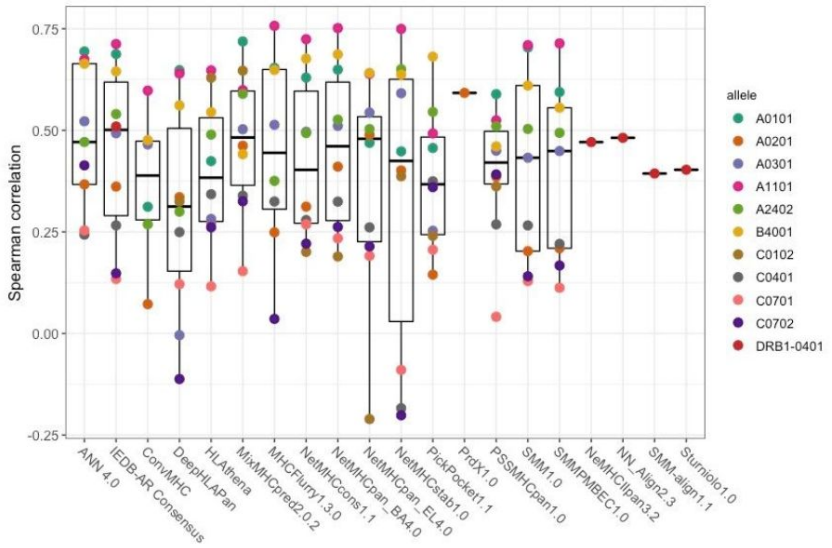
Presented at the AIMI-HIAE COVID-19 Researchathon

Center for Artificial Intelligence in Medicine & Imaging

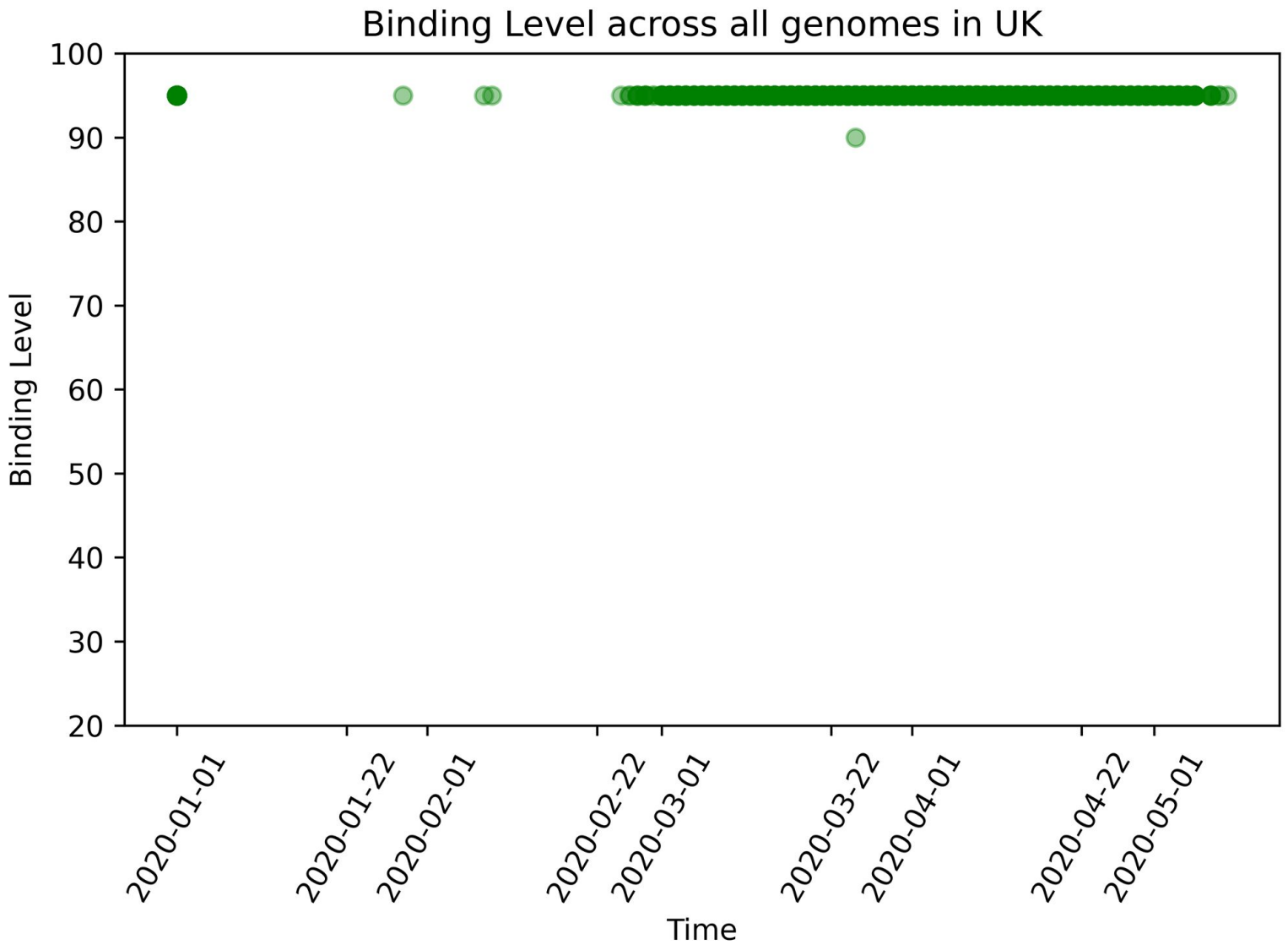
Stanford University

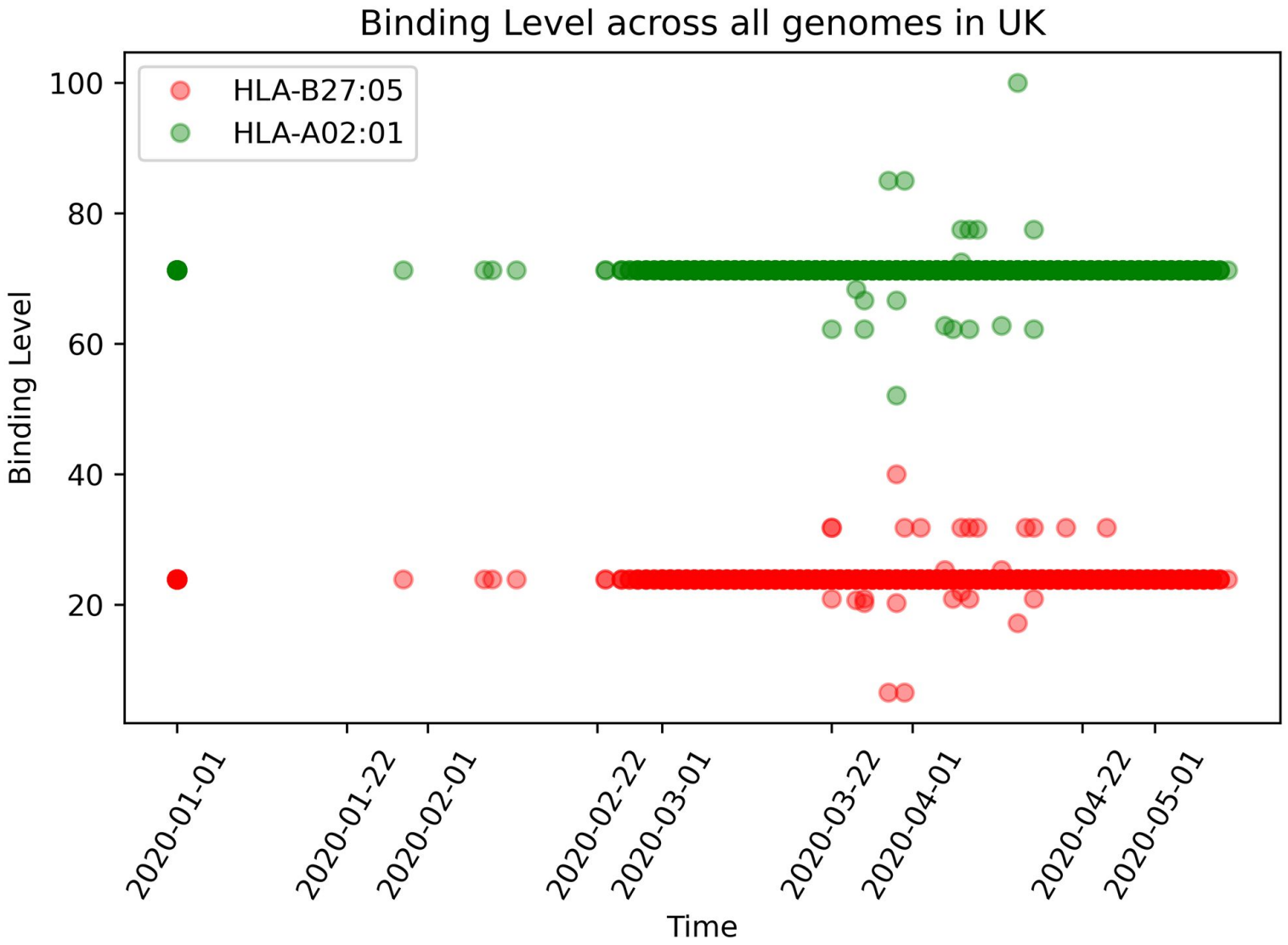
Problem

1. No peptide vaccine candidates
2. No mutation data inclusivity
3. No immunity profiling
4. Assumption that ‘one’ will fit all



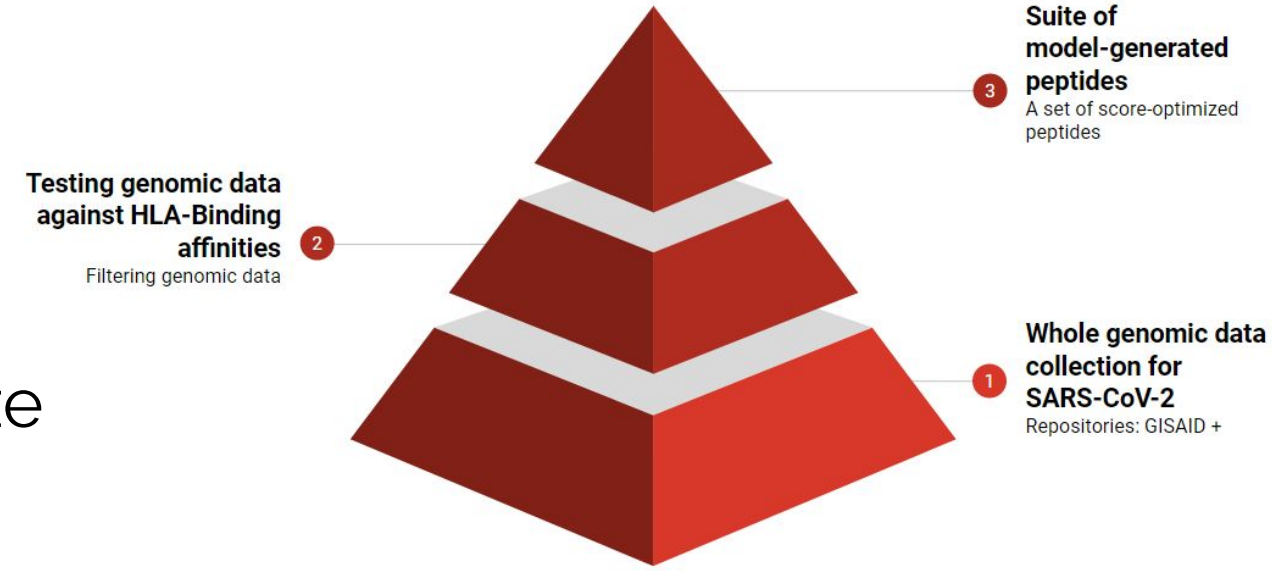
1. Top 5 candidate vaccines currently under clinical trial are viral or nucleotide based
2. Vaccines are being developed with localized genomic data, excluding potential mutations which can reduce their efficacy and applicability
3. Differences in HLA-haplotyping amidst global populations can have a direct impact on the immunological response to the vaccine


Binding Level across all genomes in UK
Binding Level
100
90
80
70
60
50
40
30
20
2020-01-01
2020-01-22
2020-02-01
2020-02-22
2020-03-01
2020-03-22
2020-04-01
2020-04-22
2020-05-01
Time


Binding Level across all genomes in UK
HLA-B27:05
HLA-A02:01
Binding Level
100
80
60
40
20
2020-01-01
2020-01-22
2020-02-01
2020-02-22
2020-03-01
2020-03-22
2020-04-01
2020-04-22
2020-05-01
Time

Solution

- Curate pipeline to generate potential peptides for vaccine
- Inclusion of large scale genomic mutation data for a robust set of mutation-resistant peptides
- Inclusion of regionally derived HLA-haplotype data to categorize peptides based on race and ethnicity



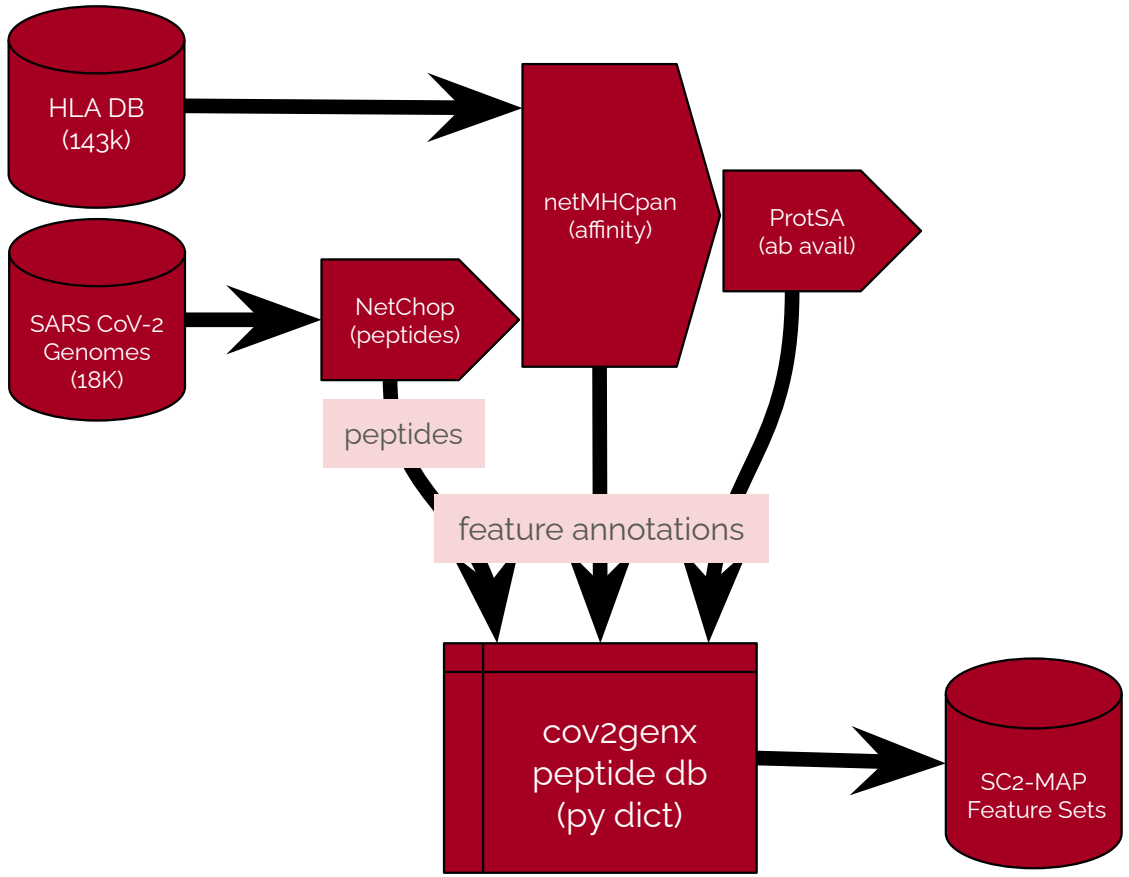
Workflow pipeline

Collect HLA-haplotype and genomic data

Apply automated pipeline for peptide identification and scoring

Curate a robust dataset with HLA-specific peptide data

Train Gradient Boosted Decision Trees to generate **mutation-aware vaccine suite**



SC2-MAP Dataset

Features – per-peptide, multi-genome, multi-HLA

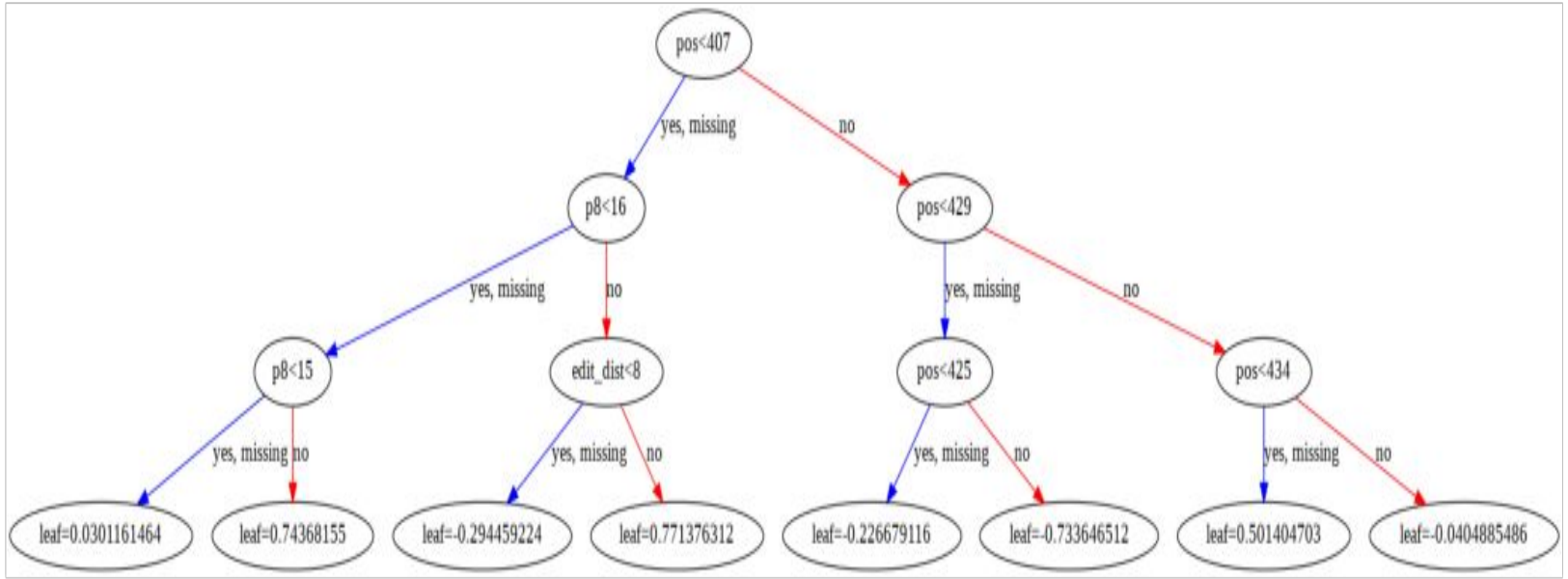
- Peptide (individual amino acids, in order) [Categorical]
- Position in protein [Integer]
- Edit distance (number of mutations in the peptide) [Integer]
- HLA (type of immunological response to the vaccine) [Categorical]
- Location (city/country of sample collection) [Categorical]
- Relative date of sample collection [Integer]

Labels

- Mutation-aware binding score [Float]

Results

- Train set: Samples collected before 1 April 2020
- Test set: Samples collected after 1 April 2020
- Trained an ensemble of 100 XGBoost GBDTs to predict mutation-aware binding scores
- Obtained a Regression NRMSE of 0.03399.
- Performed the Chi-square test to reject the null hypothesis and ensure a goodness of fit



Why GBDTs over Neural Networks

- Explainable AI (XAI) standpoint
  - Neural networks are black boxes
  - Government regulations prevent many uses in medical industry

- In contrast, GBDTs allow
  - Trace decision path
  - Provide confidence and model interpretability

- Provides estimate of feature importances
  - Debug and improve the model
  - Provide researchers insights on issues at hand (Location based mutations, etc)

Work done so far

- Identified a niche problem
- Generated a novel dataset (347,022,410 examples)
- Processed and filtered the dataset
- Trained Gradient Boosted Decision Trees

Moving forward

- Incorporate solvent accessibility
- Further train the model using a larger genome database
- Compare rank-scored peptides vs. peer reviewed scores
- *In silico* analysis of peptide suite for drug-likeness and ADMET

Meet the team

Michael D'Amour

RareKidneyCancer.org

D'Amour & Associates

Grzegorz Preibisch

MD student

BSc Math student

Co-founder of Covid Genomics

Aishwarya Chander

BSc Biomedical Sciences & Bioinformatics

Sharan Ramjee

Stanford MSCS

Purdue BS CmpE

ML Engineer

Michał Borkowski

MSc Computer Science Programmer

Piotr Grzegorczyk

BSc Math Student

Co-Founder of Covid Genomics

Questions?

Thank you.

References

1. Nguyen A, David JK, Maden SK, et al. Human leukocyte antigen susceptibility map for SARS-CoV-2. Journal of Virology. April 2020. doi:10.1128/jvi.00510-20

2. Prachar M, Justesen S, Steen-Jensen DB, et al. COVID-19 Vaccine Candidates: Prediction and Validation of 174 SARS-CoV-2 Epitopes. bioRxiv. March 2020. doi:10.1101/2020.03.20.000794

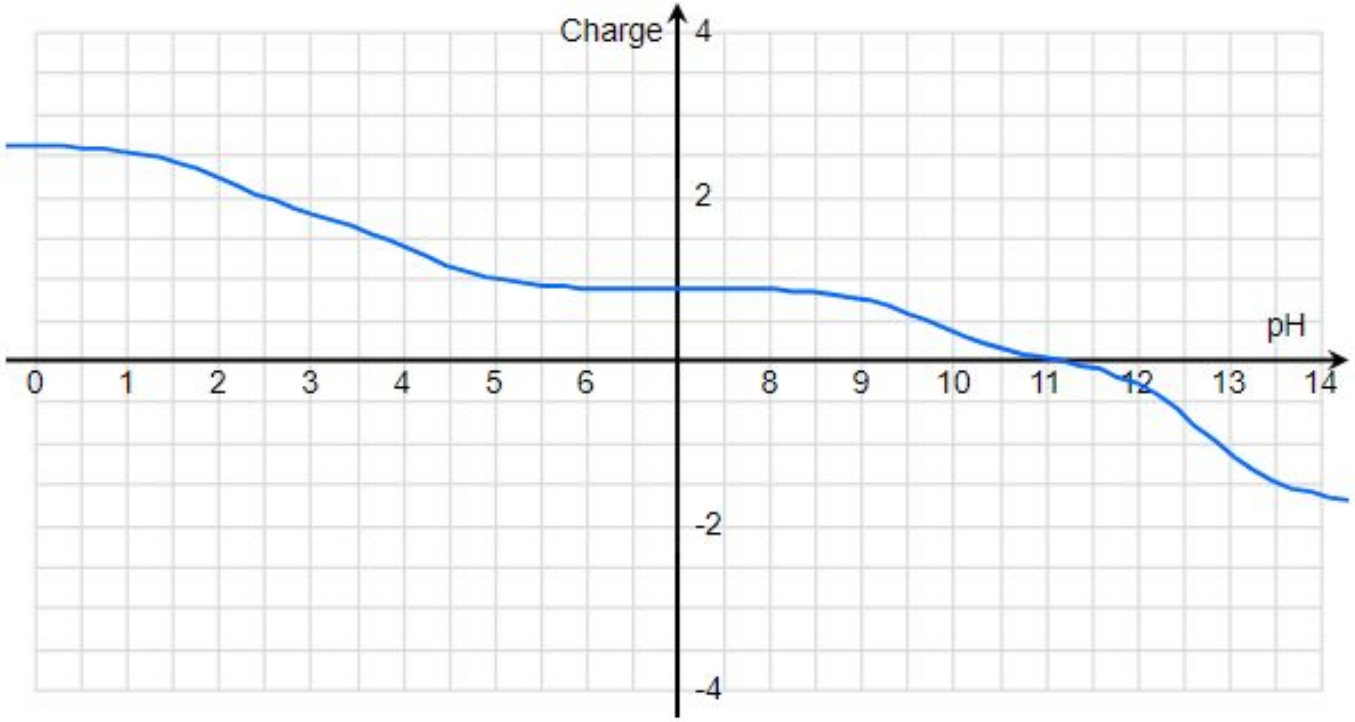
3. Kennedy RB, Ovsyannikova IG, Vierkant RA, Jacobson RM, Poland GA. Effect of human leukocyte antigen homozygosity on rubella vaccine–induced humoral and cell-mediated immune responses. Human Immunology. 2010;71(2):128-135. doi:10.1016/j.humimm.2009.11.002

4. Gartland AJ, Li S, McNevin J, et al. Analysis of HLA A*02 Association with Vaccine Efficacy in the RV144 HIV-1 Vaccine Trial. Journal of Virology. 2014;88(15):8242-8255. doi:10.1128/jvi.01164-14

Archive

Biochemical analysis of a peptide

Sequence 1



Analyzed sequence

N-Q-N-G-E-R-S-G-A-R

Sequence length	10
Hydrophobicity	0.84
GRAVY	-2.28
MW average	1088.1133 g/mol
MW monoisotopic	1087.5119
Theoretical pI	10.9

3

Suite of model-generated peptides

A set of score-optimized peptides

2

Testing genomic data against HLA-Binding affinities

Filtering genomic data

1

Whole genomic data collection for SARS-CoV-2

Repositories: GISAID +