

Model evolution in SARS-CoV-2 protein sequences using a generative neural network

Anup Kumar

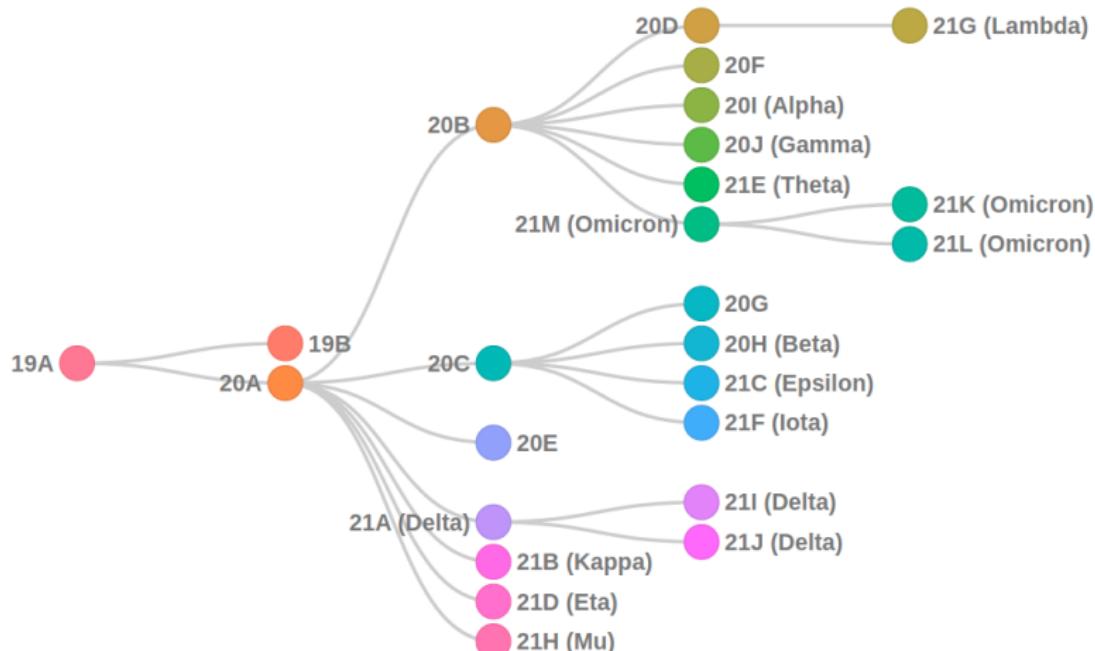
Bioinformatics group, University of Freiburg, Freiburg, Germany

March 2, 2022

Prediction of protein evolution (amino acid substitutions) in SARS-CoV-2 sequences

- Amino acid (AA) substitutions in spike protein
- Phylogenetic tree of clades
- Sequence to sequence learning with encoder-decoder neural network
- True target and generated SARS-CoV-2 AA sequences
- Frequency of substitutions per genomic position (POS)
- Substitutions from generated AA sequences (future substitutions)

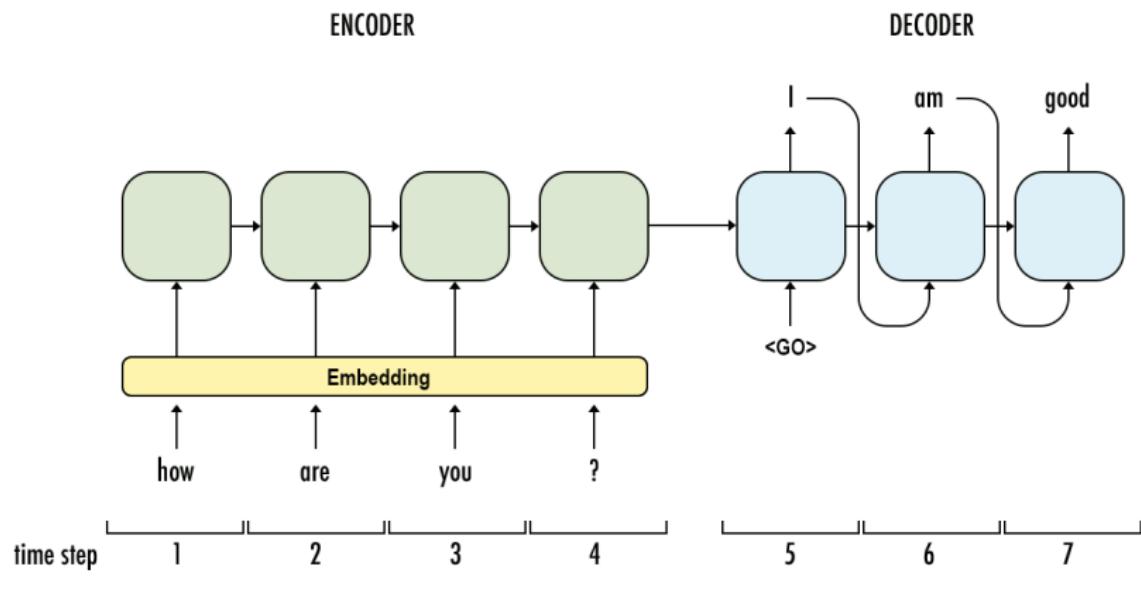
Phylogenetic tree of clades



1

¹<https://clades.nextstrain.org/>

Sequence to sequence learning



²<https://towardsdatascience.com/sequence-to-sequence-model-introduction-and-concepts-44d9b41cd42d>

Data preparation

- ① Assign clades using Galaxy and filter using 'qc.overallStatus == "good"'
- ② Select sequences belonging to source (20A) and target (20B) clades randomly
- ③ Make pairs via all-vs-all combination
- ④ Do quality control - remove pairs with Levenshtein distance in (0, 11), sequences with other than 20 amino acids (seqs with Xs), duplicate pairs
- ⑤ Preprocess - convert to Kmers ($K = 3$, 8000 combinations with 20 amino acids), represent each kmer with unique integer
- ⑥ Sequence of amino acids \downarrow sequence of kmers \downarrow sequence of integers
- ⑦ Divide pairs - train (20A-20B, size: 10,000) for learning and test (20A-20B, size: 2,000) for evaluating and generating

Clade assignments using Galaxy EU

Galaxy Europe Workflow Visualize Shared Data Help User Using 1.2 TB

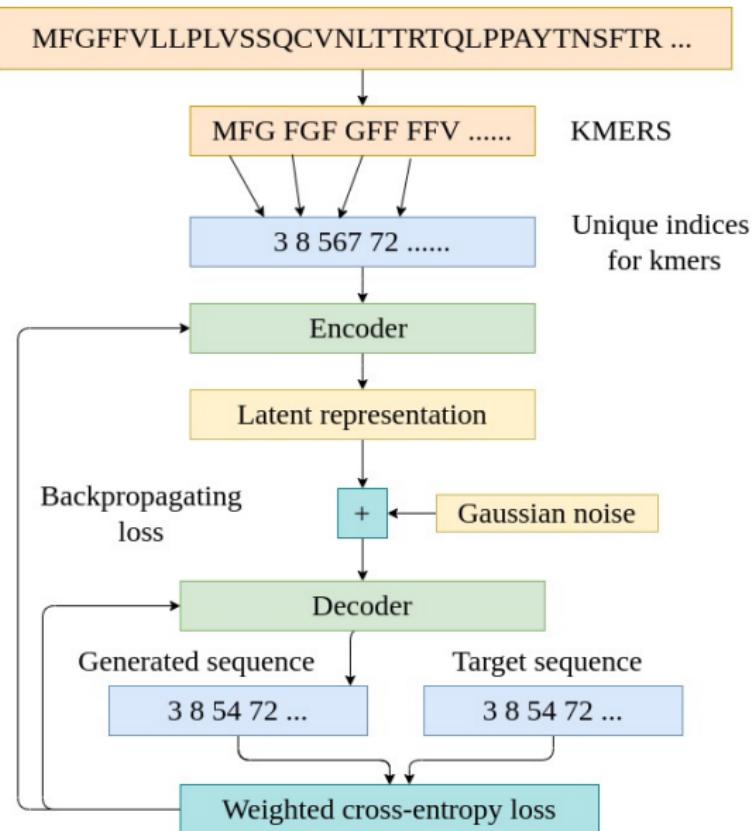
seqName	clade	qc.overallScore	qc.overallStatus	tot
Spike hCoV-19/Wuhan/WIV04/2019 2019-12-30 EPI_ISL_402124 Original hCoV-19^^Hubei Human Wuhan	19A	0.000000	good	
Spike hCoV-19/Switzerland/ZH-ETHZ-460863/2021 2021-01-16 EPI_ISL_1001482 Original hCoV-19^^Zurich Human Viollier	20A	0.000000	good	
Spike hCoV-19/Canada/QC-L00254708/2020 2020-04-27 EPI_ISL_1001069 Original hCoV-19^^Quebec Human Laboratoire	20A	0.000000	good	
"Spike hCoV-19/French (2)"	20A	0.000000	good	
Spike hCoV-19/Switzerland/JU-ETHZ-470314/2021 2021-01-24 EPI_ISL_1001983 Original hCoV-19^^Jura Human Viollier	20A	13.662606	good	
Spike hCoV-19/Canada/QC-L00252110/2020 2020-04-20 EPI_ISL_1001077 Original hCoV-19^^Quebec Human Laboratoire	20B	0.000000	good	
Spike hCoV-19/Switzerland/BL-ETHZ-470932/2021 2021-01-26 EPI_ISL_1001489 Original hCoV-19^^Basel-Landschaft Human Viollier	20A	0.000000	good	
Spike hCoV-19/Switzerland/AG-ETHZ-460935/2021 2021-01-17 EPI_ISL_1001866 Original hCoV-19^^Aargau Human Viollier	"20E (EU1)"	0.000000	good	
Spike hCoV-19/Switzerland/AG-ETHZ-461130/2021 2021-01-20 EPI_ISL_1001769 Original hCoV-19^^Aargau Human Viollier	20A	0.000000	good	
Spike hCoV-19/Switzerland/AG-ETHZ-471169/2021 2021-01-27 EPI_ISL_1001699 Original hCoV-19^^Aargau Human Viollier	20A	0.000000	good	
Spike hCoV-19/Ireland/MH-NVRL-21 RL31532/2021 2021-02-04 EPI_ISL_1001131 Original hCoV-19^^Meath Human National	20B	0.000000	good	
Spike hCoV-19/Switzerland/BE-ETHZ-450779/2021 2021-01-09 EPI_ISL_1001687 Original hCoV-19^^Bern Human Viollier	"20E (EU1)"	0.000000	good	
Spike hCoV-19/Switzerland/BL-ETHZ-480127/2021 2021-01-29 EPI_ISL_1001692 Original hCoV-19^^Basel-Landschaft Human Viollier	"20E (EU1)"	0.000000	good	
Spike hCoV-19/Switzerland/ZH-ETHZ-450888/2021 2021-01-10 EPI_ISL_1001878 Original hCoV-19^^Zurich Human Viollier	20A	0.000000	good	
Spike hCoV-19/France/BRE-IPP02819/2021 2021-01-26 EPI_ISL_1001415 Original hCoV-19^^Bretagne Human Hospital National	20A	0.000000	good	
Spike hCoV-19/Switzerland/ZH-ETHZ-481162/2021 2021-02-03 EPI_ISL_1001852 Original hCoV-19^^Zurich Human Viollier	"20E (EU1)"	0.000000	good	
"Spike hCoV-19/French (3)"	20A	0.000000	good	
Spike hCoV-19/Switzerland/ZH-ETHZ-460711/2021 2021-01-15 EPI_ISL_1001762 Original hCoV-19^^Zurich Human Viollier	"20E (EU1)"	0.000000	good	
Spike hCoV-19/USA/MI-UM-10038279312/2021 2021-02-11 EPI_ISL_1001272 Original hCoV-19^^Michigan Human University	20G	0.000000	good	
Spike hCoV-19/France/GES-IPP02600/2021 2021-01-27 EPI_ISL_1001389 Original hCoV-19^^Grand	20C	0.000000	good	
Spike hCoV-19/France/HDF-IPP02611/2021 2021-01-23 EPI_ISL_1001382 Original hCoV-19^^Hauts-de-France Human Hospital National	20B	0.000000	good	
Spike hCoV-19/Switzerland/ZH-ETHZ-461066/2021 2021-01-18 EPI_ISL_1001935 Original hCoV-19^^Zurich Human Viollier	20A	0.000000	good	
Spike hCoV-19/France/HDF-IPP02628/2021 2021-01-26 EPI_ISL_1001334 Original hCoV-19^^Hauts-de-France Human Hospital National	20B	0.000000	good	
Spike hCoV-19/France/IDF-IPP02585/2021 2021-01-29 EPI_ISL_1001300 Original hCoV-19^^Ile-de-France Human Hospital National	20B	0.000000	good	
Spike hCoV-19/USA/MI-UM-10038267086/2021 2021-02-10 EPI_ISL_1001279 Original hCoV-19^^Michigan Human University	20G	0.000000	good	
> ipike hCoV-19/Spain/CN-IBV-97016142/2021 2021-01-18 EPI_ISL_1000989 Original hCoV-19^^Las	20A	0.000000	good	

³<https://usegalaxy.eu/>

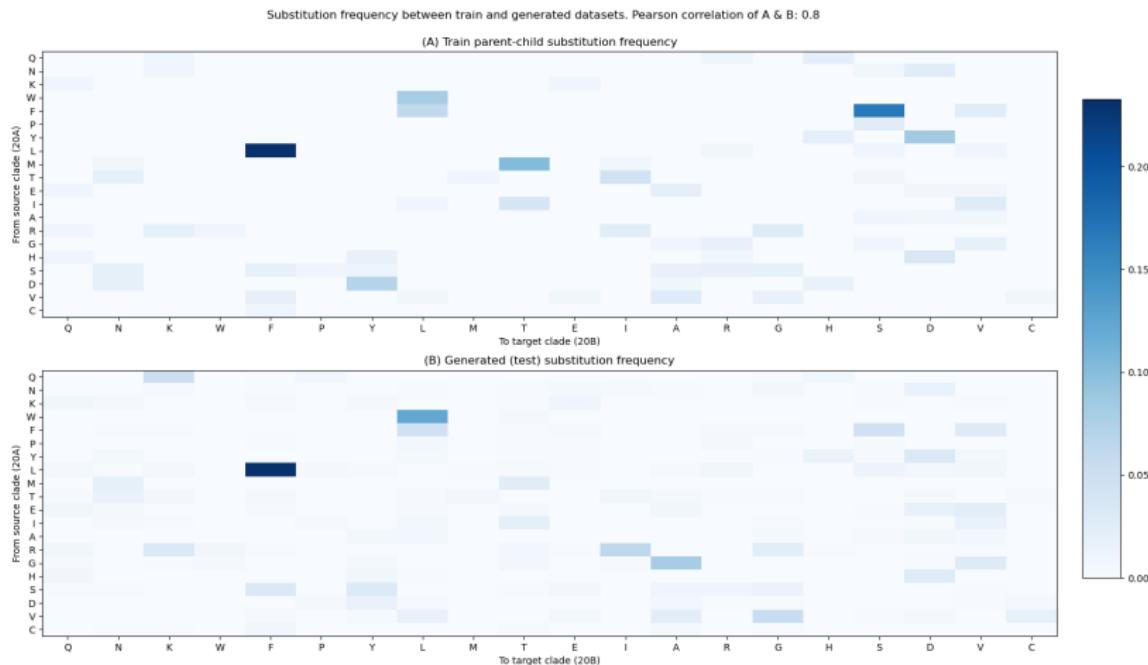
Source (20A) - target (20B) pairs

20A	20B
1. 3299,1965,7288,1759,3168,	780,7585,7685,1699,1968,7348, ...
2. 3299,1965,7288,1759,3168,	780,7585,7685,1699,1968,7348, ...
3. 3299,1965,7288,1759,3168,	780,7585,7685,1699,1968,7348, ...
4.

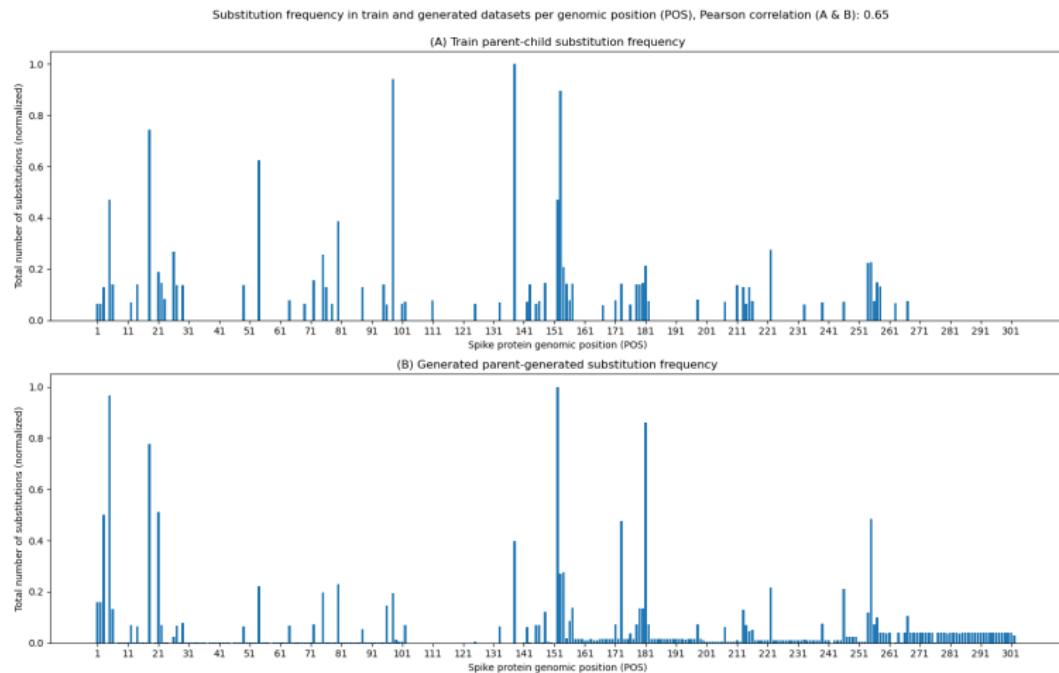
Encoder-decoder generative neural network



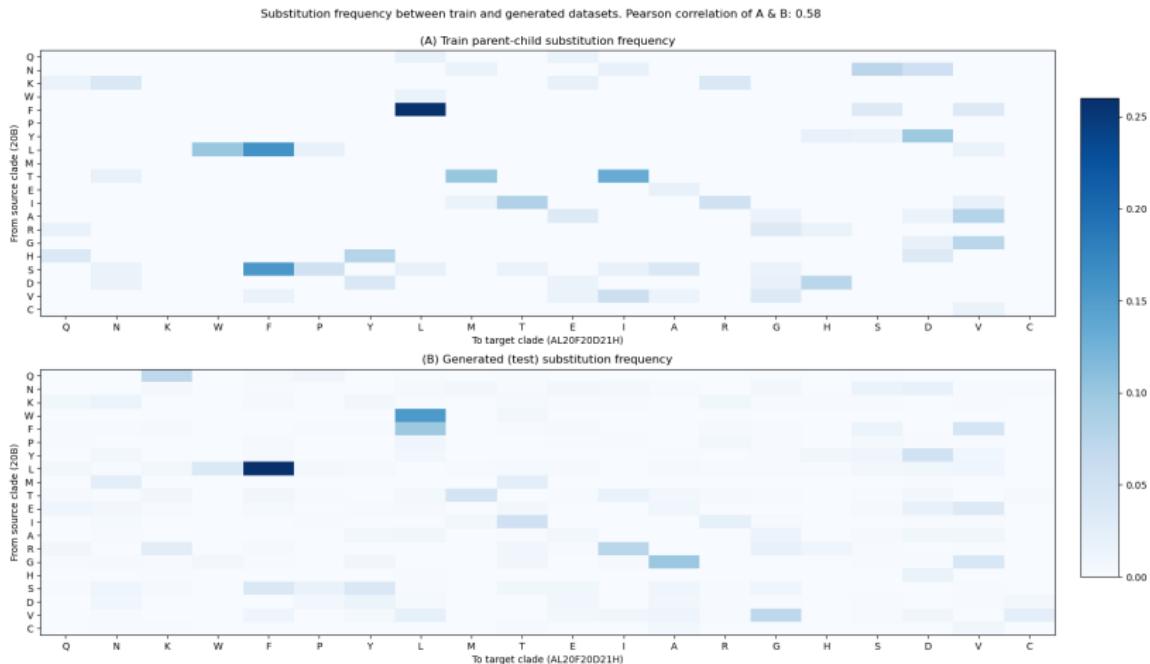
Comparison of protein evolution (for clade 20B using test sequences from 20A)



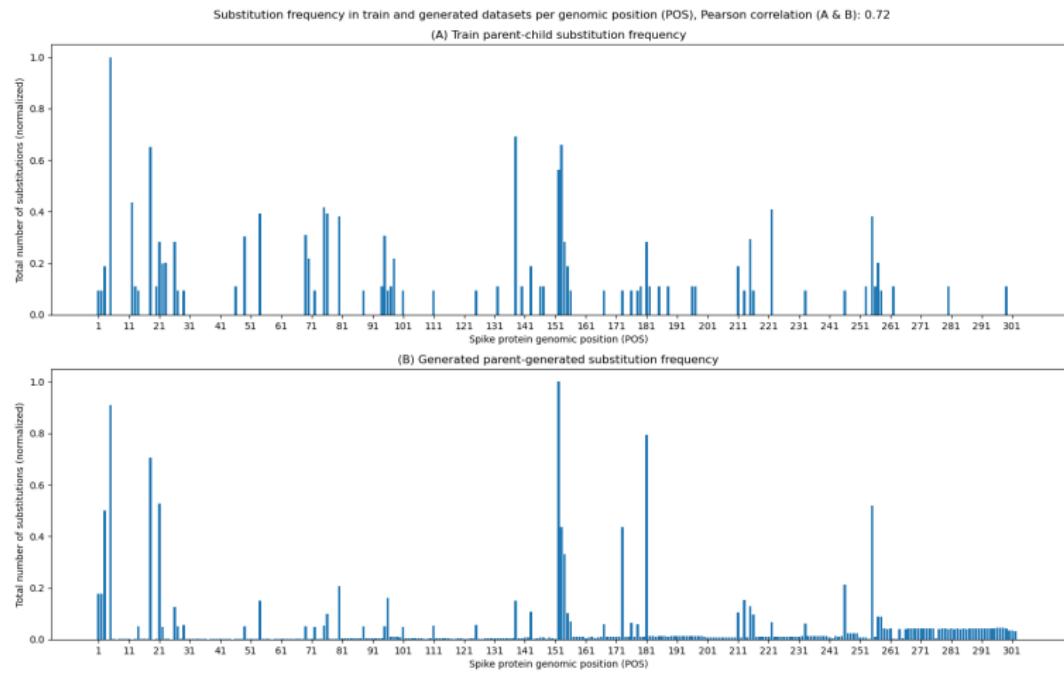
Frequency of substitution compared between training 20A-20B and test 20A-generated



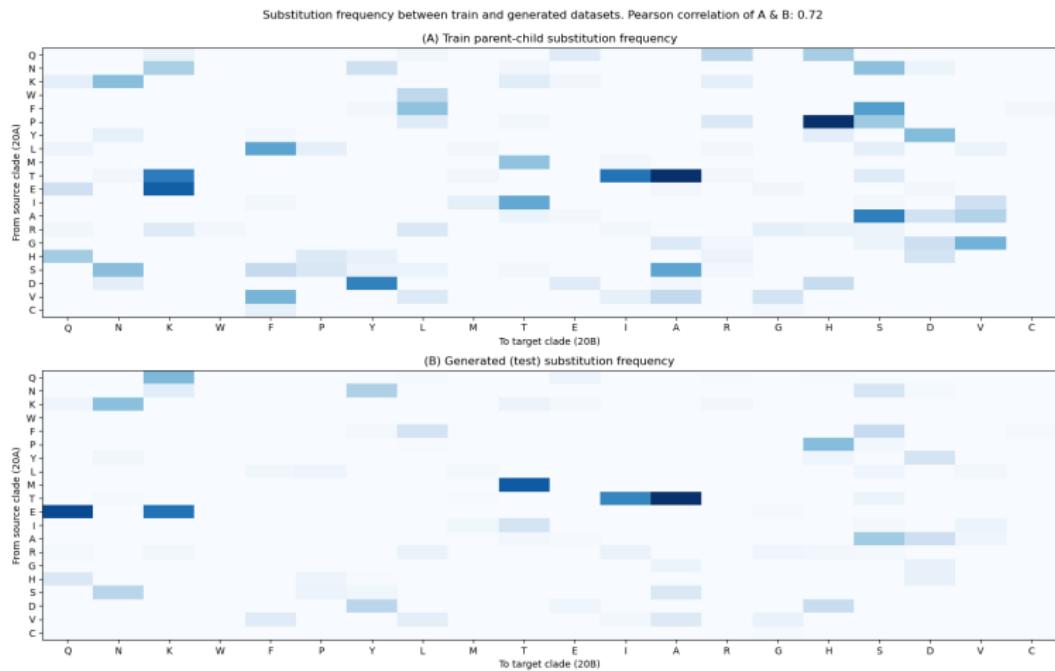
Comparison of protein evolution (for clades Alpha, Lambda, 20D, 20F and 21H using test sequences from 20B)



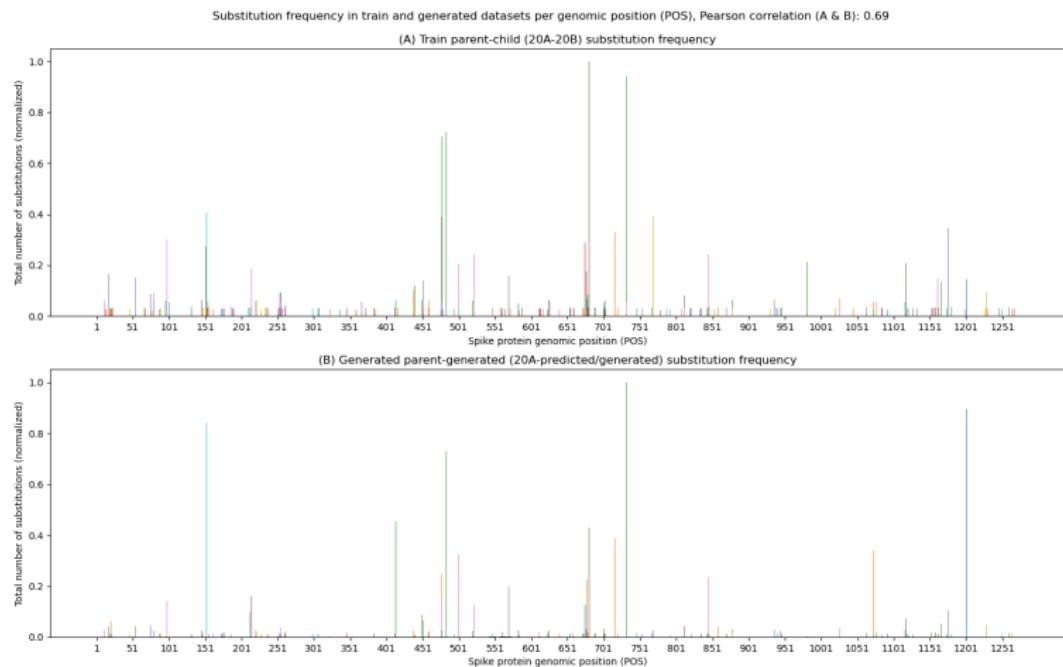
Frequency of substitution compared between test 20B-Alpha, Lambda, 20D, 20F and 21H and test 20B-generated



Comparison of protein evolution (for clade 20B using test sequences from 20A), Entire length (1273)

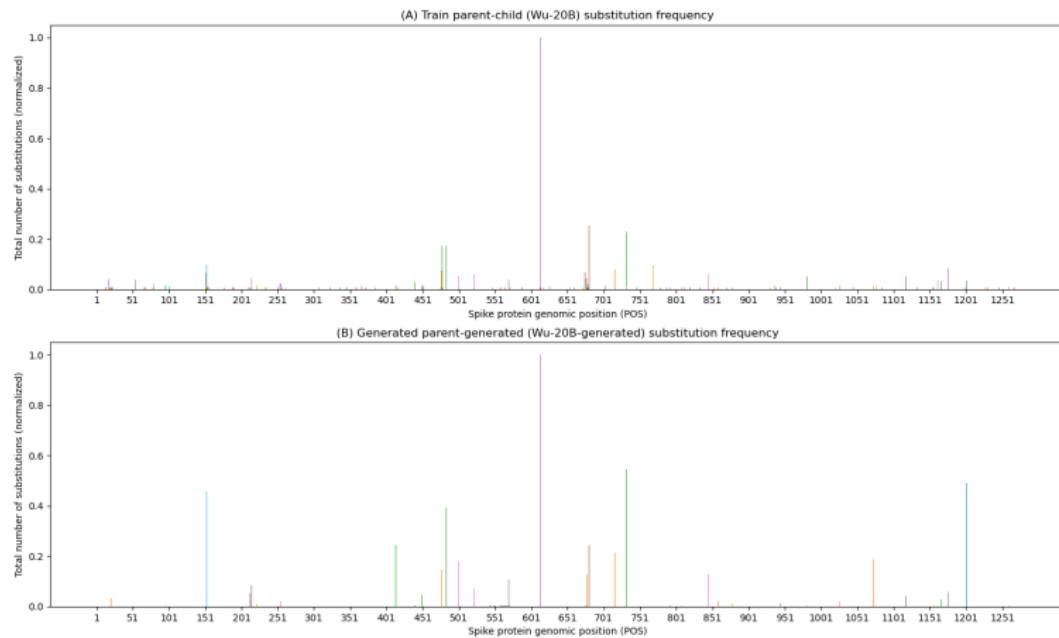


Frequency of substitution compared between training 20A-20B and test 20A-generated, Entire length (1273)

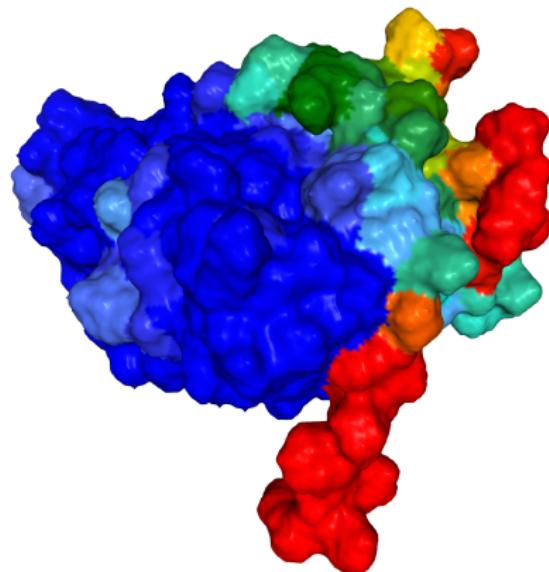


Frequency of substitution compared between training 20A-20B and test 20A-generated, Entire length (1273)

Wu-target: Substitution frequency in true target (20B) and generated target datasets per genomic position (POS). Pearson correlation (A & B):0.83



Predicted structure of a generated sequence (300) using AlphaFold



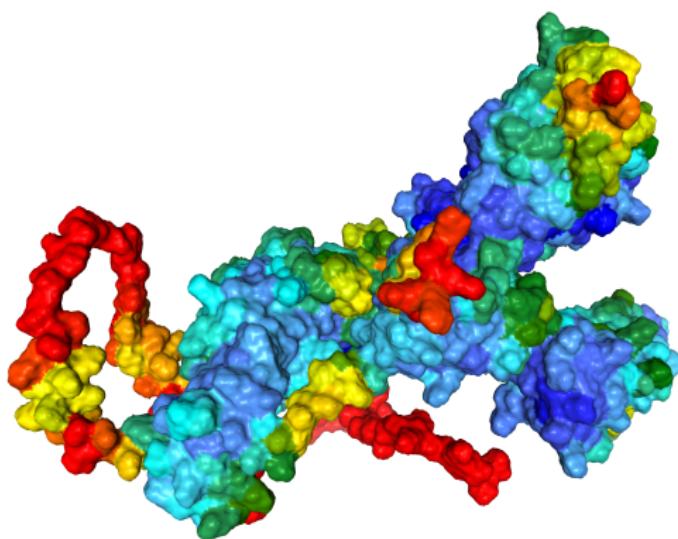
4

⁴<https://usegalaxy.eu/u/kumara/h/alphafold>,

<https://usegalaxy.eu/datasets/11ac94870d0bb33ad030384b9db286ef/display/?preview=True>



Predicted structure of a generated sequence (1273) using AlphaFold



5

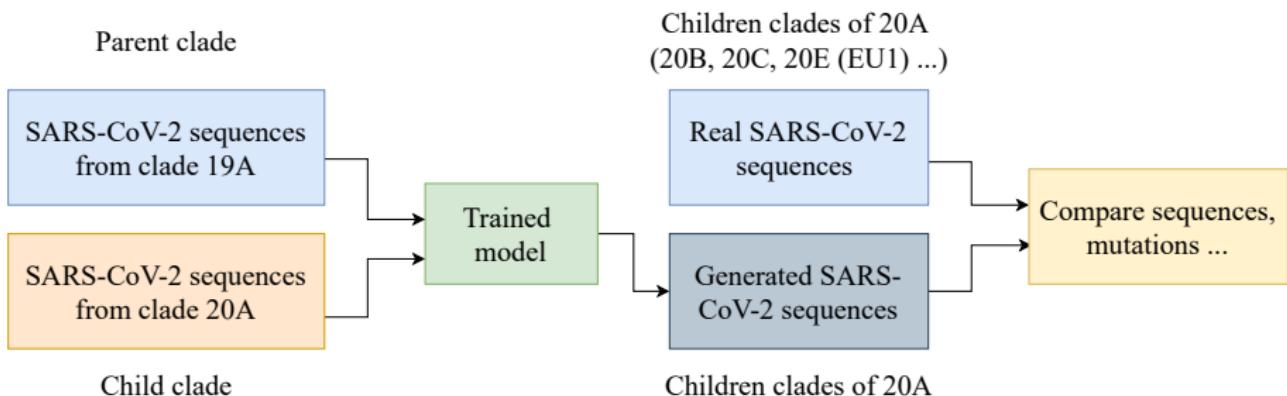
⁵<https://usegalaxy.eu/u/kumara/h/alphafold>,

<https://usegalaxy.eu/datasets/11ac94870d0bb33a03b9a0004708f93c/display/?preview=True>



Thank you for your attention. Questions?

Sequence to sequence learning with SARS-CoV-2 sequences

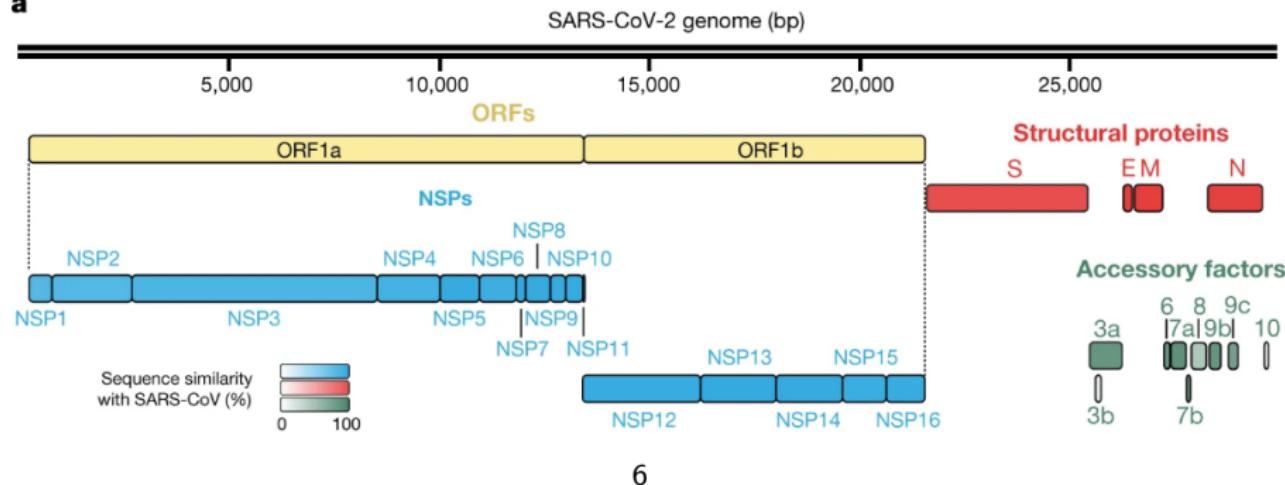


Training and prediction (generation) steps

- ① Create pairs of sequences (20A - 20B)
- ② Filter noisy pairs (Levenshtein dist > 10, sequence length > or < 1273)
- ③ 20,000 for training and 5,000 for test
- ④ Use trained model for generating sequences using test sequences from 20A
- ⑤ Use trained model for generating sequences using test sequences from 20B

Spike protein (S)

a



- Non-structural and structural proteins

⁶<https://www.nature.com/articles/s41586-020-2286-9>

Spike protein (S)

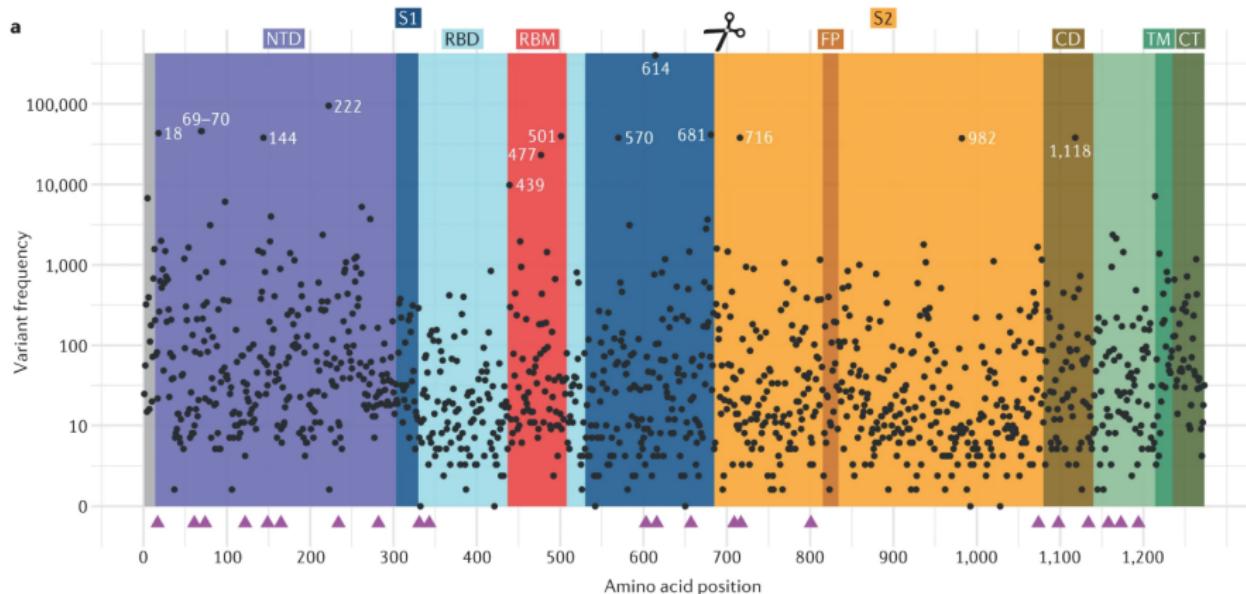
- Binds to the host cell
- Mutations may impact infectivity, transmissibility
- D614G: enhances viral replication ⁷
- N439K: enhances the binding affinity for the ACE2 receptor and reduces the neutralizing activity of antibodies ⁸
- Y453F: increased ACE2-binding affinity ⁹
- ...

⁷<https://www.nature.com/articles/s41586-020-2895-3>, <https://covariants.org/variants/20B.S.732A>

⁸<https://www.nature.com/articles/s41579-021-00573-0>

⁹<https://www.nature.com/articles/s41579-021-00573-0>

Frequency of spike mutations (substitutions and deletions)

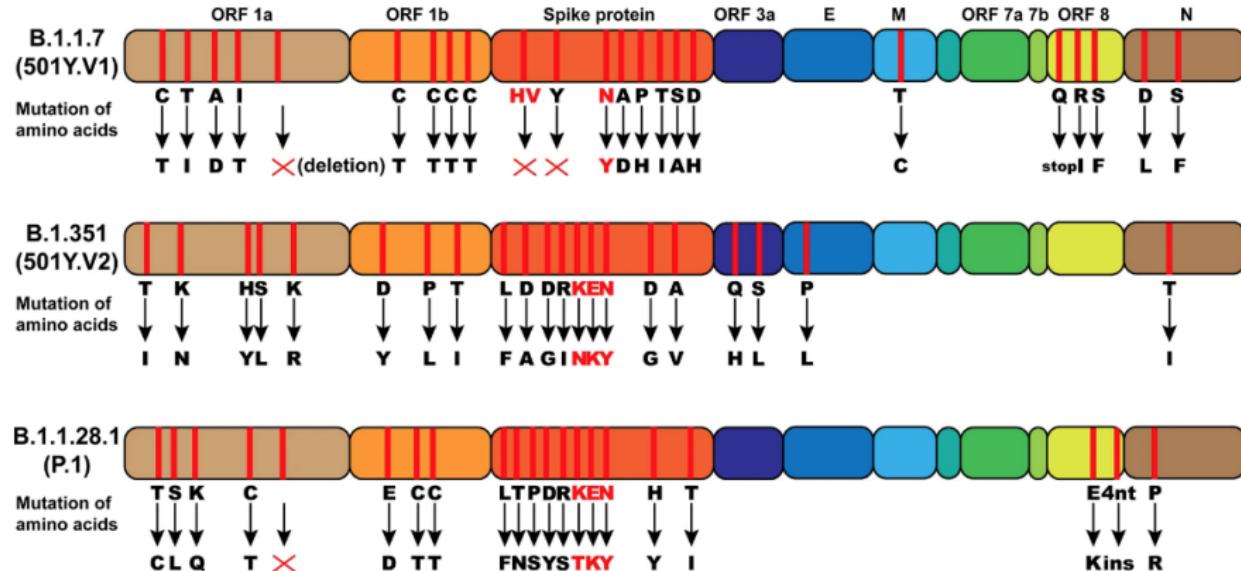


10

- 426,623 genomes, 5106 substitutions

¹⁰<https://www.nature.com/articles/s41579-021-00573-0>

Spike mutations in lineages



11

¹¹<https://www.nature.com/articles/s41392-021-00644-x>