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

Anup Kumar

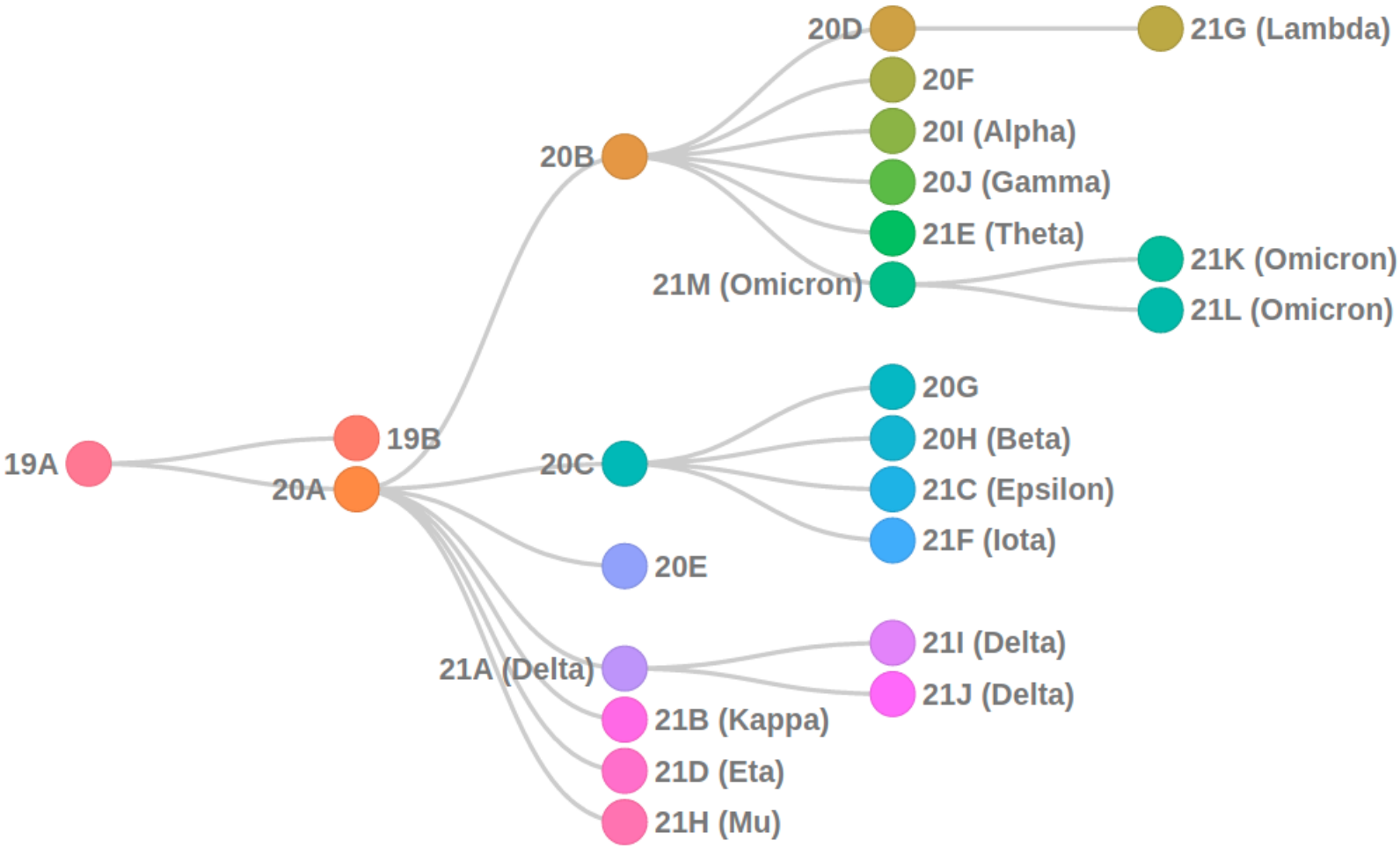
Bioinformatics group, University of Freiburg, Freiburg, Germany

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)

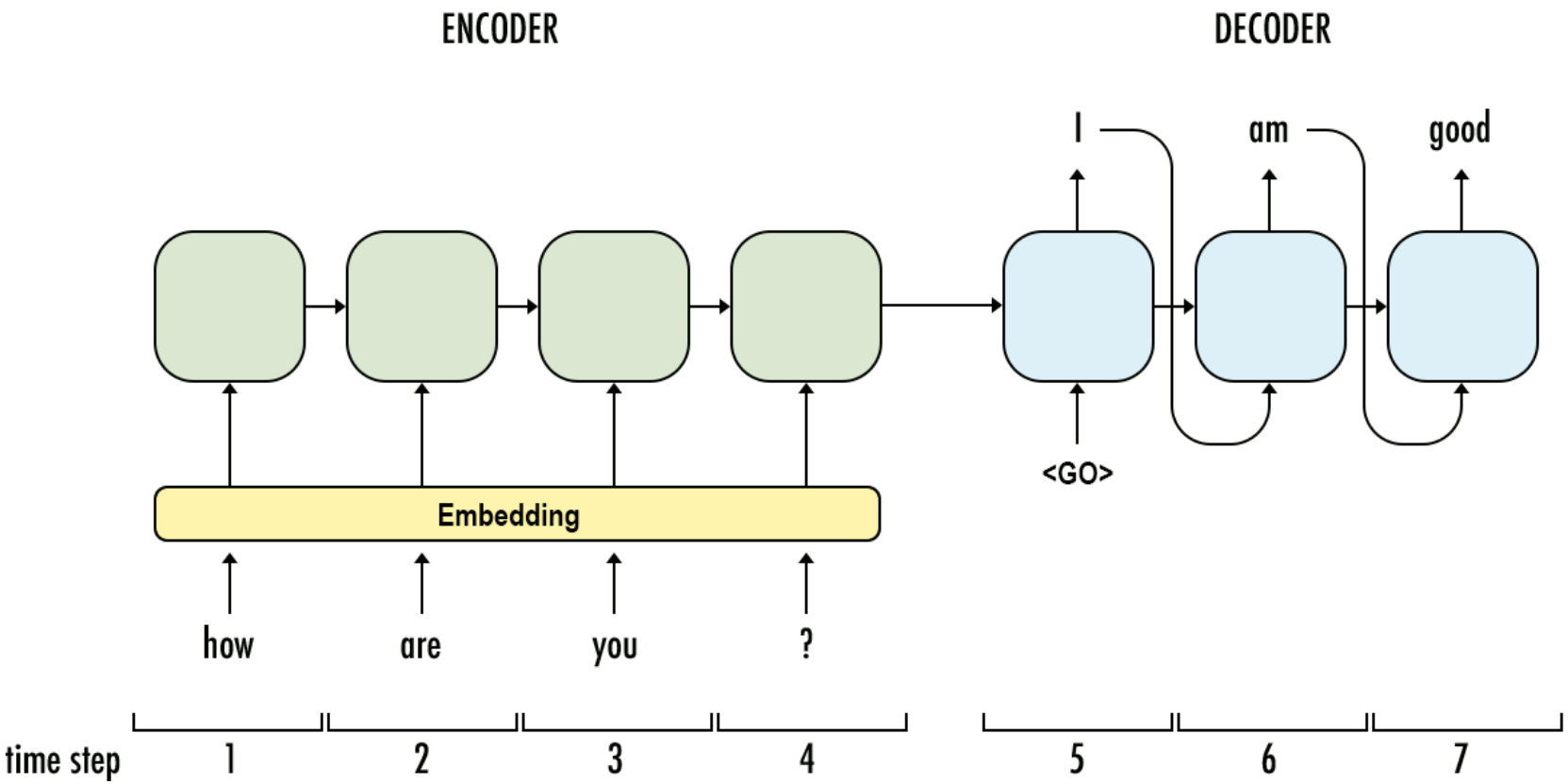
Phylogenetic tree of clades



1

[1] https://clades.nextstrain.org/

Sequence to sequence learning



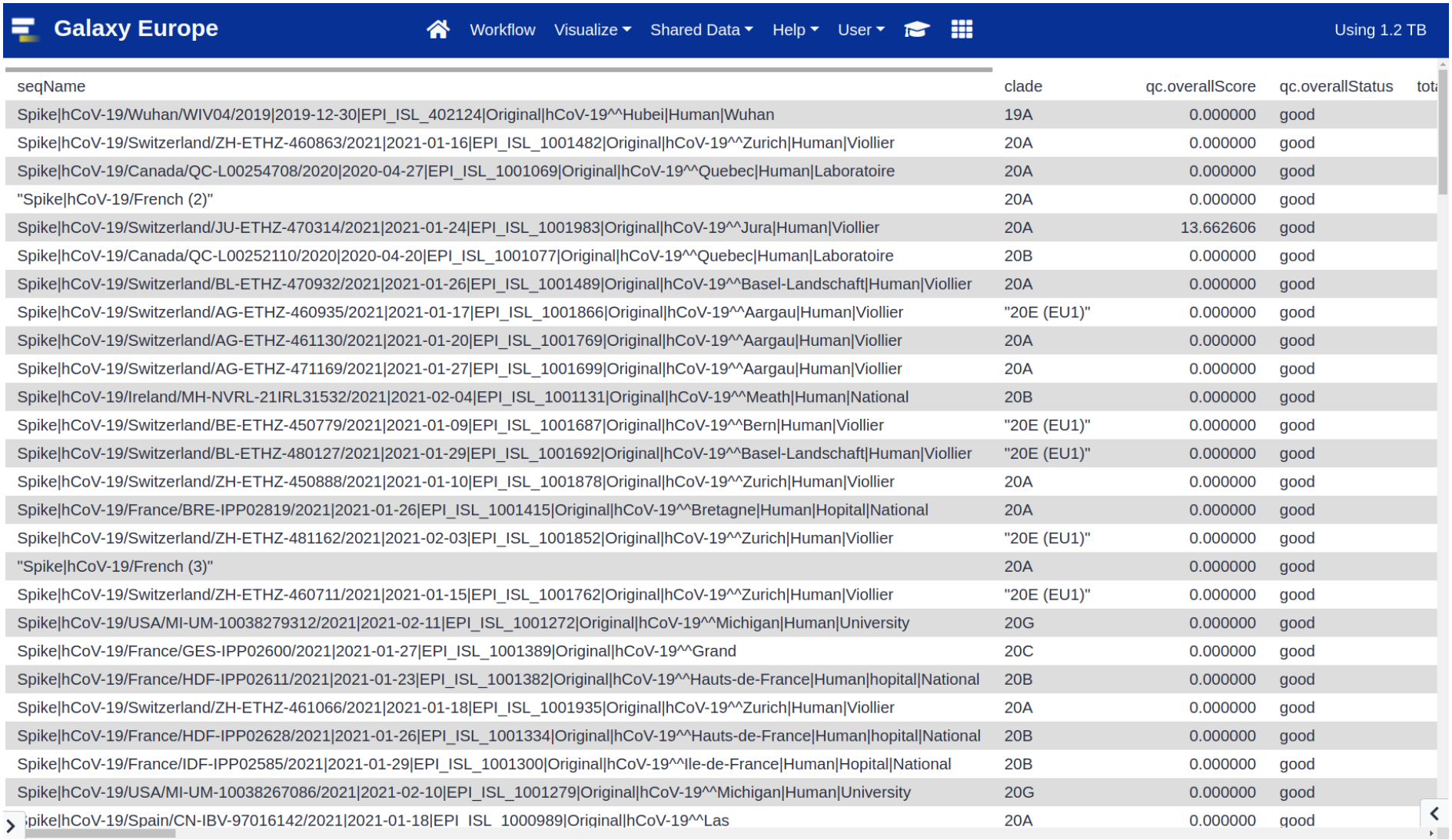
2

[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 ¿ sequence of kmers ¿ 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



seqName	clade	qc.overallScore	qc.overallStatus	tota
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-21IRL31532/2021|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|Hopital|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|hopital|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|hopital|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|Hopital|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	
spike|hCoV-19/Spain/CN-IBV-97016142/2021|2021-01-18|EPI_ISL_1000989|Original|hCoV-19^^Las	20A	0.000000	good	

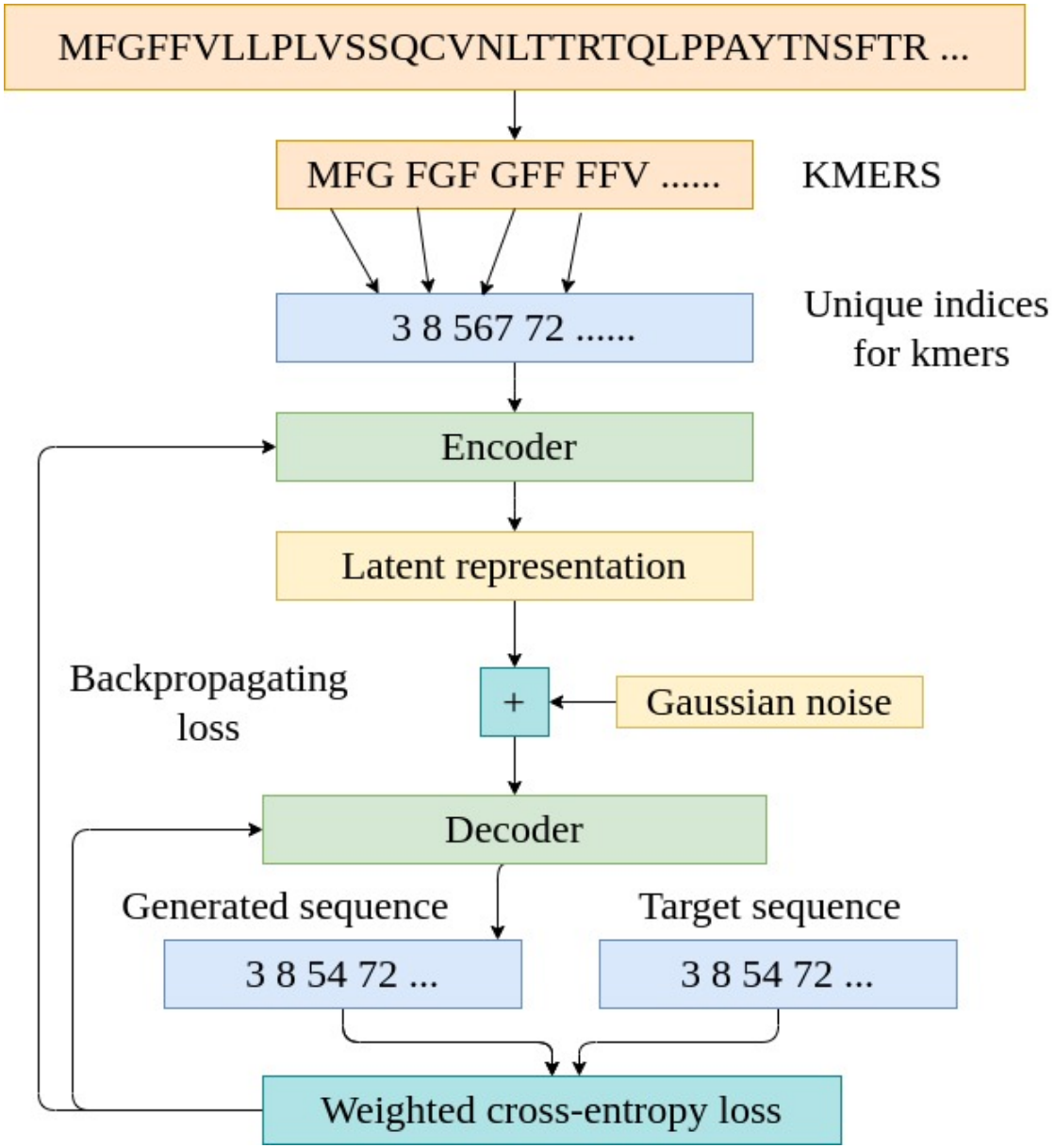
3

[3] https://usegalaxy.eu/

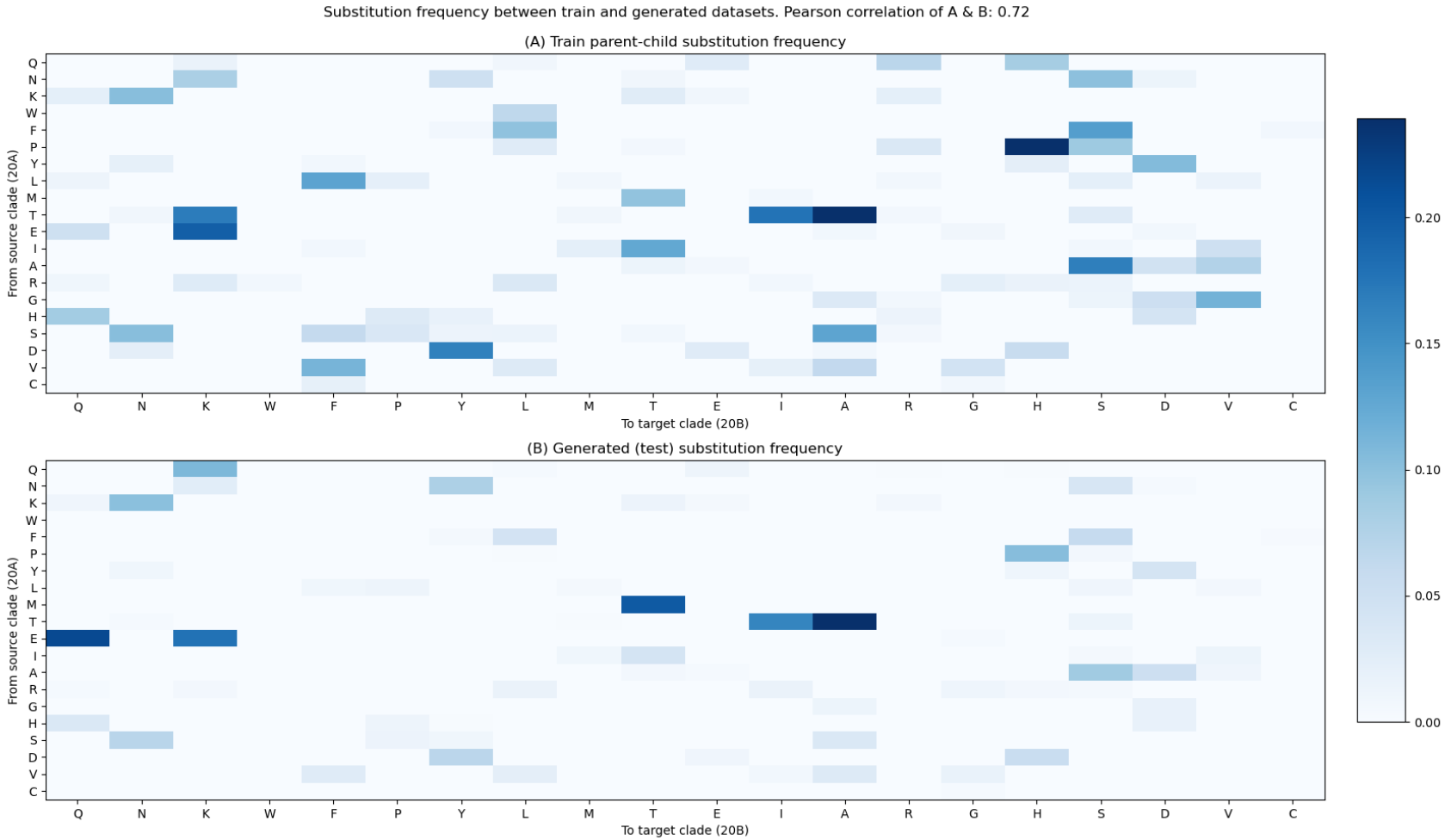
Source (20A) - target (20B) pairs

	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.	...	...

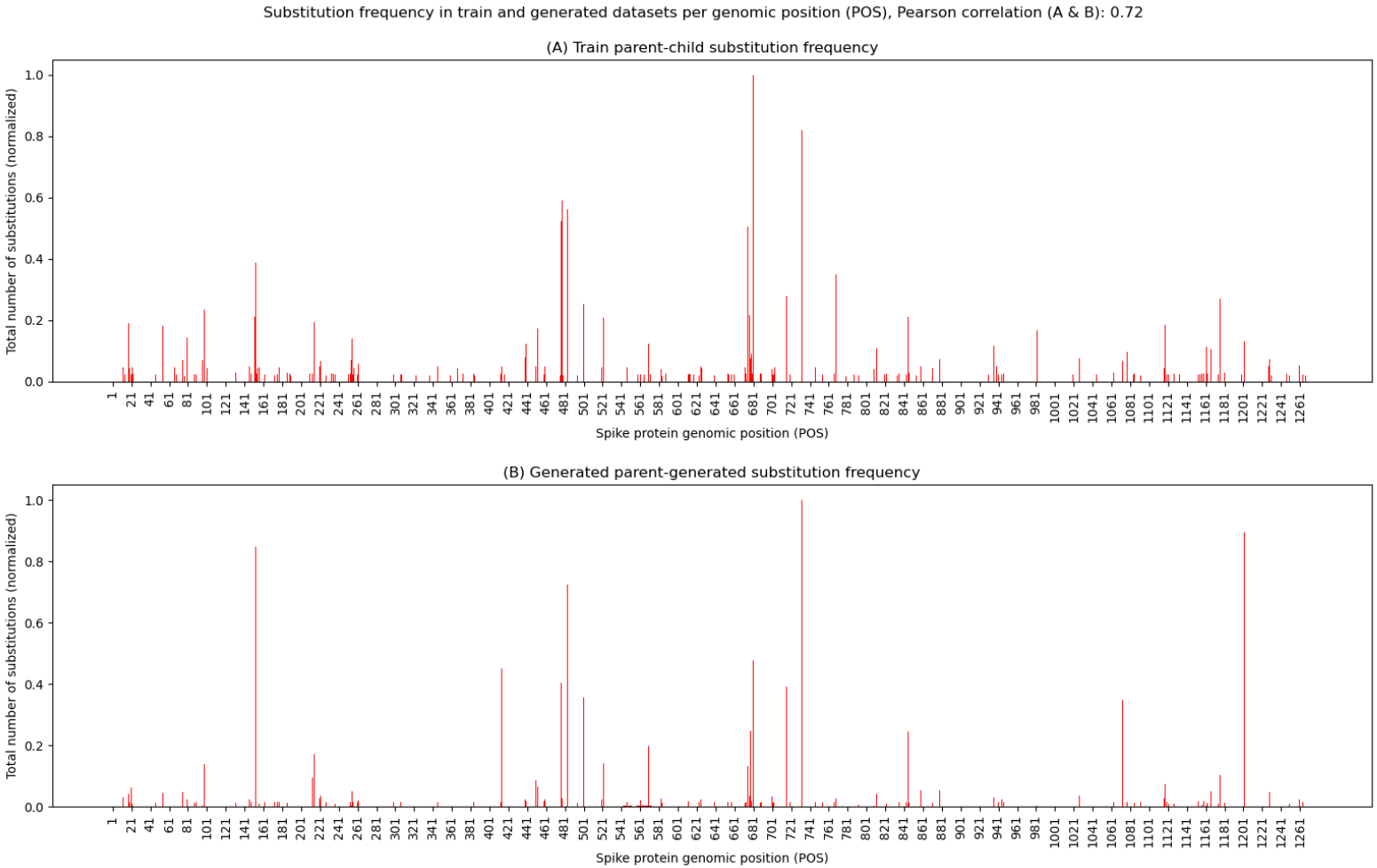
Encoder-decoder generative neural network



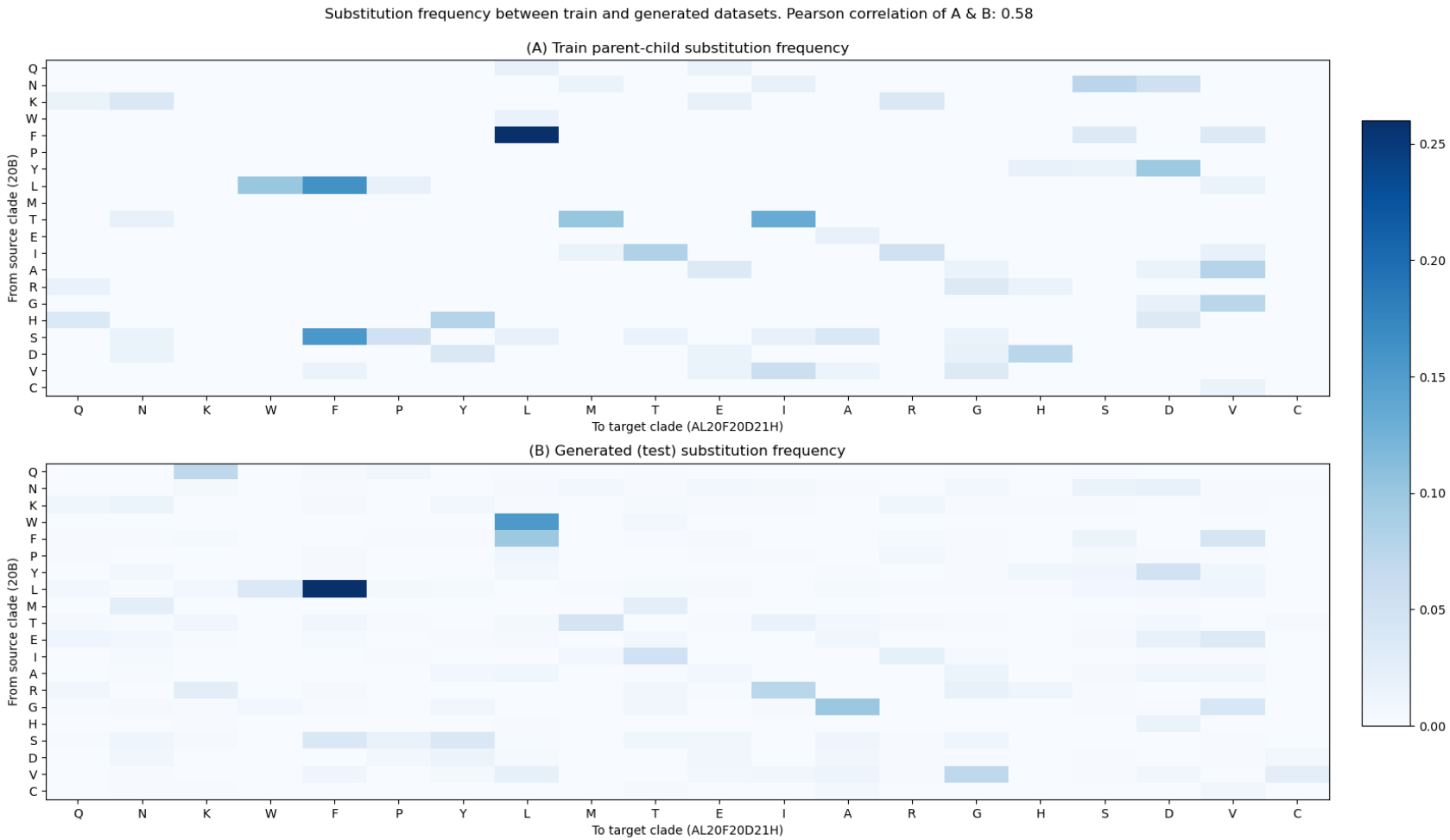
Comparison of protein evolution (for clade 20B using test sequences from 20A), Entire length (1273)



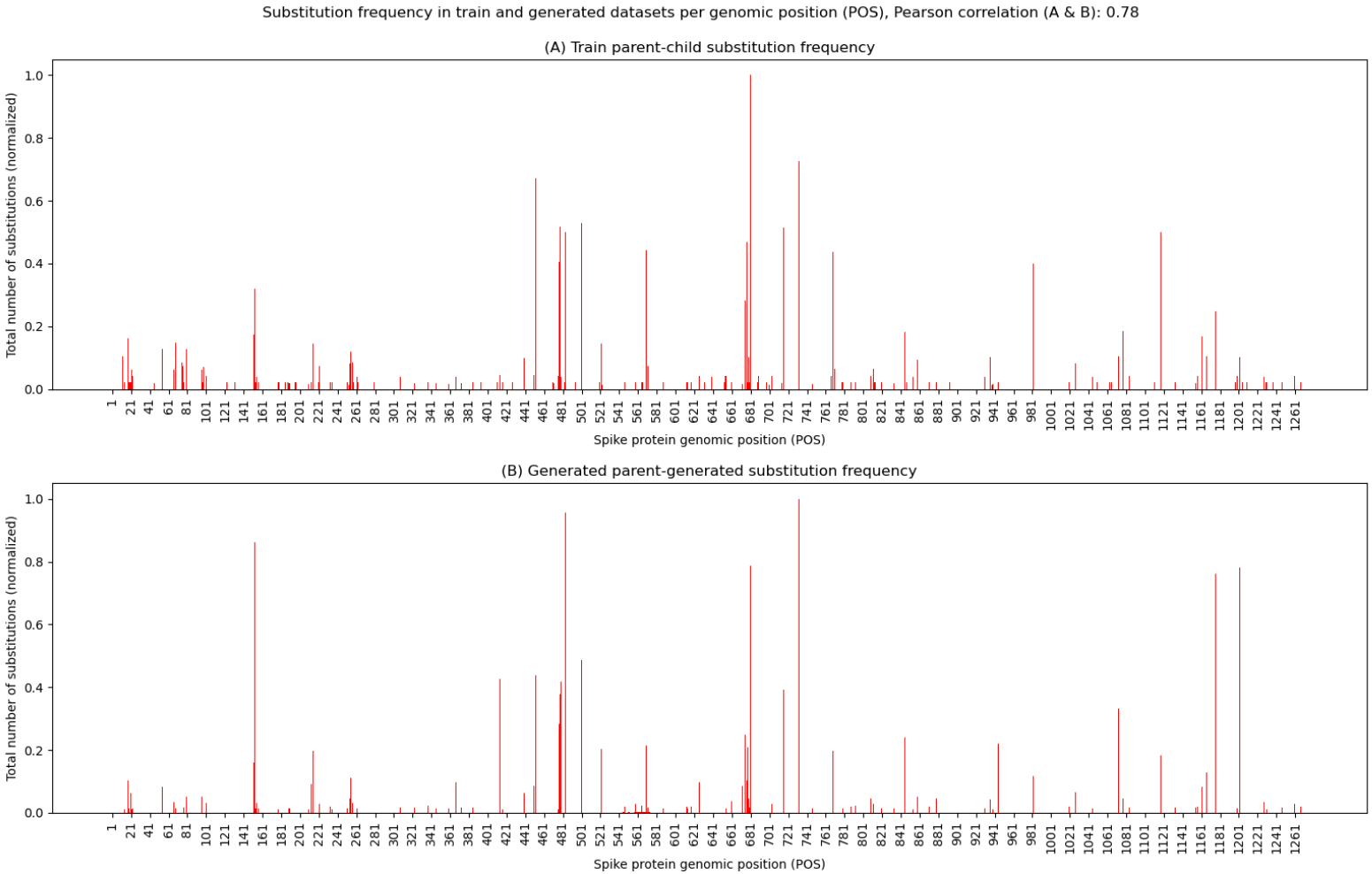
Frequency of substitution compared between training 20A-20B and test 20A-generated, Entire length (1273)



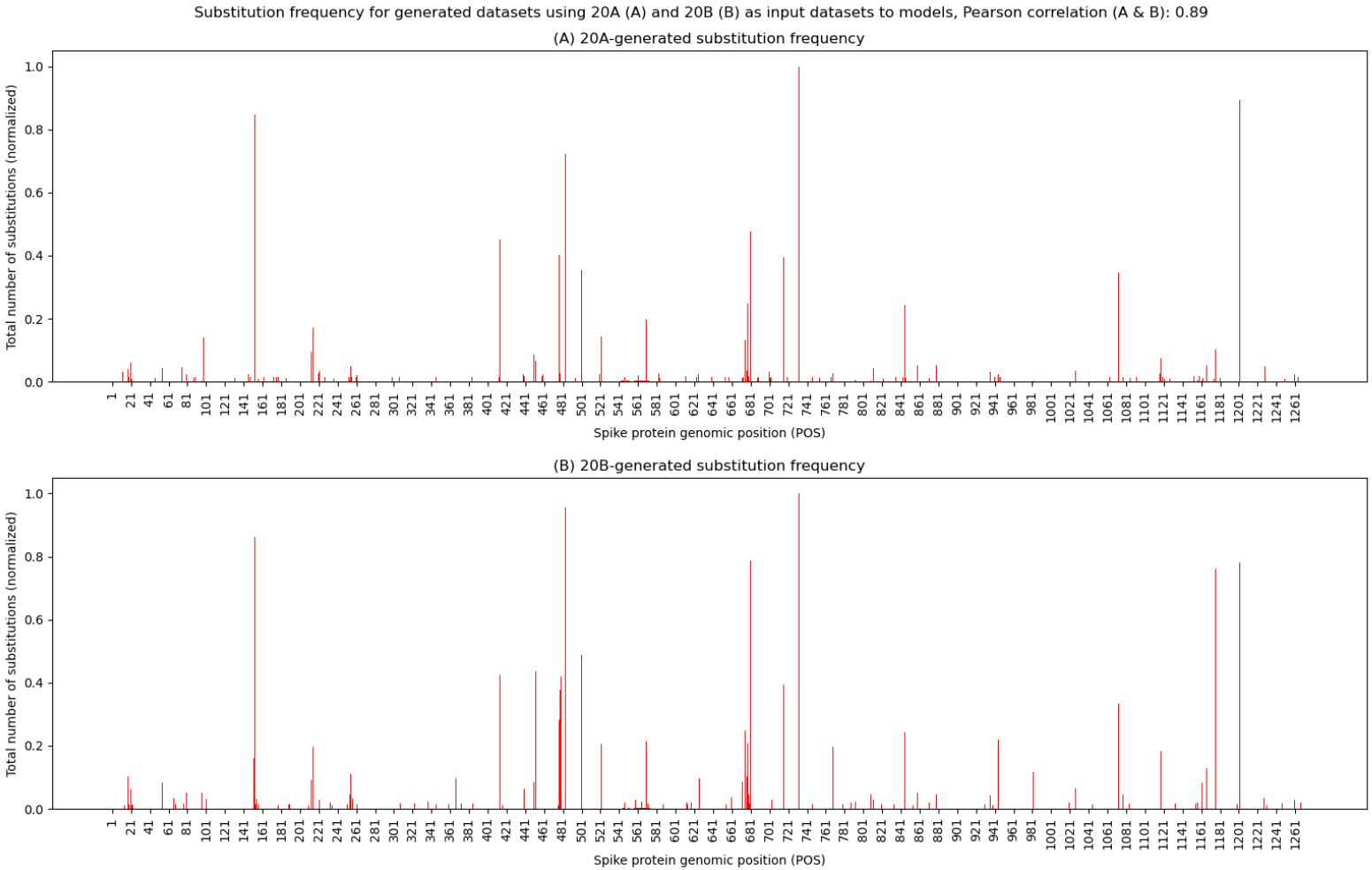
Comparison of protein evolution (for clades Alpha, Lambda, 20D, 20F and 21H using test sequences from 20B)



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



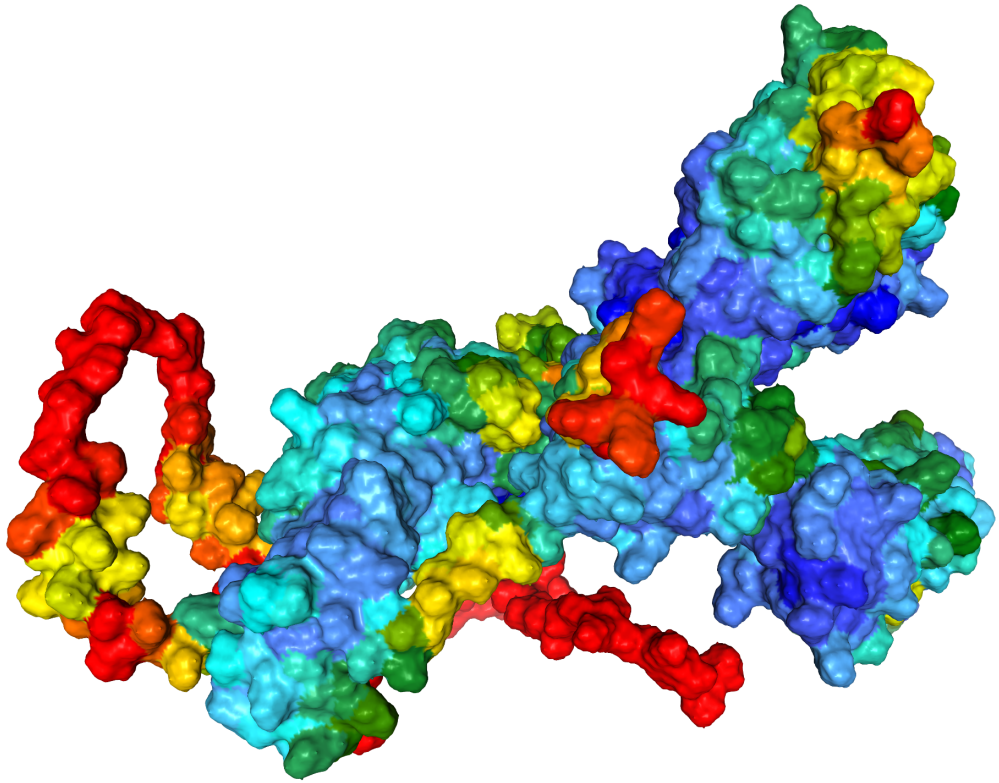
Substitution frequency between two generated datasets (from 20A and 20B) using same models



Links to tabular results

- Substitutions in training (20A-20B) and generated sequences (using 20A): https://github.com/anuprulez/clade_prediction/raw/unrolled_GAN/result_tabular_files/1273/compiled_excel_20A_20B.xlsx
- Substitutions in ground truth (20B - Alpha, Lambda, 20D, 21H, 20F) and generated sequences (using 20B): https://github.com/anuprulez/clade_prediction/raw/unrolled_GAN/result_tabular_files/1273/compiled_excel_20B_AL20F20D21H.xlsx
- Substitutions in 20B with respect to original reference: https://github.com/anuprulez/clade_prediction/blob/unrolled_GAN/result_tabular_files/1273/gen_wu_20B_gen_pos_subs.csv
- Substitutions in Alpha, Lambda, 20D, 21H, 20F with respect to original reference: https://github.com/anuprulez/clade_prediction/blob/unrolled_GAN/result_tabular_files/1273/gen_wu_AL20F20D21H_gen_pos_subs.csv

Predicted structure of a generated sequence (1273) using AlphaFold



4

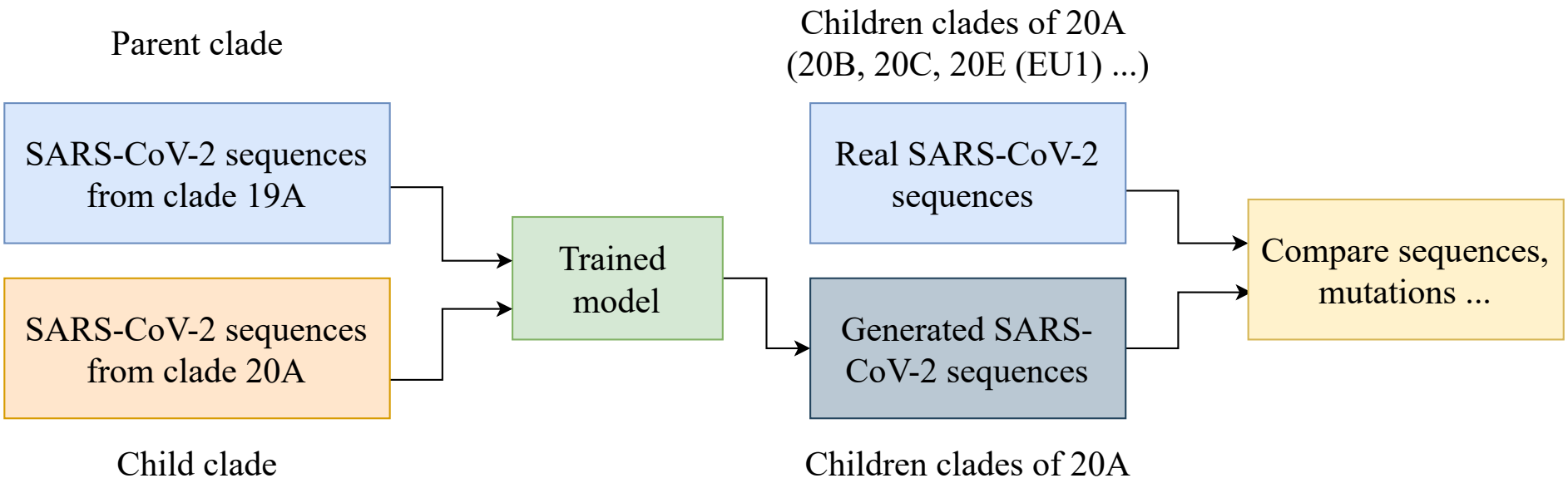
[4] https://usegalaxy.eu/u/kumara/h/alphafold, https://usegalaxy.eu/datasets/11ac94870d0bb33a03b9a0004708f93c/display/?preview=True

Future work

- Train on multiple combinations of clades and not just one branch. For example, 19A - 20A, 20A - 20B
- Train on one branch and predict on multiple branches including remote branches as well. For example, train on 20A - 20B and predict on 20C or 21A (Delta).

Thank you for your attention. Questions?

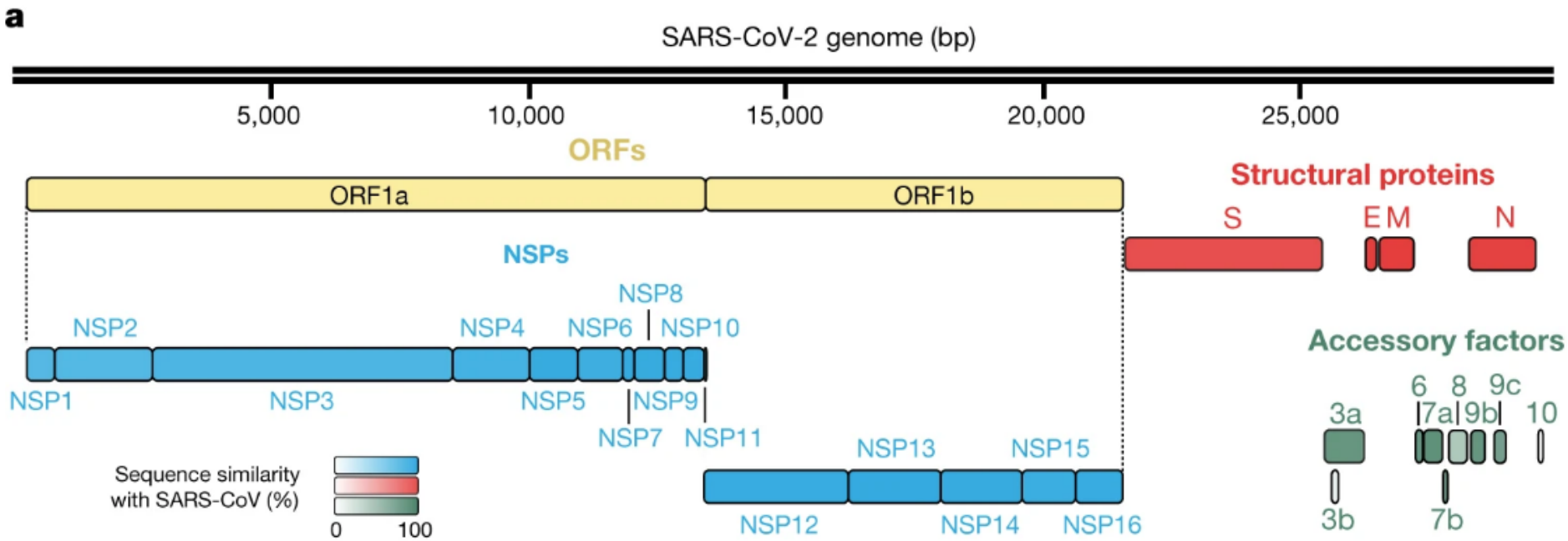
Sequence to sequence learning with SARS-CoV-2 sequences



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)



5

- Non-structural and structural proteins

[5] https://www.nature.com/articles/s41586-020-2286-9

Spike protein (S)

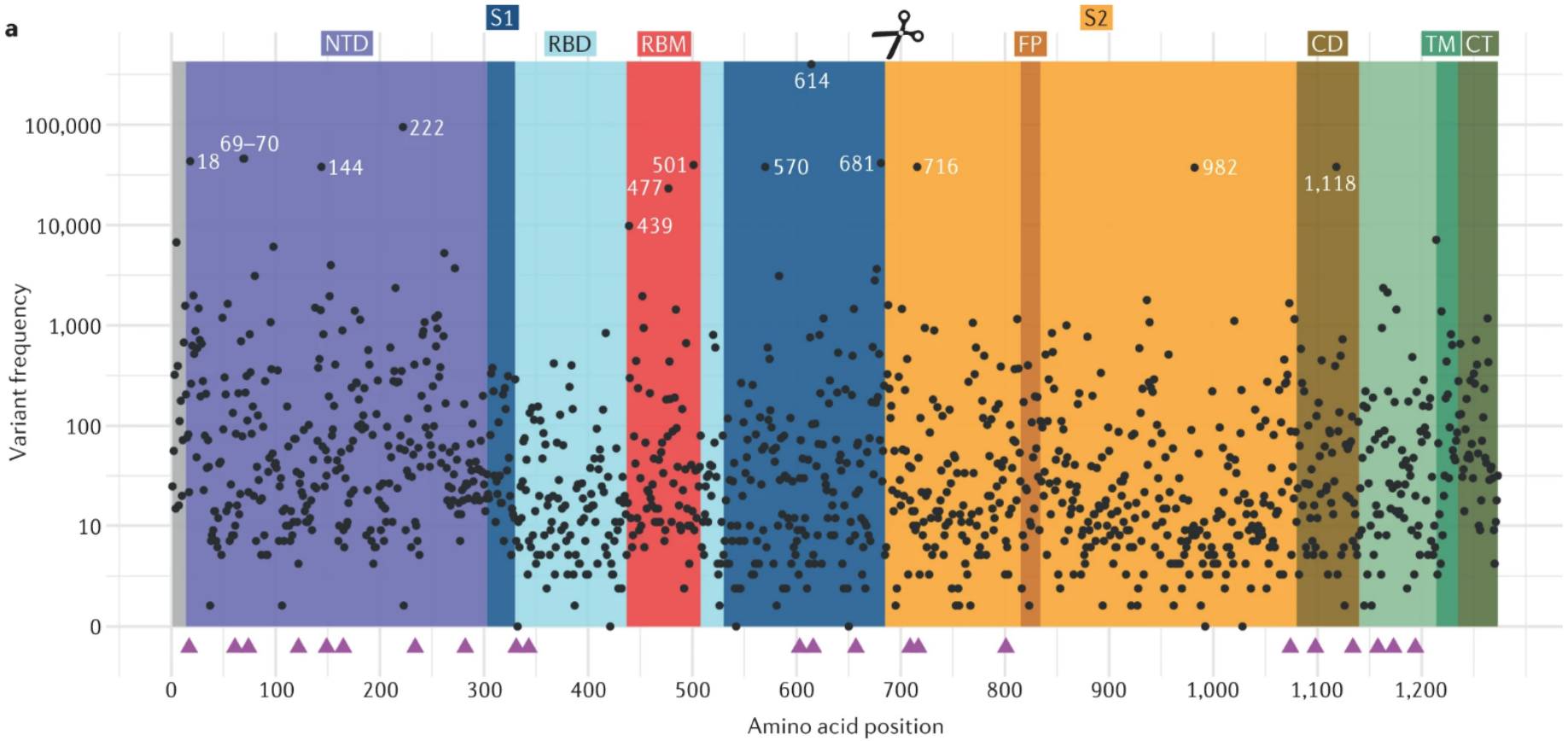
- Binds to the host cell
- Mutations may impact infectivity, transmissibility
- D614G: enhances viral replication [6]
- N439K: enhances the binding affinity for the ACE2 receptor and reduces the neutralizing activity of antibodies [7]
- Y453F: increased ACE2-binding affinity [8]
- ...

[6] https://www.nature.com/articles/s41586-020-2895-3, https://covariants.org/variants/20B.S.732A

[7] https://www.nature.com/articles/s41579-021-00573-0

[8] https://www.nature.com/articles/s41579-021-00573-0

Frequency of spike mutations (substitutions and deletions)

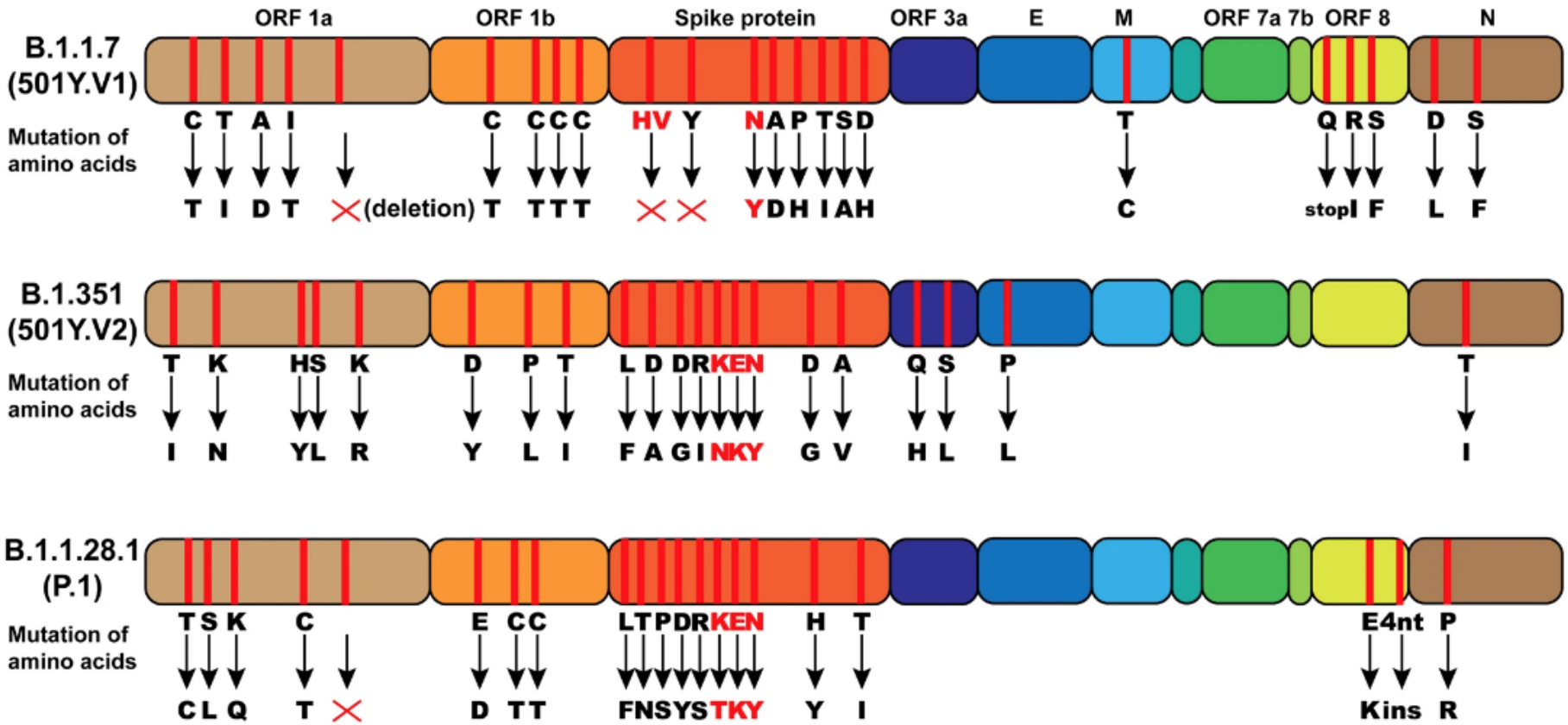


9

- 426,623 genomes, 5106 substitutions

[9] https://www.nature.com/articles/s41579-021-00573-0

Spike mutations in lineages



10

[10] https://www.nature.com/articles/s41392-021-00644-x