

# Modeling SARS-CoV-2 substitution processes: predicting the next variant

Keren Levinstein Hallak

Tel Aviv University

Saharon Rosset (✉ [saharon@post.tau.ac.il](mailto:saharon@post.tau.ac.il))

Tel Aviv University

---

## Article

**Keywords:** SARS-CoV-2, statistical modeling, variants

**Posted Date:** August 11th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-654547/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Communications Biology on March 29th, 2022. See the published version at <https://doi.org/10.1038/s42003-022-03198-y>.

<sup>1</sup> Modeling SARS-CoV-2 substitution processes:  
<sup>2</sup> predicting the next variant

<sup>3</sup> Keren Levinstein Hallak<sup>†</sup> Saharon Rosset<sup>\*†</sup>

<sup>4</sup> June 24, 2021

<sup>5</sup> **Abstract**

<sup>6</sup> We build statistical models to describe the substitution process  
<sup>7</sup> in the SARS-CoV-2 as a function of explanatory factors describing  
<sup>8</sup> the sequence, its function, and more. These models serve two differ-  
<sup>9</sup> ent purposes: first, to gain knowledge about the evolutionary biology  
<sup>10</sup> of the virus; and second, to predict future mutations in the virus,  
<sup>11</sup> in particular, non-synonymous amino acid substitutions creating new  
<sup>12</sup> variants. We use tens of thousands of publicly available SARS-CoV-2  
<sup>13</sup> sequences and consider tens of thousands of candidate models.

<sup>14</sup> Through a careful validation process, we confirm that our chosen  
<sup>15</sup> models are indeed able to predict new amino acid substitutions: can-  
<sup>16</sup> didates ranked high by our model are eight times more likely to occur  
<sup>17</sup> than random amino acid changes. We also show that named *variants*  
<sup>18</sup> *of interest* were highly ranked by our models before their appearance,  
<sup>19</sup> emphasizing the value of our models for identifying likely variants of  
<sup>20</sup> interest and potentially utilizing this knowledge in vaccine design and  
<sup>21</sup> other aspects of the ongoing battle against COVID-19.

<sup>22</sup> The intense community effort of SARS-CoV-2 sequencing has yielded a  
<sup>23</sup> wealth of information about the mutations that have occurred in the virus  
<sup>24</sup> since it first appeared in humans.

---

<sup>\*</sup>Corresponding author

<sup>†</sup>Department of Statistics and Operations Research, School of Mathematical Sciences,  
Tel-Aviv University, 6997801, Tel-Aviv, Israel

25      Understanding the evolutionary dynamics of the virus is critical for infer-  
26      ring its origin [32, 39], understanding its underlying biological mechanisms  
27      like mutagenic immune system responses [16, 29] and recombination [43, 4],  
28      predicting virus variants [5, 37, 7] and for vaccine and drug development  
29      [2, 11].

30      Recently there has been a spur of interest in analyzing substitution rates  
31      for SARS-CoV-2 [30, 10, 28]. Common analyses relate to explaining factors  
32      such as genes [22, 9, 13], CpG pairs [40, 31], context [12, 10] and codon  
33      and amino acid frequency [20, 17]. However, all previous work relied on a  
34      statistical analysis of the effect of each factor in isolation through summary  
35      statistics. If we seek to gain a deeper understanding and utility, we should  
36      consider these factors in tandem and aspire to build models that describe the  
37      entire mutation process as a function of all relevant information.

38      In this work, we employ regression in a big-data approach to identify  
39      the best statistical models for explaining the substitution rate distribution  
40      in observed sequences. We build a dataset containing 51,527 inferred substi-  
41      tutions for training the models based on a phylogenetic tree reconstruction  
42      from 61,835 available sequences [3] (as of February 8th, 2021).

43      We use the inferred substitutions in these sequences to identify the factors  
44      affecting substitution rates at different locations in the viral genome. We use  
45      our learned model to predict which sites in the genome are likely to mutate  
46      in the future and contribute to the formation of novel variants. Our methods  
47      can help vaccine design, medical research, and other tasks in the ongoing  
48      battle against COVID-19 and future viral epidemics.

49      We consider two different candidate phylogenetic trees: Tree of complete  
50      SARS-CoV-2 Sequences reconstructed by NCBI [3] and a phylogenetic tree  
51      we reconstructed by applying the sarscov2phylo method developed by Lan-  
52      fear [23] on the same sequences. Here we show results on the latter; we  
53      provide results for the NCBI phylogenetic tree in the supplementary.

54      In our models, we consider ten potential explanatory factors for explaining  
55      substitution rates based on sequence, biological function, gene location, and  
56      others. We compare 43,254 possible regression models and choose between  
57      them based on statistical goodness of fit scores.

58      We evaluate the ability of these models to predict new variants appearing  
59      in sequences that were added to the NCBI database between February 10th,  
60      2021 and April 10th, 2021 (“the test period”, see Figure 1). Our evaluation  
61      scheme does not depend on the correctness of the inferred tree or the family  
62      of regression models, thus objectively evaluating our models’ ability to rank

63 potential variants. For example, while the overall rate of occurrence of new  
64 amino acid substitutions in the test period was 2.2% among all candidate  
65 sites, the top 100 predictions of our selected model included 19 substitutions  
66 that actually occurred in the test period, for a lift (excess precision compared  
67 to random ranking) of 8.62.

## 68 Results

### 69 SARS-CoV-2 substitution model

70 We briefly describe our statistical modeling approach here; See Online Meth-  
71 ods for more details.

72 We inferred a phylogenetic tree and its mutations from the 44,080 se-  
73 quences that passed quality control (out of the 61,835 sequences available in  
74 the NCBI dataset as of 2/8/2021). We then built a training dataset describ-  
75 ing all potential substitutions in terms of the following explanatory factors:

- 76 1. Locus (Gene) of the site considered
- 77 2. Input nucleotide base (A/C/G/U)
- 78 3. Input amino acid
- 79 4. Input codon
- 80 5. The position of the site in the codon (1-3)
- 81 6. Mature peptide indicator
- 82 7. Stem loop indicator (different categorical values for each one of the  
83 stem loop genes ORF10 and ORF1ab)
- 84 8. CG pair indicator (different value for each position of the CG pair or  
85 NULL for non-CG)
- 86 9. Right neighboring nucleotide
- 87 10. Left neighboring nucleotide

We considered all possible combinations of using each factor in a generalized linear model (GLM) [24]: (−) omission, (+) as an explanatory factor, or (/) using it to split the GLM into sub-models such that a separate sub-model is built for each possible value. In our nomenclature, a *model* denotes a specific choice of inclusion (−, +, /) for each one of the categorical factors, and we fit the data the *sub-models* created by splitting according to the (/) factors. Subsequently, a total of 43,254 models were examined (each comprised of multiple sub-models). To account for over-dispersion, we considered a Negative-Binomial (NB) regression model in addition to the standard Poisson regression model in our GLM. All our models were fitted separately to synonymous and non-synonymous substitutions and accounted for the difference in rates between transitions and transversions.

Table 1 shows the top three NB, and Poisson regression models based on their AIC (penalized log-likelihood) score [1] on the training dataset. We provide all models in the supplementary material: TableS3.xlsb and TableS4.xlsb.

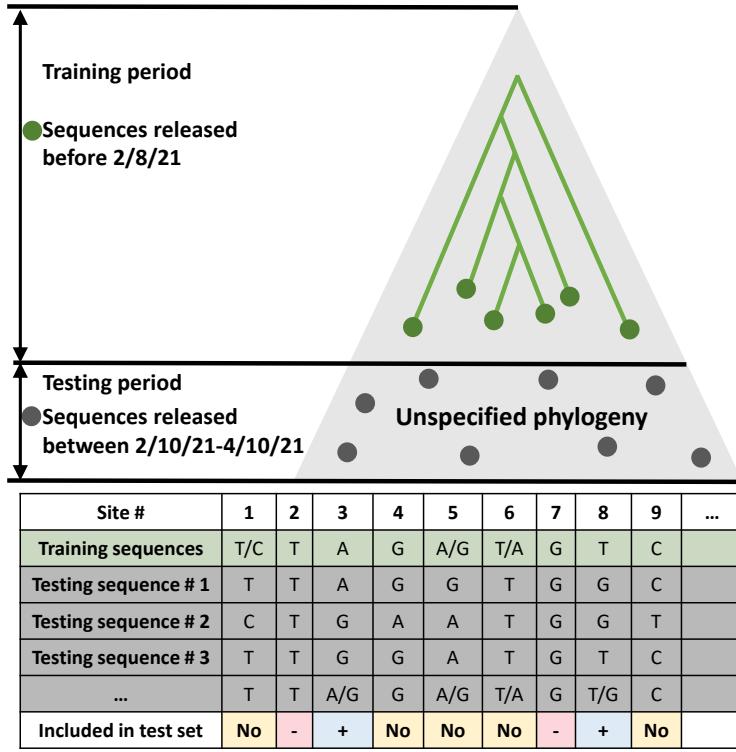
## Predictions

We next evaluated the ability of our top models to predict novel substitutions. Our prediction data set was constructed as follows. We considered the 32,495 *test* sequences that were added to the NCBI database in the period between February 10th, 2021, and April 10th, 2021. We then identified 9,696 sites with zero substitutions in the training data, i.e., identical or missing in all training data sequences. In these, we identified 2,697 sites that had at least one substitution in the test sequences. To avoid labeling sequencing errors, we required a minimum of two different test sequences with the mutated state; hence only 1,266 sites remained. Sites that had a single test sample with a mutated state were entirely ignored in the evaluation phase. For an illustration of the training and test datasets and our labeling procedure, see Figure 1.

We evaluated the ability of the top regression models to successfully rank the sites by their likelihood to mutate during the test period, thus creating new variants. Our evaluation is done at the amino acid level rather than the individual site (nucleotide) level, to express the notion that non-synonymous amino acid changes are the true object of interest in predicting new variants. The transition from predicting sites to predicting amino acids is done by careful post-processing and aggregation of the prediction model results

	Gene	Nucleotide	Amino Acid	Codon	Codon position	Mature Peptide	Stem Loop	CG Pair	Right Neighbor	Left Neighbor	# of Sub-Models
First models ranked by AIC	-	/	/	-	/	/	+	/	-	-	356
	-	/	/	-	/	/	+	/	+	+	356
	+	/	/	-	/	/	+	/	-	-	356
First models ranked by Poisson AIC	+	-	-	/	+	/	+	/	+	+	370
	+	/	/	-	/	/	+	/	+	+	356
	+	-	-	/	/	-	+	-	/	/	724

Table 1: **Top-scoring models for the training dataset.** The first three rows correspond to the top-scoring models when NB regression is applied. The next three rows correspond to the top-scoring models when Poisson regression is used. Each explaining factor is either (–) omitted from the model, (+) used as an explanatory factor, or (/) used to split the GLM into sub-models. We note that there are potential redundancies in the models. For example, the codon explaining factor contains the complete information on the amino acid and the nucleotide explaining factors (but not the other way). Our regression method of examining all inclusion possibilities for each factor considers this and produces a precise score regardless of the intertwined information.



**Figure 1: An illustration of the training and testing dataset for prediction.** Our training data consists of a phylogenetic tree reconstruction based on sequences released before February 8th, 2021 (green dots). The test data is comprised of sequences that were released between February 10th and April 10th, 2021 (gray dots). For these, we did not infer a phylogeny or rely on any other phylogenetic information. To evaluate our ability to predict new substitutions, we considered only sites for which no substitutions had occurred in the training data. The table in the figure shows examples of which substitutions are included in the test dataset. For sites 1, 5, and 6, the base is not constant for the training data set, and therefore it is not included in the test dataset. In sites 4 and 9, there is only one sequence in the test set that shows a different base from the training sequences; these sites have not been included in the test set to avoid sequencing errors. For sites 2 and 7, the base is constant for both the training and the test dataset making them negative examples in the test dataset, whereas sites 3 and 8 are positive examples, where a confirmed substitution occurred in the test period.

124 (see Online Methods). We used the area under the ROC curve (AUC) and  
125 the lift (ratio of true positives compared to a baseline model) to assess our  
126 results. The lift compares our model to two baselines: random ordering of  
127 all possible relevant substitutions and a *base model*, which takes into account  
128 *exposure*, i.e., the number of ways in which a specific amino acid can be  
129 created, and also the transition/transversion (ti/tv) ratio, but not the other  
130 explanatory factors. We compared to the base model as a sanity check that  
131 our models were indeed finding additional information to characterize amino  
132 acid substitution rates, beyond the exposure and ti/tv effect.

133 The results for our top models are shown in Table 2, both for the entire  
134 viral genome and the spike gene only, due to its biological importance [8]. For  
135 each model, we use both Poisson and Negative Binomial regressions for pre-  
136 dicting the substitution rate. Synonymous and non-synonymous substitution  
137 rates are modeled separately, due to the fundamentally different biological  
138 and evolutionary mechanisms they trigger. The community interest in non-  
139 synonymous substitutions also supports this separation [27, 19]. Note that  
140 the substitutions are aggregated per amino acid and location on the genome,  
141 as explained in the Online Methods.

142 Based on these results, we chose the third Poisson model of non-synonymous  
143 amino acid substitutions for a more detailed presentation here. The lift curves  
144 for this model are shown in Figure 2, demonstrating in more detail our mod-  
145 els' ability to identify likely substitutions. Note that in the test dataset, there  
146 are roughly 2% positives. Using the calculated lifts at 1%, the number of  
147 true positives is 7.51 times greater than the random model and 3.125 times  
148 greater than the base model. In numbers, this 1% represents 337 "candidate"  
149 substitutions, of which 50 actually occurred in the test period (compared to  
150 6.66 expected under the random model and 16 in the top base model pre-  
151 dictions). The lift curve against the base model is lower than that against  
152 the random model, yet still much higher than 1 for the highly ranked candi-  
153 dates (left side of the plot). This demonstrates that the exposure information  
154 used in the base model is essential for successful prediction, but the detailed  
155 models can still identify substantial signal beyond the exposure.

156 To help the community predict and analyze future substitutions, we pro-  
157 vide a complete list of predicted non-synonymous amino acid substitution  
158 rates in the spike protein in the supplementary. In addition, we note for each  
159 substitution whether or not it was observed in the training and test datasets  
160 (Supplementary TableS5.xlsb).

161 As an additional demonstration of our models' success in ranking amino

Model #	Non-synonymous amino acid substitutions						Synonymous amino acid substitutions						
	Poisson			Negative Binomial			Poisson			Negative Binomial			
	AUC	3% Lift Vs.		AUC	3% Lift Vs.		AUC	3% Lift Vs.		AUC	3% Lift Vs.		
		Random model	Base model		Random model	Base model		Random model	Base model		Random model	Base model	
All genes	1	0.835	4.707	2.238	0.821	4.607	1.957	0.858	3.577	1.465	0.856	3.577	1.432
	2	0.832	4.406	2.095	0.819	4.306	1.830	0.861	3.861	1.581	0.858	3.463	1.386
	3	0.836	5.358	2.548	0.826	4.557	1.936	0.847	3.520	1.442	0.846	3.690	1.477
Spike gene	1	0.814	4.062	2.667	0.786	2.538	1.250	0.867	4.748	3.333	0.861	1.899	1.333
	2	0.814	4.062	2.667	0.781	3.554	1.750	0.864	4.273	3.000	0.859	3.798	2.667
	3	0.830	4.062	2.667	0.827	3.554	1.750	0.863	4.748	3.333	0.864	4.748	3.333

Table 2: **Prediction results for the top three models.** We use the top three Poisson and Negative Binomial models from Table 1 for prediction on the test dataset. Results for the entire genome are in the first three rows, for the spike protein only in the last three. Results are shown separately for predicting amino acid substitutions (left half) and predicting synonymous substitutions (right half, these results are not discussed in the text). The first column in each quarter of the table shows the area under the ROC curve (AUC) for the corresponding prediction task and modeling approach. We highlighted the top-scoring model for every (substitution type, locus, approach) combination. Overall we obtained high AUC scores, showing the models successfully predicted many of the substitutions. The second and third columns in each quarter are 3% lift scores of each model versus the random model and the more elaborate base model (see text and Online Methods). The top models significantly outperform both baselines stressing the benefits of our approach over more naive statistical predictions. The model we analyzed further in the text (third Poisson model for non-synonymous amino acid substitutions) is also red-framed.

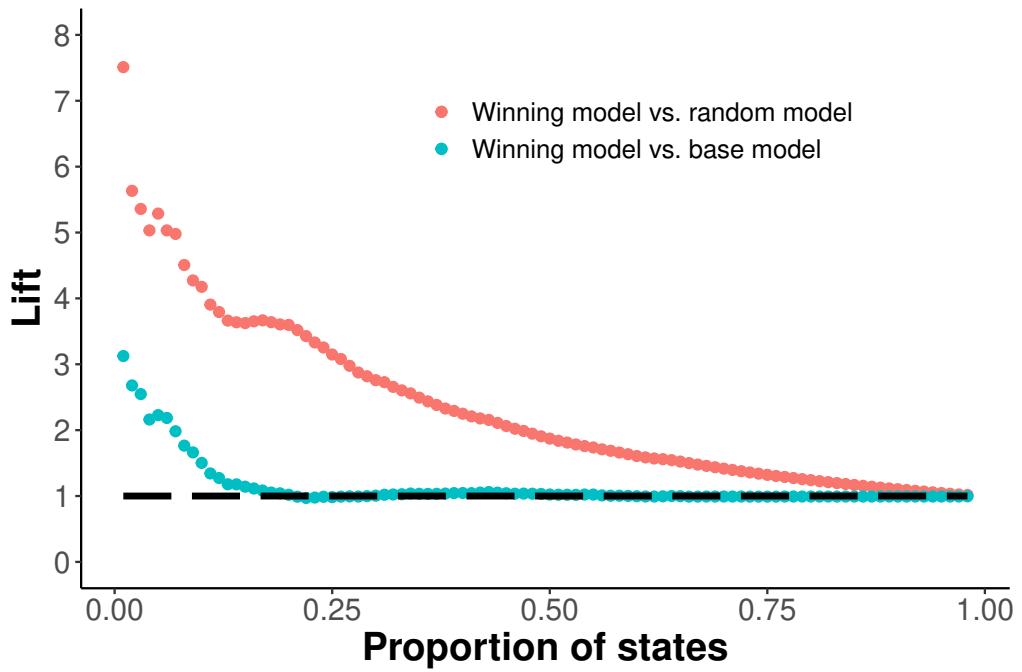


Figure 2: Lift curves of the winning model versus the random (cyan) and base models (red).

	67H	79I	95I	950N	701V	67V	477N	154K	484K	142D	80G	253G	19R
Rank	15	143	163	300	590	604	613	803	821	915	1230	1282	1456
	859N	478K	950H	565L	888I	1176F	681R	1071H	484Q	158G	157S	614G	452R
Rank	1470	1479	1847	1969	1973	2400	2698	2735	3402	4443	4643	6424	7390

Table 3: **Rank of spike protein amino acid substitutions.** Ranking was performed by our prediction model on 13,544 possible non-synonymous amino acid substitutions in the spike protein resulting from one nucleotide change. The ranks of the 26 *variants of interest* defined by the CDC are shown. The highlighted substitutions were not part of the training dataset.

acid substitutions of interest, we analyzed the substitutions comprising the variants defined by the Centers for Disease Control and Prevention (CDC) [15] as variants of interest, listed as 26 amino acid substitutions in the spike protein. Of these, 23 were included in our training data, while 3 were recorded after our training cutoff date of 2/8/2021. We examined their ranking according to our chosen model (third ranked Poisson model) in the list of all 13,544 possible spike protein amino acid substitutions. Results are given in Table 3, demonstrating that 69% of them (18/26, including all three substitutions not observed in training) were ranked in the top 2,000 (that is, top 15% of predictions) according to our model.

## Discussion

In this work, we model substitution rates in the SARS-CoV-2 as a function of several possible affecting factors describing sequence and coding information. We fit our models to *training* data that is based on inferring the phylogenetic tree connecting tens of thousands of sequences collected before February 2021 and also inferring the specific substitutions that have occurred on this tree. This phylogenetic reconstruction task is extremely challenging, and it is unlikely that the inferred tree or substitutions are completely accurate [28]. This is also evident by the different trees, substitutions, and slightly different models we get when we use the *sarscov2phylo* method [23] to reconstruct the tree, with results given in the main text, compared to using NCBI’s

183 reconstruction of the tree (results in supplementary).

184 However, a critical point is that our evaluation approach on the *test set*  
185 of sequences added after the training cutoff date does not rely on any phylo-  
186 genetic reconstruction or assumptions on the phylogenetic context between  
187 the test sequences and training sequences (as illustrated in Figure 1). The  
188 fact that the test set shows high AUC and lift curves demonstrates that  
189 regardless of doubts about the accuracy of the training phylogenetic recon-  
190 struction, the models we fit to the training data are indeed useful to predict  
191 future substitutions.

192 The specific substitutions we include in the test set were carefully chosen  
193 to avoid sequencing errors and phylogenetic uncertainty in the evaluation.  
194 However, we emphasize that our models can be used to predict the likelihood  
195 of all possible substitutions and variants, including ones that have already  
196 appeared in the training data (as we did in our analysis of known variants in  
197 Table 3). Furthermore, the nucleotide level predictions we generate can be  
198 easily transformed into amino acid level predictions, as we did in our actual  
199 evaluation and AUC and lift calculations (with the methodology described  
200 in Online Methods). This is critical since the discussion of variants in the  
201 literature is typically focused on the amino acid level [33, 34].

202 Our top regression models shown in Table 1 suggest that all of the factors  
203 we consider are potentially useful for predicting future substitutions and  
204 variants, but some are more important than others. Specifically, most of the  
205 best models split into sub-models by amino acid rather than by codon (as  
206 shown by their designation as / in all top models according to NB AIC),  
207 suggesting that codon usage bias effects such as those described in [27, 13]  
208 may not be major.

209 An important property of our regression approach is that regression mod-  
210 els consider all candidate explanatory factors at once. They are thus able  
211 to identify factors that appear essential when considered on their own but  
212 whose effect can be *explained away* by other, better factors. For instance,  
213 the neighboring nucleotides identities (context) seem to have a minor role  
214 once the amino acid and codon position are taken into account and are not  
215 included at all in some of our top models (as indicated by their designation  
216 as – in two of the top three models). While it is true that in an analysis  
217 examining only the connection between neighbors and likelihood of substi-  
218 tution, the context would appear very significant (results not shown), this  
219 effect is mitigated and may disappear when taking into account the better  
220 factors.

221 In summary, our statistical modeling approach offers two significant ben-  
222 efits: A better understanding and modeling of the factors affecting substi-  
223 tution rates in the SARS-CoV-2 virus, and by implication in other viruses; and  
224 the resulting predictive models, which can be used to rank future variants by  
225 their likelihood. We hope and expect that both of these contributions will  
226 serve the scientific and medical communities in the ongoing battle against  
227 the COVID-19 epidemic caused by this virus.

## 228 **Code availability**

229 The code used in this work is available at:  
230 <https://github.com/Kerenlh/sarscov2predictions.git>

## 231 **Funding**

232 This work was supported in part by a fellowship from the Edmond J. Safra  
233 Center for Bioinformatics at Tel-Aviv University.

## 234 **Online Methods**

### 235 **Phylogeny of SARS-CoV-2**

236 The sequences used in this work were all downloaded from the NCBI website<sup>1</sup>  
237 [3]. As a training set, we used 61,835 available sequences as of February 8th,  
238 2021. For a test set, we used 32,495 sequences released between February  
239 10th, 2021, and April 10th, 2021. We used two phylogenetic reconstructions  
240 of SARS-CoV-2 following related works in the literature [10, 36, 26]:

- 241 1. The tree of complete SARS-CoV-2 Sequences by NCBI <sup>2</sup>.
- 242 2. A tree reconstructed by us using the sarscov2phylo method developed  
243 by Lanfear <sup>3</sup> [23].

244 NCBI's tree and the sarscov2phylo method exclude noisy sequences. These  
245 include low quality sequences and sequences missing sufficient data so that

---

<sup>1</sup><https://www.ncbi.nlm.nih.gov/sars-cov-2/>

<sup>2</sup><https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/precomptree>

<sup>3</sup><https://github.com/roblanf/sarscov2phylo>

246 it is hard to place them meaningfully in the phylogeny. We used the global  
247 sequence alignment method implemented in the sarscov2phylo method which  
248 aligns every sequence to the reference sequence (accession *NC\_045512.2*) from  
249 NCBI and then joins the individually aligned sequences into a global alignment  
250 using MAFFT v7.471 [21], faSplit <sup>4</sup>, faSomeRecords <sup>5</sup> and GUPparallel  
251 [35].

## 252 Internal Nodes Reconstruction

253 The internal nodes of the tree phylogeny are necessary to infer the substitutions  
254 that occurred on the tree edges. We now describe our heuristic, inspired  
255 by Fitch's algorithm [14], used to reconstruct the sequences in the internal  
256 nodes.

257 Every site holds a probability vector over the bases A/C/G/U defined as  
258 follows:

- 259 1. For every leaf, assign probability 1 to the base in the respective site  
260 and probability 0 to all other bases. Whenever there is base ambiguity,  
261 the probability is split uniformly among the possible bases.
- 262 2. **Pass from bottom to top.** The probability vector of an internal  
263 node is the average of the probability vectors of its children.
- 264 3. **Pass from top to bottom.** We descend the tree from the root and  
265 add to each node  $\epsilon = 1/(\# \text{ of children})$  multiplied by its parent's prob-  
266 ability vector (and normalize by  $1 + \epsilon$  to keep it in the  $l_1$ -simplex).
- 267 4. The chosen base at every node is determined by the highest probability  
268 value. This procedure also solves ambiguous sites in the leaves.

269 By doing this, we break ties between the highest probabilities (such ties are  
270 frequent) and allow information to flow between nodes that have a common  
271 ancestor.

272 Finally, we applied a battery of statistical tests to validate the phylo-  
273 genetic tree and its internal nodes. For example, multiple back mutations  
274 might imply that the internal node reconstruction is faulty, so we examined  
275 the number of back mutations in the two phylogenetic trees. In the tree

---

<sup>4</sup>(<http://hgdownload.soe.ucsc.edu/admin/exe/>)

<sup>5</sup>(<https://github.com/ENCODE-DCC/kentUtils>)

276 reconstructed according to Lanfear’s method, there were no back mutations,  
277 while in the tree reconstructed by NCBI, there were only two with no obvious  
278 alternative (examined manually).

## 279 Substitution Model

280 By reconstructing the tree’s internal nodes, we can generate a tabular dataset  
281 consisting of the list of factors and the number of substitutions that occurred  
282 for each instantiation of these factors. We use the multiple regression ap-  
283 proach described in [25] which considers for every factor in the tabular data  
284 the options to either join in the regression linearly (marked  $\textcolor{blue}{+}$ ), not join at  
285 all (marked  $\textcolor{red}{-}$ ), or to partition the data according to it (marked  $\textcolor{green}{/}$ ). We use  
286 the term model to denote a specific choice of inclusion for each categorical  
287 factor that might affect the substitution rate as listed.

288 A partitioning ( $\textcolor{green}{/}$ ) splits the regression model into multiple smaller re-  
289 gressions, where each factor gets one of its values. Consider, for example,  
290 that there are only two factors, the base, and the codon position. If both  
291 are ( $\textcolor{blue}{+}$ ), then only one regression will be applied with a one-hot encoding of  
292 both factors. However, if the base is ( $\textcolor{green}{/}$ ), we will use four regression models to  
293 partition the data according to the base (A/C/G/U). We use the term sub-  
294 model for each of the actual models fitted after splitting. The AIC [1] score  
295 is given by  $AIC = 2k - 2 \log(\hat{L})$  where  $k$  is the number of free parameters  
296 and  $\hat{L}$  is the maximum likelihood. Then, the AIC scores of these sub-models  
297 are summed up to form one unified score for this model.

298 Consequently, the number of models we consider is, in theory, combina-  
299 torial in the number of values each factor can have. However, the number  
300 of models can be substantially reduced since some factors are dependent on  
301 one another (for example, the codon determines the amino acid and base).  
302 In our data, we score 43,254 models. We apply both Poisson regression and  
303 Negative-Binomial regression [18] for each model, where the latter is used  
304 to account for overdispersion, specifically to account for latent factors not  
305 included in the model. The complete list of factors is given in the main  
306 paper. Finally, our experiments infer different regression coefficients for syn-  
307 onymous and non-synonymous sub-models and combine the AIC scores. We  
308 also considered doing the same for transitions/transversions and different  
309 output nucleotides, but we got strictly worse AIC scores.

310 Another critical notion is that of *exposure* [6], which weights the states  
311 we train on according to the frequency of their occurrence. For instance, a

312 specific combination of frequently appearing factors in the dataset has rela-  
313 tively higher exposure than a rare set. When we learn the regression model,  
314 taking exposure into account is crucial to reduce bias in the dataset and im-  
315 prove the predictions. The exposure is proportional to the total amount of  
316 time a specific set of factors was observed. To calculate that duration, we  
317 summarize the lengths of relevant branches in the phylogenetic tree and use  
318 the sum as an offset variable in the regression.

Finally, we apply additional normalization. We first define the non-synonymous ti/tv ratio [42]:

$$r_{ti:tv}^{non-syn} = \frac{\# \text{ Non-synonymous transitions}}{\# \text{ Non-synonymous transversions}}$$

319 in the training data. Then, we count the number of possible transitions and  
320 transversions per state for each state and normalize the substitution rate  
321 accordingly. For example, the codon GCG in the first codon position has  
322 one possible non-synonymous transition and two possible non-synonymous  
323 transversions. The non-synonymous substitution rate for that state is hence  
324 normalized by  $1 + 2/r_{ti:tv}^{non-syn}$ . An identical procedure is applied to the syn-  
325 onymous substitutions.

## 326 Prediction

327 Our main prediction task is focused on predicting amino acid substitutions.  
328 As our basic predictions are always at the single nucleotide level, we care-  
329 fully aggregate them to form amino acid predictions – the substitution rate  
330 of an amino acid output at a given location is the sum of the rates of all  
331 the substitutions leading to it. Note that in most but not all cases, there  
332 is only a simple correspondence, in that there is a single non-synonymous  
333 nucleotide substitution that leads to a given amino acid change. However,  
334 more complex settings can occur, such as the substitution from Histidine to  
335 Glutamine through four different non-synonymous transversions in the third  
336 codon position.

337 To test the performance of our predictions, we compare them to two base-  
338 lines. The first baseline is the *random model* which places equal probability  
339 on all amino acid substitutions. While a naive random model would consider  
340 all 21 amino acids per location, we permit only one substitution per codon  
341 since multiple substitutions per codon are highly unlikely (less than 0.5% of  
342 the substitutions occurred at adjacent sites in the same tree branch). This

343 limitation drastically improves the random model’s predictions and reduces  
344 possible amino acid substitutions throughout the molecule from 121,653 to  
345 33,684.

346 The second baseline model is called *base model*. This model takes into  
347 account the exposure and ti/tv normalization for each substitution and uses it  
348 for prediction. Hence it is a lot less naive than the random model and relies  
349 on careful evaluation of the different likelihood for different substitutions  
350 based on the observed states in the tree and the ti/tv effect. It differs from  
351 our “true” prediction models in ignoring the ten potential affecting factors,  
352 and comparing to it is our way to quantify the contribution of these factors  
353 to predictive power within our regression approach.

354 To compare the top models to the baseline models, we use two scoring  
355 methods – AUC and lift (we emphasize here again that all comparisons are  
356 made on data in the test period not used for building the models, as explained  
357 in Figure 1 of the main text). First, we transform the predicted substitution  
358 rate into a binary prediction vector of 0/1 predictions. We do this by applying  
359 a threshold on the predicted substitution rate where all rates above a specific  
360 value are deemed positive. By varying the threshold, we can derive the  
361 ROC curve (using the test dataset as the ground truth), from which we can  
362 calculate the AUC score. Lift [41, 38] measures how well a targeting model  
363 performs at predicting compared to a random choice method. We compute  
364 the lift for each threshold by taking the ratio of “precision at x%” between  
365 our model and each baseline model separately.

## 366 References

- 367 [1] H. Akaike. A new look at the statistical model identification. *IEEE*  
368 *Trans. Automat. Contr.*, 19(6):716–723, dec 1974.
- 369 [2] Fatima Amanat and Florian Krammer. Sars-cov-2 vaccines: status re-  
370 port. *Immunity*, 52(4):583–589, 2020.
- 371 [3] DA Benson, M Cavanaugh, K Clark, I Karsch-Mizrachi, DJ Lipman,  
372 J Ostell, and EW Sayers. Genbank nucleic acids res 41 (d1). *D36–D42*,  
373 2013.
- 374 [4] Maciej F Boni, Philippe Lemey, Xiaowei Jiang, Tommy Tsan-Yuk Lam,  
375 Blair W Perry, Todd A Castoe, Andrew Rambaut, and David L Robert-

- 376 son. Evolutionary origins of the sars-cov-2 sarbecovirus lineage respon-  
377 sible for the covid-19 pandemic. *Nature Microbiology*, 5(11):1408–1417,  
378 2020.
- 379 [5] Rachele Cagliani, Diego Forni, Mario Clerici, and Manuela Sironi. Com-  
380 putational inference of selection underlying the evolution of the novel  
381 coronavirus, severe acute respiratory syndrome coronavirus 2. *Journal  
382 of virology*, 94(12):e00411–20, 2020.
- 383 [6] Harvey Checkoway, Neil Pearce, and David Kriebel. *Research methods  
384 in occupational epidemiology*, volume 34. Monographs in Epidemiology  
385 and, 2004.
- 386 [7] Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations  
387 strengthened sars-cov-2 infectivity. *Journal of molecular biology*,  
388 432(19):5212–5226, 2020.
- 389 [8] Xiangyang Chi, Renhong Yan, Jun Zhang, Guanying Zhang, Yuanyuan  
390 Zhang, Meng Hao, Zhe Zhang, Pengfei Fan, Yunzhu Dong, Yilong Yang,  
391 et al. A neutralizing human antibody binds to the n-terminal domain  
392 of the spike protein of sars-cov-2. *Science*, 369(6504):650–655, 2020.
- 393 [9] Marti Cortey, Yanli Li, Ivan Diaz, Hepzibar Clilverd, Laila Darwich,  
394 and Enric Mateu. Sars-cov-2 amino acid substitutions widely spread in  
395 the human population are mainly located in highly conserved segments  
396 of the structural proteins. *bioRxiv*, 2020.
- 397 [10] Nicola De Maio, Conor R Walker, Yatish Turakhia, Robert Lanfear,  
398 Russell Corbett-Detig, and Nick Goldman. Mutation rates and selection  
399 on synonymous mutations in sars-cov-2. *Genome Biology and Evolution*,  
400 13(5):evab087, 2021.
- 401 [11] Bethany Dearlove, Eric Lewitus, Hongjun Bai, Yifan Li, Daniel B  
402 Reeves, M Gordon Joyce, Paul T Scott, Mihret F Amare, Sandhya  
403 Vasan, Nelson L Michael, et al. A sars-cov-2 vaccine candidate would  
404 likely match all currently circulating variants. *Proceedings of the Na-  
405 tional Academy of Sciences*, 117(38):23652–23662, 2020.
- 406 [12] Salvatore Di Giorgio, Filippo Martignano, Maria Gabriella Torcia, Gior-  
407 gio Mattiuz, and Silvestro G Conticello. Evidence for host-dependent

- 408 rna editing in the transcriptome of sars-cov-2. *Science Advances*,  
409 6(25):eabb5813, 2020.
- 410 [13] Maddalena Dilucca, Sergio Forcelloni, Alexandros G Georgakilas, An-  
411 drea Giansanti, and Athanasia Pavlopoulou. Codon usage and pheno-  
412 typic divergences of sars-cov-2 genes. *Viruses*, 12(5):498, 2020.
- 413 [14] Walter M Fitch. Toward defining the course of evolution: minimum  
414 change for a specific tree topology. *Systematic Biology*, 20(4):406–416,  
415 1971.
- 416 [15] Centers for Disease Control and Prevention.  
417 <https://www.cdc.gov/coronavirus/2019-ncov/variants/variant->  
418 info.html, 2021.
- 419 [16] Alex Graudenzi, Davide Maspero, Fabrizio Angaroni, Rocco Piazza, and  
420 Daniele Ramazzotti. Mutational signatures and heterogeneous host re-  
421 sponse revealed via large-scale characterization of sars-cov-2 genomic  
422 diversity. *Iscience*, 24(2):102116, 2021.
- 423 [17] Haogao Gu, Daniel KW Chu, Malik Peiris, and Leo LM Poon. Multivariate  
424 analyses of codon usage of sars-cov-2 and other betacoronaviruses.  
425 *Virus evolution*, 6(1):veaa032, 2020.
- 426 [18] Joseph M Hilbe. *Negative binomial regression*. Cambridge University  
427 Press, 2011.
- 428 [19] Elio Issa, Georgi Merhi, Balig Panossian, Tamara Salloum, and Sima  
429 Tokajian. Sars-cov-2 and orf3a: nonsynonymous mutations, functional  
430 domains, and viral pathogenesis. *Msystems*, 5(3):e00266–20, 2020.
- 431 [20] Mahmoud Kandeel, Abdelazim Ibrahim, Mahmoud Fayez, and Mo-  
432 hammed Al-Nazawi. From sars and mers covs to sars-cov-2: Moving  
433 toward more biased codon usage in viral structural and nonstructural  
434 genes. *Journal of medical virology*, 92(6):660–666, 2020.
- 435 [21] Kazutaka Katoh and Daron M. Standley. MAFFT multiple sequence  
436 alignment software version 7: Improvements in performance and usabil-  
437 ity. *Mol. Biol. Evol.*, 30(4):772–780, apr 2013.

- 438 [22] Neha Kaushal, Yogita Gupta, Mehendi Goyal, Svetlana F Khaiboullina,  
439 Manoj Baranwal, and Subhash C Verma. Mutational frequencies of  
440 sars-cov-2 genome during the beginning months of the outbreak in usa.  
441 *Pathogens*, 9(7):565, 2020.
- 442 [23] R. Lanfear. <https://github.com/roblanf/sarscov2phylo>, 2021.
- 443 [24] Keren Levinstein Hallak, Shay Tzur, and Saharon Rosset. Big data  
444 analysis of human mitochondrial DNA substitution models: a regression  
445 approach. *BMC Genomics*, 19(1):759, dec 2018.
- 446 [25] Keren Levinstein Hallak, Shay Tzur, and Saharon Rosset. Big data  
447 analysis of human mitochondrial DNA substitution models: a regression  
448 approach. *BMC Genomics*, 19(1):759, dec 2018.
- 449 [26] Tingting Li, Dongxia Liu, Yadi Yang, Jiali Guo, Yujie Feng, Xinmo  
450 Zhang, Shilong Cheng, and Jie Feng. Phylogenetic supertree reveals  
451 detailed evolution of sars-cov-2. *Scientific reports*, 10(1):1–9, 2020.
- 452 [27] Yashpal Singh Malik, Mohd Ikram Ansari, Jobin Jose Kattoor, Rahul  
453 Kaushik, Shubhankar Sircar, Anbazhagan Subbaiyan, Ruchi Tiwari,  
454 Kuldeep Dhama, Souvik Ghosh, Shailly Tomar, et al. Evolutionary  
455 and codon usage preference insights into spike glycoprotein of sars-cov-  
456 2. *Briefings in bioinformatics*, 22(2):1006–1022, 2021.
- 457 [28] Benoit Morel, Pierre Barbera, Lucas Czech, Ben Bettisworth, Lukas  
458 Hübner, Sarah Lutteropp, Dora Serdari, Evangelia-Georgia Kostaki,  
459 Ioannis Mamais, Alexey M Kozlov, et al. Phylogenetic analysis of sars-  
460 cov-2 data is difficult. *Molecular biology and evolution*, 38(5):1777–1791,  
461 2021.
- 462 [29] Tobias Mourier, Mukhtar Sadykov, Michael J Carr, Gabriel Gonzalez,  
463 William W Hall, and Arnab Pain. Host-directed editing of the sars-cov-2  
464 genome. *Biochemical and biophysical research communications*, 2020.
- 465 [30] Matías J Pereson, Laura Mojsiejczuk, Alfredo P Martínez, Diego M  
466 Flichman, Gabriel H Garcia, and Federico A Di Lello. Phylogenetic  
467 analysis of sars-cov-2 in the first few months since its emergence. *Journal  
468 of medical virology*, 93(3):1722–1731, 2021.

- 469 [31] Mukhtar Sadykov, Tobias Mourier, Qingtian Guan, and Arnab Pain.  
470 Short sequence motif dynamics in the sars-cov-2 genome suggest a role  
471 for cytosine deamination in cpg reduction. *BioRxiv*, 2020.
- 472 [32] Muhammad Adnan Shereen, Suliman Khan, Abeer Kazmi, Nadia  
473 Bashir, and Rabeea Siddique. Covid-19 infection: Origin, transmis-  
474 sion, and characteristics of human coronaviruses. *Journal of advanced*  
475 *research*, 24:91–98, 2020.
- 476 [33] Joshua Singer, Robert Gifford, Matthew Cotten, and David Robertson.  
477 Cov-glue: a web application for tracking sars-cov-2 genomic variation.  
478 *Preprints*, 2020.
- 479 [34] Julian W Tang, Paul A Tambyah, and David SC Hui. Emergence of a  
480 new sars-cov-2 variant in the uk. *The Journal of infection*, 2020.
- 481 [35] Ole Tange et al. Gnu parallel-the command-line power tool. *The*  
482 *USENIX Magazine*, 36(1):42–47, 2011.
- 483 [36] Yatish Turakhia, Bryan Thornlow, Angie S Hinrichs, Nicola De Maio,  
484 Landen Gozashti, Robert Lanfear, David Haussler, and Russell Corbett-  
485 Detig. Ultrafast sample placement on existing trees (usher) enables real-  
486 time phylogenetics for the sars-cov-2 pandemic. *Nature Genetics*, pages  
487 1–8, 2021.
- 488 [37] Lucy van Dorp, Mislav Acman, Damien Richard, Liam P Shaw, Char-  
489 lotte E Ford, Louise Ormond, Christopher J Owen, Juanita Pang,  
490 Cedric CS Tan, Florencia AT Boshier, et al. Emergence of genomic  
491 diversity and recurrent mutations in sars-cov-2. *Infection, Genetics and*  
492 *Evolution*, 83:104351, 2020.
- 493 [38] Miha Vuk and Tomaz Curk. Roc curve, lift chart and calibration plot.  
494 *Metodoloski zvezki*, 3(1):89, 2006.
- 495 [39] Hongru Wang, Lenore Pipes, and Rasmus Nielsen. Synonymous muta-  
496 tions and the molecular evolution of sars-cov-2 origins. *Virus evolution*,  
497 7(1):veaa098, 2021.
- 498 [40] Yong Wang, Jun-Ming Mao, Guang-Dong Wang, Zhi-Peng Luo, Liu  
499 Yang, Qin Yao, and Ke-Ping Chen. Human sars-cov-2 has evolved to

- 500 reduce cg dinucleotide in its open reading frames. *Scientific Reports*,  
501 10(1):1–10, 2020.
- 502 [41] Ian H Witten and Eibe Frank. Data mining: practical machine learning  
503 tools and techniques with java implementations. *Acm Sigmod Record*,  
504 31(1):76–77, 2002.
- 505 [42] Ziheng Yang and Anne D Yoder. Estimation of the transition/  
506 transversion rate bias and species sampling. *Journal of Molecular  
507 Evolution*, 48(3):274–283, 1999.
- 508 [43] Zhao Zhang, Libing Shen, and Xun Gu. Evolutionary dynamics of mers-  
509 cov: potential recombination, positive selection and transmission. *Sci-  
510 entific reports*, 6(1):1–10, 2016.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TableS5.xlsb](#)
- [Supplementary.pdf](#)
- [TableS3.xlsb](#)
- [Supplementary.pdf](#)
- [TableS3.xlsb](#)
- [TableS4.xlsb](#)
- [TableS4.xlsb](#)
- [TableS5.xlsb](#)