PhyloTransformer: A Self-supervised Discriminative Model for SARS-CoV-2 Viral Mutation Prediction Based on a Multi-head Self-attention Mechanism

Yingying Wu[1,2,†], Shusheng Xu[3,4,†], Shing-Tung Yau[1] and Yi Wu[3,4]

[1]Harvard University, Center of Mathematical Sciences and Applications
[2]University of Houston, Houston, U.S.
[3]Tsinghua University, Institute for Interdisciplinary Information Sciences, Beijing, China
[4]Shanghai Qi Zhi Institute, Shanghai, China

Abstract

In this article, we developed PhyloTransformer, a Transformer-based self-supervised discriminative model, which can model genetic mutations that may lead to viral reproductive advantage. We trained PhyloTransformer on 1,765,297 severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) sequences to infer fitness advantages, by directly modeling the nucleic acid sequence mutations. PhyloTransformer utilizes advanced techniques from natural language processing to enable efficient and accurate intra-sequence dependency modeling over the entire RNA sequence. We measured the prediction accuracy of novel mutations and novel combinations using our method and baseline models that only take local segments as input. We found that PhyloTransformer outperformed every baseline method with statistical significance. We also predicted the occurrence of mutations in each nucleotide of the receptor binding motif (RBM) and predicted modifications of *N*-glycosylation sites. We anticipate that the viral mutations predicted by PhyloTransformer may identify potential mutations of threat to guide therapeutics and vaccine design for effective targeting of future SARS-CoV-2 variants.

Keywords

COVID-19, SARS-CoV-2, spike protein, variants of concern, PhyloTransformer, self-supervised neural network

Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is the causative agent of Coronavirus disease 2019 (COVID-19). The unprecedented COVID-19 pandemic is one of three major pathogenic zoonotic disease outbreaks caused by β-coronaviruses in the past two decades [1, 2]. Severe acute respiratory syndrome coronavirus (SARS-CoV) emerged in 2002, infecting 8,000 people with a 10% mortality rate [3, 4]. Middle East respiratory syndrome coronavirus (MERS-CoV) emerged in 2012 with 2,300 cases and a 35% mortality rate [5]. The third outbreak, mediated by SARS-CoV-2, emerged in 2019 with a mortality rate of 3.6% [6] and 219 million cases have been reported as of October 2021.

After the emergence of SARS-CoV-2 in late 2019, the virus exhibited relative evolutionary stasis for approximately 11 months. Since the end of 2020, SARS-CoV-2 has consistently acquired approximately two mutations per month [7] resulting in novel variants of concern (VOCs). As more individuals became vaccinated against SARS-CoV-2, the viral evolution has been characterized by the emergence of sets of mutations, probably in response to the changing immune profile of the human population. Currently, the main focus is to identify critical SARS-CoV-2 countermeasures, including vaccines, therapeutics, and diagnostics.

Since coronaviruses have proofreading functions [8], most mutations in the SARS-CoV-2 genome are expected to comprise neutral amino acid changes with little or no impact on fitness advantages [9]. However, the evolutionary diversity introduced by a small minority of mutations may impact the virus phenotype and promote virus fitness. Some of the SARS-CoV-2 mutations displayed positive selection with improved pathogenicity, infectivity [10], transmissibility [11, 12], angiotensin converting enzyme 2 (ACE2) binding affinity [13], or antigenicity [14]. In addition, other SARS-CoV-2 mutations introduced an optimized trade-off to improve overall fecundity. Heavily mutated lineages have also been reported, such as the lineage B.1.1.298, which harbors the following four amino acid substitutions: ΔH69–V70, Y453F, I692V, and M1229I [15]. Some mutations may amplify other mutations, providing an improved fitness advantage. For example, the combination of E484K, K417N, and N501Y results in the highest degree of conformational alterations compared to either E484K or N501Y alone [16]. Accumulating evidence suggests that mutations which require

KDH@IJCIA 2023: The 6th international workshop on knowledge discovery from healthcare data, Macao, S.A.R.: 19-20 August, 2023.
[†]These authors contributed equally.
ywu68@uh.edu (Y. Wu); xssstory@gmail.com (S. Xu); yau@math.harvard.edu (S. Yau); jxwuyi@gmail.com (Y. Wu)

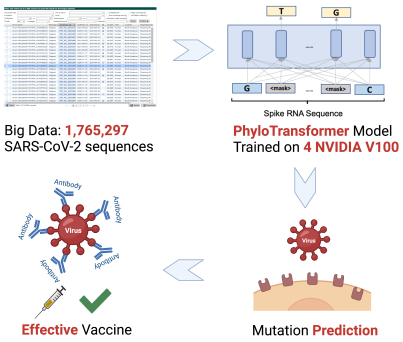


Figure 1: PhyloTransformer prediction paradigm.

immediate attention are circulating, which highlights the urgent need to develop effective prevention and treatment strategies.

While vaccination has been the most important and effective preventive measure, it is also facing challenges. The mRNA vaccine BNT162b2 (Pfizer–BioNTech) has 95% efficacy against COVID-19 [17]. However, the estimated effectiveness of the vaccine against the B.1.1.7 variant was 89.5% (95% CI, 85.9 to 92.3) at 14 or more days after the second dose and 75.0% (95% CI, 70.5 to 78.9) against the B.1.351 variant [18] at 14 or more days after the second dose. Several studies have characterized multiple mutations that change the antigenic phenotype. Thus, these studies elucidate how these mutations affect antibody-mediated neutralization. Variants containing these mutations are potentially highly virulent and have received much recent attention. However, it remains unknown whether more infectious variants exist along with the likelihood that they will appear and transmit. Designing vaccines after a novel variant has emerged is not optimal because the variant could potentially compromise existing vaccines and spread among the population. Thus, more infections might generate further variants, leading to a never-ending pandemic.

In order to win the race against the rapidly evolving SARS-CoV-2, an intelligent system capable of forecasting potential VOCs before they actually appear is urgently required. Therefore, in order to infer fitness advantages, we proposed PhyloTransformer, which models constraints from natural sequences, including long-range dependencies between positions. We hope that PhyloTransformer can be used to predict novel mutations and novel combinations of mutations in SARS-CoV-2, as depicted in Fig. 1. Thus, we anticipate that when variants of high consequence arise, existing vaccines based on PhyloTransformer predictions will have already been developed that target those strains.

Results

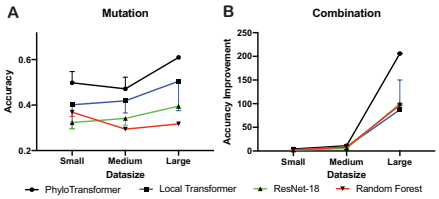
We used the hCoV-19/Wuhan/WIV04/2019 sequence (WIV04) as our reference sequence, which is the official reference sequence employed by GISAID (EPI_ISL_402124). WIV04 represented the consensus of several early submissions for the β-coronavirus responsible for COVID-19 [19], which was isolated by the Wuhan Institute of Virology from a clinical sample of a bronchoalveolar lavage fluid for RNA extraction and metagenomic next-generation sequencing. The consensus sequence was obtained by *de novo* assembly [20]. Based on WIV04, we define a *mutation* as the change in a nucleotide at a particular position that is different from the reference sequence. We define a mutation at a particular position that only occurs in the testing set but does not occur within the training set as a *novel mutation,* which signifies a mutation that is novel for the training set. We define all the novel mutations over an RNA sequence as a *novel combination,* i.e., a combination of mutations that do not occur in the training data. The prediction of *novel mutations* aims to predict single mutations, while the prediction of *novel combinations* aims to predict a collection of single mutations that jointly occur in a mutated sequence.

The prediction accuracies of *novel mutations* and *novel combinations* were evaluated after the predicting models PhyloTransformer, Local Transformer, and ResNet-18 converged. We first performed lag 1 autocorrelation to test the correlation between accuracy scores obtained from models that are one checkpoint apart. The autocorrelation tests were performed on *small, medium,* and *large* datasets for predicting *novel mutations* and *novel combinations,* with a total of 18 tests. We found no time dependency between the 10 accuracy scores in each of these 18 tests. For other classical machine learning models, we repeated the experiment 10 times for each dataset. The details are reported in Box 1C.

In this section, we first evaluated PhyloTransformer-generated predictions of *novel mutations* and *novel combinations.* Next, we compared the accuracy of each prediction with those obtained from baseline models. We then reported our predictions in the receptor binding motif (RBM). Finally, we predicted modifications of *N*-glycosylation sites to help identify mutations associated with altered glycosylation that might be favored during viral evolution. The detailed model architecture and training process are reported in the Methodology section.

Predicting Novel Mutations

We evaluated the efficacy of PhyloTransformer to predict *novel mutations* and compared it to baseline model predictions from three datasets with different sizes spanning different time frames. The prediction results are



C	Novel Mutation				Novel Combination			
	Mean	SD	Sig. Phylo	Sig. Local	Mean	SD	Sig. Phylo	Sig. Local
PhyloTransformer	0.61	0.00	-	< 0.0001	205.98	0.60	-	< 0.001
Local Transformer	0.50	0.13	< 0.001	-	87.31	62.85	< 0.001	-
ResNet-18	0.40	0.01	< 0.001	< 0.001	99.82	1.06	< 0.001	< 0.001
MLP	0.37	0.08	< 0.001	< 0.001	3.34	1.71	< 0.001	< 0.001
Logistic Regression	0.44	0.00	< 0.001	< 0.001	0.11	0.00	< 0.001	< 0.001
KNN	0.32	0.00	< 0.001	< 0.001	9.34	0.26	< 0.001	< 0.001
Random Forest	0.32	0.00	< 0.001	< 0.001	96.70	0.13	< 0.001	< 0.001
Gradient Boosting	0.44	0.00	< 0.001	< 0.001	0.09	0.00	< 0.001	< 0.001

: Box 1 | Prediction Accuracy. **A.** Prediction accuracy of *novel mutations* from the *small, medium,* and *large* datasets based on PhyloTransformer and the best baseline methods. **B.** Prediction accuracy of *novel combinations* trained with the *small, medium,* and *large* datasets based on PhyloTransformer and the best baseline methods. The accuracy improvement for each indicated model was calculated based on dividing the number of correct predictions by the expected number of correct random guesses. **C.** Prediction accuracy of PhyloTransformer–and baseline method–generated predictions of *novel mutations* and *novel combinations.* Sig. Phylo: p-value with respect to PhyloTransformer, compared to random guessing resulted in an accuracy of 0.26% with an SD = 0.012%. Sig. Local: p-value with respect to Local Transformer.

reported in Box 1. For each mutation, we masked the raw nucleotide in the reference sequence and predicted which nucleotide it would mutate to, and we selected the nucleotide with the highest confidence as our prediction. The prediction accuracy is the proportion of positions that are predicted correctly among all novel positions in the testing set. The prediction accuracy of random guessing is exactly 1/3. We evaluated the prediction efficacy averaged over 10 checkpoints after the convergence of PhyloTransformer, Local Transformer, and our baseline models on three datasets with the variance marked either below or above. Next, we reported the model predictions from each dataset, which is displayed in Box 1A.

We performed a two-sample z-test of proportions and found that for each model, the best prediction accuracy of *novel mutations* from the *large* dataset among the 10 checkpoints significantly less than PhyloTransformer. Local Transformer had the best performance among baseline models, but the average over 10 checkpoints was still 11% lower than that of PhyloTransformer on the *large* dataset with statistical significance, as shown in Box 1C. Table 1 reports the 20 *novel mutations* predicted by training PhyloTransformer with greatest probability using the *large* dataset.

Predicting Novel Combinations

If a sequence in the testing set does not exist in the training set, we compared it to the reference sequence, then masked the mutated positions and generated predictions at these positions. If the model predicts *all* the mutations correctly in this sequence, we say that it predicted a *novel combination* correctly. The accuracy of predicting *novel combinations* is the proportion of the number of sequences whose combinations are predicted correctly to all the sequences in the testing set.

The difficulty of predicting *novel combinations* changes as the size of the dataset changes, so we measure our prediction efficacy by accuracy improvement, which is defined as the following:

$$\text{Improvement} := \frac{\text{Model Acc.}}{\text{Random Guessing Acc.}}.$$

For the *small* dataset, there were 2.26 mutations on average with a standard deviation (SD) = 5.06; for the *medium* dataset, there were 3.06 mutations on average with an SD = 2.56; and for the *large* dataset, there were 8.75 mutations on average with an SD = 2.87. For the *small* dataset, random guessing resulted in an accuracy of 13.30% with an SD = 1.12%; for the *medium* dataset, random guessing resulted in an accuracy of 5.42% with an SD = 0.12%; and for the *large* dataset, random guessing resulted in an accuracy of 0.26% with an SD = 0.012%. The predicted results are summarized in Box 1B, where the accuracy improvement value was defined as follows: given the dataset (*small, medium,* or *large*), take the number of correct predictions generated by the indicated model and divide that value by the expected number of correct random guesses.

We performed a two-sample z-test of proportions to determine whether the accuracy of predicting *novel combinations* by PhyloTransformer significantly less than baseline models on the *large* dataset. The prediction accuracy of PhyloTransformer among the 10 checkpoints was higher than that generated by all of the baseline models with statistical significance. Local Transformer was no longer the best baseline model, while ResNet-18 and random forest outperformed Local Transformer for the task of predicting *novel combinations.*

Predictions in the Spike Protein RBM

SARS-CoV-2 infects human cells by binding of the viral surface protein spike to its receptor on human cells, the ACE2 protein. Because of its role in viral entry, the RBD is a dominant determinant of zoonotic cross-species transmission. Although SARS-CoV-2 does not cluster within SARS and SARS-related coronaviruses, the RBD of SARS-CoV and SARS-CoV-2 share structural similarities,

Table 1
Top 20 *novel mutations* predicted by training PhyloTransformer with the *large* dataset. Ref.: reference sequence hCoV-19/Wuhan/WIV04/2019 sequence (WIV04).

	In Amino Acid			In Nucleotide							
Rank	Loc.	Ref.	Pred.	Loc.	Ref.			Pred.		
1	587	I	T	1759	A	T	T	A	C	T
2	742	I	T	2224	A	T	T	A	C	T
3	538	C	R	1611	T	G	T	C	G	T
4	1080	A	V	3238	G	C	C	G	T	C
5	720	I	T	2158	A	T	T	A	C	T
6	851	C	R	2550	T	G	T	C	G	T
7	423	Y	H	1266	T	A	T	C	A	T
8	377	F	S	1129	T	T	T	T	C	T
9	823	F	L	2466	T	T	C	C	T	C
10	488	C	R	1461	T	G	T	C	G	T
11	819	E	G	2455	G	A	A	G	G	A
12	617	C	F	1849	T	G	C	T	T	C
13	749	C	R	2244	T	G	C	C	G	C
14	873	Y	H	2616	T	A	C	C	A	C
15	1059	G	V	3175	G	G	T	G	T	T
16	539	V	A	1615	G	T	C	G	C	C
17	421	Y	H	1260	T	A	T	C	A	T
18	877	L	P	2629	C	T	G	C	C	G
19	418	I	T	1252	A	T	T	A	C	T
20	1145	L	S	3433	T	T	A	T	C	A

probably due to their shared zoonotic ancestry. This similarity implies convergent evolution for improved binding to ACE2 between the SARS-CoV and SARS-CoV-2 RBDs. Therefore, we focused our predictions on the spike protein RBD. The total length of the SARS-CoV-2 spike protein is 1,273 amino acids, and its structural features are listed below:

- A signal peptide is located at the N-terminus (1–13 residues).
- The S1 subunit (14–685 residues) is responsible for receptor binding. The S1 subunit contains an N-terminal domain (14–305 residues), a C-terminal domain 0 (306-330 residues), an RBD (331-527 residues), a C-terminal domain 1 (528-590 residues), and a C-terminal domain 2 (591-685 residues).
- The S2 subunit (686–1273 residues) is responsible for receptor binding and membrane fusion. The S2 subunit contains cleavage sites (686-815 residues) at S1/S2 and S2', a fusion peptide (816–855 residues), a fusion peptide region (856-911 residues), a heptapeptide repeat sequence 1 (912–984 residues), a center helix (985-1034 residues), a connector domain (1035-1080 residues), a connector domain 1 (1081-1147 residues), a heptapeptide repeat sequence 2 (1163–1213 residues), a transmembrane domain (1213–1237 residues), and a cytoplasmic domain (1237–1273 residues) [21].

The spike protein RBM comprises amino acids 438 to 506. Yi et al. [22] compared the SARS-CoV-2 and SARS-CoV RBD affinity for hACE2 by creating single amino acid substitution mutations in the SARS-CoV and SARS-CoV-2 RBM sequences. The authors found that receptor binding was enhanced by introducing amino acid changes at P499, Q493, F486, A475, and L455, which are all localized to the RBM. PhyloTransformer trained with the *large* dataset predicted only two mutations. The first mutation was predicted at amino acid 488, changing it from C to R, which is closely adjacent to F486. The second mutation was predicted at amino acid 497, changing it from F to S, once again right next to P499. The close proximity of the introduced mutations and predicted mutations indicated that PhyloTransformer is potentially capable of capturing meaningful genetic phenomena and can generate effective predictions. Our prediction results are reported in Table 2.

Table 2
Prediction of spike protein RBM mutations. Ref.: reference sequence hCoV-19/Wuhan/WIV04/2019 sequence (WIV04).

In Amino Acid			In Nucleotide						
Loc.	Ref.	Pred.	Loc.	Ref.			Pred.		
488	C	R	1461	T	G	T	C	G	T
497	F	S	1489	T	T	C	T	C	C

Prediction of Glycosylation Site Modifications

The SARS-CoV-2 spike protein is heavily glycosylated. Viral glycosylation plays a vital role in viral pathobiology, including antibody resistance, target recognition, viral entry, and host immune modulation [23]. Glycosylation sites facilitate immune evasion by shielding epitopes from antibody neutralization; therefore, they are under selective pressure. Since glycosylation site modifications of the SARS-CoV-2 spike protein will likely impact the overall activities of SARS-CoV-2 replication and escape from immune surveillance [24], we examined glycosylation site model predictions. We reported our results on the *N*-glycosylation sites to help identify mutations associated with altered glycosylation that are favored during viral evolution. PhyloTransformer predicted three mutations of the following glycosylation sites: N122, N331, and N343. Table 3 shows the predicted mutations in the spike protein changing N to a different amino acid. Figure 3 summarizes the predicted mutations, including existing mutations (left) and *novel mutations* (right), with predictions mutating away from amino acid N highlighted.

Table 3
Predictions of glycosylation sites and N-mutations. *First three rows:* predicted glycosylation site mutations. *N-mutation sites:* other predictions with mutations of N. Ref.: reference sequence hCoV-19/Wuhan/WIV04/2019 sequence (WIV04).

Sites	In Amino Acid			In Nucleotide						
	Loc.	Ref.	Pred.	Loc.	Ref.			Pred.		
Glycosy.	122	N	D	363	A	A	C	G	A	C
Glycosy.	331	N	D	990	A	A	T	G	A	T
Glycosy.	343	N	S	1027	A	A	C	A	G	C
N-mut.	422	N	D	1263	A	A	T	G	A	T
N-mut.	542	N	D	1623	A	A	C	G	A	C
N-mut.	542	N	S	1624	A	A	C	A	G	C
N-mut.	953	N	D	2856	A	A	C	G	A	C

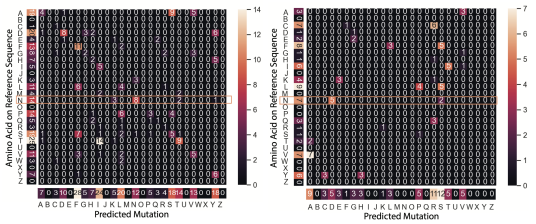


Figure 3: Predicting amino acid changes. Left: existing mutations. **Right:** predicted *novel mutations* (total: 196 predictions). **Row:** amino acids of the SARS-CoV-2 reference sequence (total: 69 predictions). **Column:** predicted amino acids.

Methodology

Technical Background

In this section, we will briefly review the history of sequence models that led to the development of Transformer and then introduce our PhyloTransformer model. The recurrent neural network (RNN) is the standard neural sequence model which extends the conventional feed-forward neural network with a recurrent hidden state dependent on the previous timestep. RNN and its variants, such as the long short-term memory (LSTM) [25] and the gated recurrent unit (GRU) [26], have been widely applied to important AI tasks, including language modeling [27], speech recognition [28], handwriting recognition [29], and machine translation [30]. However, RNNs are difficult to train in practice since the gradients tend to either vanish or explode as the sequence length increases [31]. In addition, these models encode a source sequence into a fixed-length vector, which becomes a bottleneck when tackling particularly long sequences. Therefore, the attention mechanism was introduced [32] to augment RNNs with an additional variable-length representation when encoding the input sequence. The attention mechanism allows the model to only focus on a subset of the input sequence for decoding. The Transformer model comprises a purely attention-based network architecture without RNN backbones to directly capture intra-position dependencies via the self-attention mechanism [33]. In self-attention, each sequence item has direct access to all the other positions, which yields a more powerful global representation of the sequence. This feature also inspires biological applications due to the long-range interactions of genetic sequences. However, the following challenges in modeling mutations on RNA sequences remain:

- **Length adaptation:** most natural language processing (NLP) models deal with sequence lengths of a few hundred to a thousand, but the RNA sequence of SARS-CoV-2 is much longer: the genome of SARS-CoV-2 is 29,903 nucleotides in length [34], and the spike protein has 3,819 nucleotides.
- **Mutation sparsity:** due to the proofreading functions of coronaviruses [8], mutations in the SARS-CoV-2 genome are rare. Our dataset shows consistency in this regard.

Regular Transformer scales quadratically with respect to the input sequence length, and the sparsity of mutations might lead to the generative Transformer model overfitting the identical parts while ignoring the mutations. Therefore, to adapt to biological problems and address issues regarding genetic mutations, a new model that tackles the length and sparsity issues commonly encountered in existing deep neural network architectures is required. To address these two challenges, we propose PhyloTransformer, which is a linear time complexity discriminative model based on the Transformer architecture. The time and space linearity are achieved by adopting FAVOR+ from Performer [35], which performs an unbiased fast attention approximation with low variance. The mutation sparsity issue is addressed by directly modeling the mutations using the MLM training objective from BERT [36], which is a discriminative variant of Transformer for supervised NLP tasks. A detailed description of PhyloTransformer architecture is presented in the next section.

Model Development

We adopted a discriminative approach to model the mutation probability at a particular position in the RNA sequence. Let $p(x_i = A|X)$ denote the probability of the i^{th} nucleotide changing to A given the reference sequence X. We will demonstrate how to predict $p(x_i|X)$ by PhyloTransformer and other baseline models in this section.

The PhyloTransformer Model

The PhyloTransformer model adopts a Transformer-based network, which utilizes the full spike sequence of 3,819 nucleotides as input and generates output mutation probabilities at particular positions. We followed the MLM pre-training objective from BERT [36]. Note that the attention mechanism in Transformer [33] calculates attention matrices with a shape of $L \times L$ (where L is the length of the sequence) to capture the relationship between nucleotides. In order to reduce the computation complexity of the attention matrix, we adopted the FAVOR+ technique from Performer [35], which performs approximate attention computation in linear time. In the following content, we first present the network architecture of PhyloTransformer. Next, we introduce FAVOR+ for fast low-rank approximation of the regular full-rank attention computation in linear time. Finally, the overall training process will be discussed in detail.

Bidirectional Transformer Encoder: Let $X = (x_1, x_2, ..., x_L)$ denote the reference sequence, where x_i is the nucleotide at position i in the RNA sequence. We first applied trainable projections to map each x_i with its position information to three embedding vectors, q_i, k_i and v_i, for attention computation. Suppose the dimension of each embedding is d. The output of the attention layer is computed by the following equation:

$$\text{Attention}(Q, K, V) = A \cdot V = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (1)$$

where $A \in \mathbb{R}^{L\times L}$ is the attention matrix. $Q = [q_1; q_2; ...; q_L], K = [k_1; k_2; ...; k_L]$, and $V = [v_1; v_2; ..., v_L]$ are embedding matrices in $\mathbb{R}^{L\times d}$, where q_i, k_i, and v_i are row vectors representing three embeddings. After the attention layer is computed, we further applied a feed-forward layer with a residual connection. An attention layer and a feed-forward layer compose a single Transformer module. We stacked the N Transformer modules as the overall network architecture of our PhyloTransformer model.

FAVOR+: In the original attention mechanism, the time complexity of computing the attention layer by Equation (1) is $O(L^2 d)$, which becomes computationally intractable when L is large. The Performer [35] model proposed kernelizable attention by deriving a mapping ϕ to decouple the attention matrix A into Q' and K', where $q_i' = \phi(q_i), k_i' = \phi(k_i)$ and $Q', K' \in \mathbb{R}^{L\times r}, r \ll L$. In this case, the attention layer can be computed by the following equation:

$$\text{Attention}(Q, K, V) = D^{-1}(Q'((K')^T V)), \quad (2)$$

$$D = \text{diag}(Q'((K')^T \mathbf{1}_L)) \quad (3)$$

where $\mathbf{1}_L$ is an all-ones vector of length L. Since $Q', K' \in \mathbb{R}^{L\times r}, V \in \mathbb{R}^{L\times d}$, the computation complexity decreases to $O(rLd)$ with respect to a small constant r, making it computationally feasible to handle particularly long sequences such as RNA data.

Training process: We denoted the reference sequence as $X = (x_1, x_2, ..., x_L)$ and the mutated sequence as $Y = (y_1, y_2, ..., y_L)$, where x_i and y_i refer to the nucleotide at position i. On average, there were 0.0592% mutations in the small dataset, 0.0801% mutations in the medium dataset, and 0.2291% mutations in the large dataset. These numbers refer to the average number of y_is that are different from the number of x_is in the respective dataset. During the training process, we masked certain positions in X, and used the model to predict nucleotides in Y at those masked positions. Fig. 4 shows the workflow of our model. Specifically, we first identified the set of mutated positions $\mathcal{P}_m = (P_1, \ldots, P_k)$, where $P_1, \ldots, P_k$ are

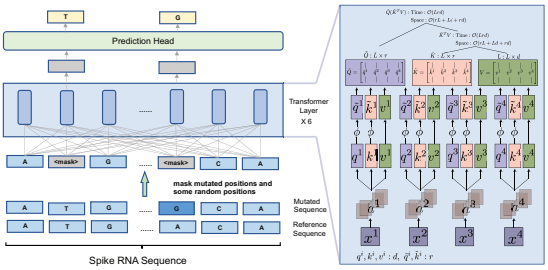


Figure 4: The training scheme of PhyloTransformer. We compared the reference sequence to a mutated sequence and focused on the mutations. We masked the mutated positions along with some random positions and processed the masked sequence with stacked Transformer modules. Each masked position's final hidden state was used as the aggregate representation for the mutation prediction task.

nucleotide positions. In addition, we randomly chose several unchanged positions $\mathcal{P}_u = (P_1', \ldots, P_r')$ such that $\frac{|\mathcal{P}_m \cup \mathcal{P}_u|}{L} = 1.5\%$. Next, we applied a masking function $f_m(x_i)$ to each nucleotide x_i at the masked positions. Namely, $\forall P_i \in \mathcal{P}_m \cup \mathcal{P}_u$, the masking function f_m changes the nucleotide x_i at position P_i to:

$$f_m(x_{P_i}) = \begin{cases} <mask> & 80\% \text{ of cases,} \\ x_{P_i} & 10\% \text{ of cases,} \\ \text{Random}(\{A, T, C, G\}) & 10\% \text{ of cases,} \end{cases} \quad (4)$$

where $<mask>$ is a special masking token. The masking function f_m acts on 1.5% of the entire nucleotides and further randomly maps each nucleotide from this masking subset to (1) a special token $<mask>$ (80% chance), (2) a random substitution (10%), or (3) itself (10%).

Denoting the masked sequence as $\widetilde{X}$, we encode $\widetilde{X}$ with stacked Transformer modules and represent each nucleotide as a hidden vector h_i from the model output. Next, the probability distribution of the ith nucleotide position over {A, T, C, G} is computed as follows:

$$P(y_i|\widetilde{X}) = \text{softmax}(W_o h_i) \quad \forall i \in \mathcal{P}_m \cup \mathcal{P}_u, \quad (5)$$

where W_o are trainable parameters. The probability of all the masked nucleotides is the following equation:

$$P(Y|\widetilde{X}) = \prod_{i\in\mathcal{P}_m\cup\mathcal{P}_u} P(y_i|\widetilde{X}) \quad (6)$$

The model is optimized to minimize the negative log probability over all the mutated sequences from the training set $\mathcal{Y}$ with respect to different masking positions, as determined by the equation:

$$L(\theta) = -\sum_{Y\in\mathcal{Y}} \mathbb{E}_{f_m}\left[\log P(Y|\widetilde{X})\right]. \quad (7)$$

$$L(\theta) = -\mathbb{E}_{Y\in\mathcal{Y}}\left\{\mathbb{E}_{i\in\mathcal{P}_m\cup\mathcal{P}_u}\left[\log P(y_i|\widetilde{X})\right]\right\}. \quad (8)$$

Since most of the masked positions are mutated positions, our model is trained to concentrate on mutation predictions. Meanwhile, the randomly chosen positions (i.e., $\mathscr{P}_u$) also improved the robustness of our model.

Local models

In addition to PhyloTransformer, which considers the full sequence, we also examined baseline methods, which predict $p(x_i|X)$ based on local segments from the spike RNA sequence. There is a total of 3,819 nucleotides in the spike sequence. We can obtain a local segment of 15 nucleotides centered around each nucleotide with sequence padding. Thus, we can obtain 3,819 segments of 15 nucleotides from the full spike RNA sequence. The center position of each segment is masked. We adopted various classification methods (including neural models and non-neural methods) to predict the center nucleotide based on other nearby nucleotides. During the training phase, we split all training spike RNA sequences into segments and generated a local dataset with repeated segments filtered out. The training process is shown in Appendix A, where any classification method could be used, such as the standard Transformer, ResNet-18, MLP, logistic regression, KNN, random forest, and gradient boosting.

Conclusion

The overall goal of our research is to train a state-of-the-art sequence model using existing viral genetic sequence data to identify SARS-CoV-2 variants that may have evolutionary advantages and become the emerging VOCs. In this paper, we developed the PhyloTransformer model, a novel deep neural network with a multi-headed self-attention mechanism. PhyloTransformer was subjected to an advanced training methodology to predict potential mutations that may lead to enhanced virus transmissibility or resistance to antisera. Our computational platform may be helpful in guiding the design of therapeutics and vaccines for effective targeting of emerging SARS-CoV-2 VOCs, as well as novel mutants of other viruses that may cause pandemics.

Ethics Statement This research was based on the SARS-CoV-2 sequences in the Global Initiative for Sharing All Influenza Data (GISAID) database (https://www.gisaid.org/). There is no human information involved in the data.

References

[1] J. Cui, F. Li, Z.-L. Shi, Origin and evolution of pathogenic coronaviruses, Nature Reviews Microbiology 17 (2019) 181–192.

[2] E. De Wit, N. Van Doremalen, D. Falzarano, V. J. Munster, SARS and MERS: recent insights into emerging coronaviruses, Nature Reviews Microbiology 14 (2016) 523–534.

[3] M. S. Patel, M. J. Gutman, J. A. Abboud, Orthopaedic considerations following COVID-19: lessons from the 2003 SARS outbreak, JBJS reviews 8 (2020) e20.

[4] D. S. Hui, E. I. Azhar, T. A. Madani, F. Ntoumi, R. Kock, O. Dar, G. Ippolito, T. D. Mchugh, Z. A. Memish, C. Drosten, et al., The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health—the latest 2019 novel coronavirus outbreak in wuhan, china, International journal of infectious diseases 91 (2020) 264–266.

[5] R. L. Graham, R. S. Baric, Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission, Journal of virology 84 (2010) 3134–3146.

[6] D. Baud, X. Qi, K. Nielsen-Saines, D. Musso, L. Pomar, G. Favre, Real estimates of mortality following COVID-19 infection, The Lancet infectious diseases 20 (2020) 773.

[7] M. Worobey, J. Pekar, B. B. Larsen, M. I. Nelson, V. Hill, J. B. Joy, A. Rambaut, M. A. Suchard, J. O. Wertheim, P. Lemey, The emergence of SARS-CoV-2 in europe and north america, Science 370 (2020) 564–570.

[8] E. C. Smith, H. Blanc, M. Vignuzzi, M. R. Denison, Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: evidence for proofreading and potential therapeutics, PLoS pathogens 9 (2013) e1003565.

[9] O. A. MacLean, R. J. Orton, J. B. Singer, D. L. Robertson, No evidence for distinct types in the evolution of SARS-CoV-2, Virus Evolution 6 (2020) veaa034.

[10] L. Yurkovetskiy, X. Wang, K. E. Pascal, C. Tomkins-Tinch, T. P. Nyalile, Y. Wang, A. Baum, W. E. Diehl, A. Dauphin, C. Carbone, et al., Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant, Cell 183 (2020) 739–751.

[11] Y. J. Hou, S. Chiba, P. Halfmann, C. Ehre, M. Kuroda, K. H. Dinnon, S. R. Leist, A. Schäfer, N. Nakajima, K. Takahashi, et al., SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo, Science 370 (2020) 1464–1468.

[12] E. Volz, V. Hill, J. T. McCrone, A. Price, D. Jorgensen, Á. O'Toole, J. Southgate, R. Johnson, B. Jackson, F. F. Nascimento, et al., Evaluating the effects of SARS-CoV-2 spike mutation D614G on transmissibility and pathogenicity, Cell 184 (2021) 64–75.

[13] T. N. Starr, A. J. Greaney, S. K. Hilton, D. Ellis, K. H. Crawford, A. S. Dingens, M. J. Navarro, J. E. Bowen, M. A. Tortorici, A. C. Walls, et al., Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ace2 binding, Cell 182 (2020) 1295–1310.

[14] E. C. Thomson, L. E. Rosen, J. G. Shepherd, R. Spreafico, A. da Silva Filipe, J. A. Wojcechowskyj, C. Davis, L. Piccoli, D. J. Pascall, J. Dillen, et al., Circulating SARS-CoV-2 spike N439K variants maintain fitness while evading antibody-mediated immunity, Cell 184 (2021) 1171–1187.

[15] J. Fonager, S. scientist Ria Lassaunière1, senior scientist Jannik Fonager1, S. scientist Morten Rasmussen, A. Frische, S. S. C. P. Strandh, S. scientist veterinarian Thomas Bruun Rasmussen, C. veterinarian Anette Bøtner, C. V. A. Fomsgaard, Working paper on SARS-CoV-2 spike mutations arising in danish mink, their spread to humans and neutralization data., ???? URL: https://files.ssi.dk/Mink-cluster-5-short-report_AFO2(2020).

[16] G. Nelson, O. Buzko, P. R. Spilman, K. Niazi, S. Rabizadeh, P. R. Soon-Shiong, Molecular dynamic simulation reveals e484k mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y. V2 variant) induces conformational change greater than n501y mutant alone, potentially resulting in an escape mutant, BioRxiv (2021).

[17] F. P. Polack, S. J. Thomas, N. Kitchin, J. Absalon, A. Gurtman, S. Lockhart, J. L. Perez, G. P. Marc, E. D. Moreira, C. Zerbini, et al., Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine, New England Journal of Medicine (2020).

[18] L. J. Abu-Raddad, H. Chemaitelly, A. A. Butt, Effectiveness of the BNT162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants, New England Journal of Medicine (2021).

[19] P. Okada, R. Buathong, S. Phuygun, T. Thanadachakul, S. Parnmen, W. Wongboot, S. Waicharoen, S. Wacharapluesadee, S. Uttayamakul, A. Vachiraphan, et al., Early transmission patterns of coronavirus disease 2019 (COVID-19) in travellers from wuhan to thailand, january 2020, Eurosurveillance 25 (2020) 2000097.

[20] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang, H.-R. Si, Y. Zhu, B. Li, C.-L. Huang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, nature 579 (2020) 270–273.

[21] Y. Huang, C. Yang, X.-f. Xu, W. Xu, S.-w. Liu, Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19, Acta Pharmacologica Sinica 41 (2020) 1141–1149.

[22] C. Yi, X. Sun, J. Ye, L. Ding, M. Liu, Z. Yang, X. Lu, Y. Zhang, L. Ma, W. Gu, et al., Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ace2 and neutralizing antibodies, Cellular & molecular immunology 17 (2020) 621–630.

[23] K. J. Doores, The hiv glycan shield as a target for broadly neutralizing antibodies, The FEBS journal 282 (2015) 4679–4691.

[24] D. Hoffmann, S. Mereiter, Y. J. Oh, V. Monteil, R. Zhu, D. Canena, L. Hain, E. Laurent, C. Gruber, M. Novatchkova, et al., Identification of lectin receptors for conserved SARS-CoV-2 glycosylation sites, bioRxiv (2021).

[25] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780.

[26] K. Cho, B. Van Merriënboer, D. Bahdanau, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint arXiv:1409.1259 (2014).

[27] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, S. Khudanpur, Recurrent neural network based language model, in: Eleventh annual conference of the international speech communication association, 2010.

[28] A. Graves, A.-r. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE international conference on acoustics, speech and signal processing, Ieee, 2013, pp. 6645–6649.

[29] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, J. Schmidhuber, A novel connectionist system for unconstrained handwriting recognition, IEEE transactions on pattern analysis and machine intelligence 31 (2008) 855–868.

[30] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1700–1709.

[31] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks 5 (1994) 157–166.

[32] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473 (2014).

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[34] D. Kim, J.-Y. Lee, J.-S. Yang, J. W. Kim, V. N. Kim, H. Chang, The architecture of sars-cov-2 transcriptome, Cell 181 (2020) 914–921.

[35] K. M. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Q. Davis,

A. Mohiuddin, L. Kaiser, D. B. Belanger, L. J. Colwell, A. Weller, Rethinking attention with performers, in: International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=Ua6zuk0WRH.

[36] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

Table 4
Datasets used during training. The analysis was based on the GISAID database. Each dataset was evenly split into training data and testing data while retaining their temporal order.

Dataset	Start Date	Train End Date	Test End Date	Total	Training Set	Testing Set
Small	01/01/2020	03/20/2020	03/31/2020	24,951	12,475	12,476
Medium	01/01/2020	04/22/2020	09/30/2020	134,704	67,352	67352
Large	01/01/2020	02/17/2021	05/31/2021	1,765,297	882,648	882,649

	In Training Set			In Testing Set		
VOC	VOC Mutation	Unmutated	Other Mutation	VOC Mutation	Unmutated	Other Mutation
K417N$^{\beta}$	0.81%	99.05%	0.14%	1.67%	95.48%	2.85%
K417T$^{\gamma}$	0.14%	99.05%	0.81%	2.85%	95.48%	1.67%
T478K$^{\delta}$	0.46%	99.48%	0.06%	3.46%	96.48%	0.06%
L452R$^{\delta}$	2.38%	97.57%	0.05%	6.41%	93.40%	0.20%
E484K$^{\alpha,\beta,\gamma}$	1.60%	98.33%	0.07%	8.78%	90.86%	0.36%
N501Y$^{\alpha,\beta,\gamma}$	18.51%	81.24%	0.24%	73.96%	25.89%	0.15%
D614G$^{\alpha,\beta,\gamma,\delta}$	96.44%	3.54%	0.02%	99.39%	0.60%	0.01%
P681H$^{\alpha}$	19.34%	80.30%	0.36%	73.12%	24.39%	2.49%
P681R$^{\delta}$	0.30%	80.30%	19.41%	2.47%	24.39%	73.14%

A. Dataset Details

We list the details of our three datasets in table 4.

B. Local Models

Fig. 5 depicts the training scheme of local models. We first split the spike RNA sequence into 3,819 segments with padding. We then masked the middle nucleotide in each segment and adopted a classification model to predict the masked nucleotides. In this training scheme, any classification model could be used.

C. Training details

For the PhyloTransformer model, we stacked six Transformer modules with eight attention heads and a hidden size of $1,024$. We optimized the model following the loss function with Adam ($\beta_1 = 0.9, \beta_2 = 0.999$). We chose a learning rate of $3e-5$ for all three datasets. The batch size was 16 for the *small* and *medium* datasets, and the batch size was 32 for the *large* dataset. For the *large* dataset, we trained the PhyloTransformer model with 8 Nvidia V100 GPUs for 10 epochs, which took 13 hours per epoch. For Local Transformer, we stacked 12 Transformer modules with eight attention heads and a hidden size of 768 for better representation capability. We employed a standard classification loss and optimized the model via Adam ($\beta_1 = 0.9, \beta_2 = 0.98$). We used the learning rate of $1e-4$ with a batch size of 128 for all three datasets on a single Nvidia 3080 GPU with 100 training epochs. For the *large* dataset, each epoch was completed in approximately 20 minutes. For the ResNet model, a popular variant of the convolutions neural network, we employed the ResNet-18 architecture as our backbone, and Adam ($\beta_1 = 0.9, \beta_2 = 0.999$) was utilized as the optimizer with a learning rate of $5e-5$, and a batch size of 128. We trained the ResNet model for 100 epochs on a single Nvidia 3080 GPU. In the *large* dataset, a single epoch was completed in approximately five minutes. For other methods, we used scikit-learn (0.23.2) [1] with its default settings.

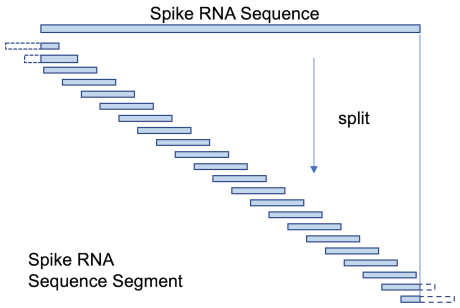


Figure 4: Step 1: The spike RNA sequence was split into 3,819 segments with padding.

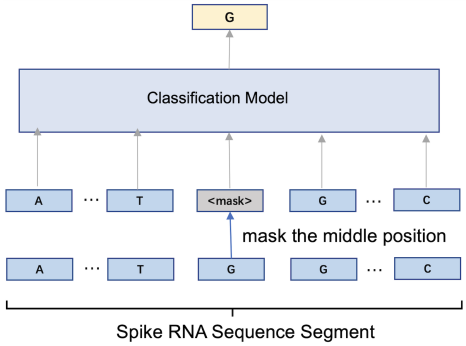


Figure 4: Step 2: The middle nucleotide was masked in each segment. A classification model was adopted to predict the masked nucleotides.

Figure 5: The training scheme of local models. Any classification method could be used, such as the standard Transformer, ResNet-18, MLP, logistic regression, KNN, random forest, and gradient boosting.

[1] https://scikit-learn.org/stable/