An Interpretable Deep Learning Model for Predicting the Risk of Severe COVID-19 from Spike Protein Sequence

Bahrad A. Sokhansanj (bahrad@molhealtheng.com)
Drexel University
Zhengqiao Zhao
Drexel University
Gail L. Rosen
Drexel University

Research Article

Keywords:

Posted Date: January 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1234007/v1

An Interpretable Deep Learning Model for Predicting the Risk of Severe COVID-19 from Spike Protein Sequence

Bahrad A. Sokhansanj[1,*], **Zhengqiao Zhao**[1], **and Gail L. Rosen**[1]

[1]Drexel University, Ecological and Evolutionary Signal-Processing and Informatics Laboratory, Department of Electrical & Computer Engineering, College of Engineering, Philadelphia, PA 19104, USA
[*]bahrad@molhealtheng.com

ABSTRACT

Throughout the COVID-19 pandemic, the virus has mutated in ways that affect its ability to infect people, cause severe disease, and escape immunity. It can be costly and time-consuming to experimentally study viral mutations. Sequencing genetic code is cheaper, and millions of SARS-CoV-2 genome sequences are available. With the quickly changing dynamics of SARS-CoV-2 evolution and patient outcomes, we need fast ways to translate sequence data to biologically meaningful and clinically relevant information. Inspired by advances in natural language processing, we design a deep learning architecture that can be visualized at multiple scales to interpret trained models. We train a model to predict the risk of severe disease based on genetic changes in the SARS-CoV-2 spike protein, which plays a key role in infection and immune response. Trained solely on spike protein sequences from pre-Omicron infections (i.e., acquired before any empirical data for Omicron was available), the model predicts Omicron sequences with a reduced risk of severe disease (by 40-50%) relative to Delta. Testing on Omicron sequences collected so far, the deep learning model's predictions agree with real world observations, suggesting that the methodology can be applied to future variants.

Introduction

The emergence of the COVID-19 pandemic has coincided with the widespread availability of lower cost, rapid whole genome sequencing. As of October 2021, when the training data for the models presented in this paper were collected, nearly 4 million SARS-CoV-2 sequences were available to researchers from the GISAID website (http://www.gisaid.org).[1] By December 22, 2021, that number had grown to nearly 6.4 million sequenced genomes. The critical challenge is to translate abundant genomic sequence data to as much biological insight and actionable clinical information as possible. This goal is particularly urgent as the virus continues to rapidly mutate and change in virulence—and experimental tools other than sequencing are costly, time-consuming, and often technically difficult to perform.

Connecting genomic sequence data to phenotype typically relies on phylogenetic methods and sequence alignment to: (i) group viral sequence variants into strains or lineages whose properties can be measured collectively, and (ii) identify specific mutations with potential impact on virulence, transmission or immune escape. However, the quantity, diversity, and pace of changes in the SARS-CoV-2 genome—driven by its novelty to the human immune system and sheer number of infections worldwide—has strained computational methods.[2] Indeed, so much data has been generated that it has spurred the development of computational tools for high frequency variant tracking.[3] While tree-based phylogenetic methods such as the NextStrain project's visualizations[4] (https://nextstrain.org/) have been powerful tools to trace the transmission and origins of SARS-CoV-2, they are not as useful in classifying genomes because the virus can accumulate and drop mutations in parallel across clades and subclades.[5] The principal way to classify the viral lineages has instead been the Pango nomenclature (https://cov-lineages.org/), derived from machine learning classification of viral "lineages" based on constituent sequence mutations with a phylogenetic relationship.[6,7] Particularly significant Pango lineages have been identified by the World Health Organization (WHO) as variants of concern (VOC) and given Greek letter designations[8], such as Alpha (Pango lineage designation B.1.1.7), Beta (B.1.351), Delta (B.1.167.2), and, recently, Omicron (B.1.1.529).

Changes to SARS-CoV-2 properties, however, often implicate combinations of multiple mutations that emerge simultaneously—and then sometimes revert in whole or in part as the virus continues to evolve.[9] Widespread lineages, such as Delta (Pango designation B.1.617.2), have spawned complex sublineages with distinct immune evasion and virulence properties—and can genetically share more in common with other lineages than their parents.[10] During a long-term infection, a spike protein may emerge with multiple variations, i.e., a "long branch" divergence from the phylogenetic tree, which may explain the origin of the Omicron variant.[11] The recurrence and reversion of mutations in complex combinations make it hard to connect static

sequenced classifications based on phylogenetic relationships to phenotype.[2] Alternatives based on bar-coding[12,13] may also fail as viral evolution becomes more complex.[14]

This paper focuses specifically on sequence variants of the SARS-CoV-2 spike protein. The spike protein covers the surface of the virus particle, binds to host cell receptors, and mediates cell entry.[15] Early in the pandemic, spike protein sequence variants began to be observed, such as the substitution D614G that rapidly became dominant as the pandemic spread from Asia to Europe.[2] Since then, the amount of further spike variation has been substantial, likely because the virus has been exposed to so many immune-naive hosts around the world.[16] The spike protein will likely continue to change. The spike gene of another human coronavirus, HCoV-OC43, has a higher rate of substitution and positive selection sites than the virus's other genes, suggesting that spike genetic drift plays a role in adaptive evolution.[17] One study comparing SARS-CoV-2 to other related coronavirus genomes estimated that, as of July 2021, SARS-CoV-2 had only "explored" 31% of the potential space for spike sequence variation.[18] Spike mutations impacts key virus properties, including virulence and immune evasion. For example, the Delta (B.1.617.2) variant has increased transmissibility and risk of severe disease based on both epidemiological data[19] and laboratory studies that show increased fitness over previous variants, such as enhanced viral replication due to modification in the furin cleavage site of the spike protein.[20,21]

In this paper, we present here a deep learning model built to predict disease severity of sequences by using patient metadata available for a subset of sequences in the GISAID database.

After the model was trained and validated, using data available as of October 2021, a new SARS-CoV-2 lineage emerged, initially designated B.1.529 and quickly identified as a Variants of Concern (VOC) and named Omicron. First reported on November 25, 2021, the Omicron variant has since exhibited rapid growth around the world.[22] Omicron contains numerous spike protein mutations, immediately causing concern that not only could it evade the immune response in previously infected and vaccinated patients, but it could also have different infectivity and virulence properties.[23] Omicron represents a particular challenge for machine learning models, since Omicron's sequence differs significantly from the data used to train the model. Because we did not consider any sequence data in GISAID after October 2021, the model's training data excludes Omicron sequences. Thus, Omicron sequences severity prediction task is used as a real world evaluation of our deep learning model. We show that the model's predictions about Omicron's relative risk of disease severity agree with recent experimental data, despite our model never having seen an example in training. This proof-of-concept shows the potential for a predictive computational model of variant properties that can provide a starting point for epidemiological guidance before health observations and laboratory data are collected.

Model Design

We propose deep neural network models of protein sequence as a complementary way to translate sequence data to viral phenotype, e.g., virulence, immune evasion, and other functional properties. Fig. 1 shows the model architecture. Amino acids are encoded analogously with the words in a sentence in language processing models, by assigning them integer tokens, as described in the Methods section. The model can then be trained to predict function from sequence.

We employ a Transformer architecture for sequence encoding, first described for an encoder-decoder architecture to machine translate text.[24] The Transformer is a modular multi-head structure, each head consisting of an attention layer and feed forward neural network, wherein head outputs are added and normalized to provide a sequence encoding.[24] The token and position embedding dimension (N_E) is large, because the spike protein is (typically) 1273 amino acids, and the embedding must include both positional and token information. We employ a convolutional neural network (CNN) layer with a kernel width of 1 and N_C filters before the transformer as a way to reduce the required size of the transformer heads, allowing us to load our model on a single processing unit for efficient computation.

While neural networks excel at fitting data, they suffer from a "black box" problem: it is very hard to explain a neural network's predictions.[25] We add two layers to not only help classify data but also allow us to visualize and potentially interpret the models that we train. The first is a forward attention layer[26], based on a structure used to analyze text[27] that can also be applied to biological sequences.[28,29] Here, we apply this structure to be able to readily access sequence-level classification, since the attention heads in Transformers cannot individually, nor in sum, provide full attention visualization across the sequence (see, e.g.,[30,31]). The second is an intermediate N_H-dimension densely connected layer, which can project an N_H-dimensional embedding of the sequence as a whole. The embeddings of a group of sequences post-training can be visualized or clustered to find patterns at the group level.[29,32] As further explained below, we concatenate integer-valued demographic variables (age, date, gender) from patient metadata with the L-length Transformer output and feed it to the N_H-dimensional embedding layer. The resulting embedded vectors will thus reflect both sequence and sample demographics. Finally, a sigmoid activation node can provide the class prediction.

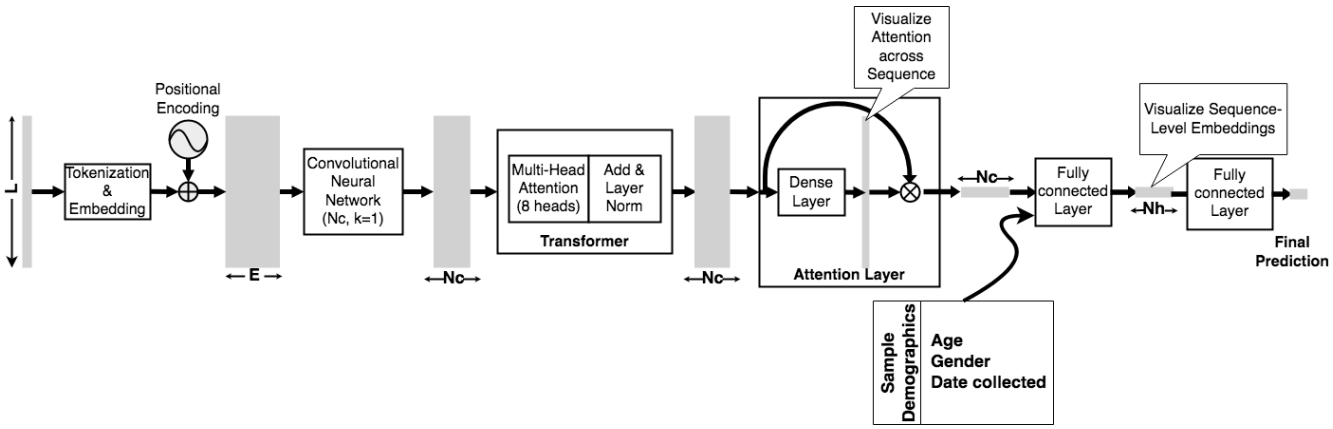

Figure 1. Deep learning architecture used in this paper, further described in the accompanying text.

Results

Training and Validating a Deep Learning Model for the Severity of Patient Outcomes

The deep learning model is trained on the fraction of GISAID sequence data which has patient status metadata. Approximately 147,000 of the 3.9 million samples available on the October 2021 cutoff date for inclusion in our study include both patient metadata and spike protein sequence information. Of the metadata entries, many are simply "unknown" or contain irrelevant information, such as the sample collection method or location. Additionally, many descriptions cannot be used to infer disease severity. For example, a patient with a metadata entry of "Alive" may have had a wide range of outcomes from asymptomatic to ICU admission. After excluding all samples that could not be assigned to "Mild" or "Severe" following the scheme in Table S1, we were left with 54,081 samples for training and testing the model.

When developing a model to link sequence to disease outcomes, we discovered that we had to also consider demographic variables. Throughout the pandemic, there has been consistent epidemiological and clinical evidence that age and male sex/gender (GISAID includes a metadata field for "gender") are major risk factors for more severe COVID-19 symptoms.[33] Earlier studies have shown that GISAID data also supported an empirical correlation between age and data to clinical severity.[34,35] Figs. 2A and 2B show the relationship between these variables and clinical severity, which show, as expected, a correlation with age (increasing age results in increasing mean severity) and gender (male patients have a greater prevalence of severe outcomes). Notably, the age data do not show a consistent trend at the extremes of young and old. Extreme age values, however, are represented by far fewer samples in the data set, and thus may be susceptible to study bias of which samples were sequenced. For example, if more infants (with sequenced samples in the GISAID database) were hospitalized, even for incidental reasons, that would be reflected as a more severe case. Or, if very old patients were sequenced as part of a study of elderly patients who had survived or had a milder course of disease than expected for their age, that would skew a small number of samples milder.

Besides patient demographics, we find a notable trend in the frequency of severe outcomes based on sample collection date. Fig. 2C shows how the number of samples with severe disease in the GISAID database drops as a function of time through the pandemic. There has been a particularly sharp decrease since approximately February 2021. This trend has been consistent even as the prevalent genetic compositions of the virus have changed in many different ways over time. For example, the B.1 lineage (D614G) was first detected January 1, 2020 and no longer highly prevalent by mid-2020; eventually it was supplanted by Alpha (B.1.1.7) and Delta (B.1.617.2) and Delta sublineages like AY.4. Correspondingly, in the GISAID validation data set, 73% of samples designated as B.1 (first detected January 1, 2020) are categorized as Severe, dropping to 60% for the later-emerging Alpha (B.1.1.7) variant and again to 42% for Delta (B.1.617.2) and 26% for the AY.4 sublineage of Delta. This contradicts repeated studies showing that Alpha resulted in elevated hospitalizations, ICU admissions, and other markers of severe outcomes as compared to ancestral lineages, that Delta was yet more severe, and that overall other variants of concern emerging later in the pandemic lead to more severe outcomes overall.[36–39]

The trend of declining severity rate later in the pandemic coincides with changes in how COVID-19 is treated and can be prevented. For example, the early decline in 2020 in average case severity is consistent with a Canadian study showing that the case fatality rate (CFR) decreased between the first and second waves prior to any vaccination, even when controlling for age and increased testing.[40] The latter decline may be accounted for by better understanding of how to treat COVID-19, and, later in 2020, the emergence of monoclonal antibody therapies. The accelerated decline in case severity starting in early 2021 coincides with increasing vaccination rates, particularly of older and more vulnerable people. In addition to treatment and

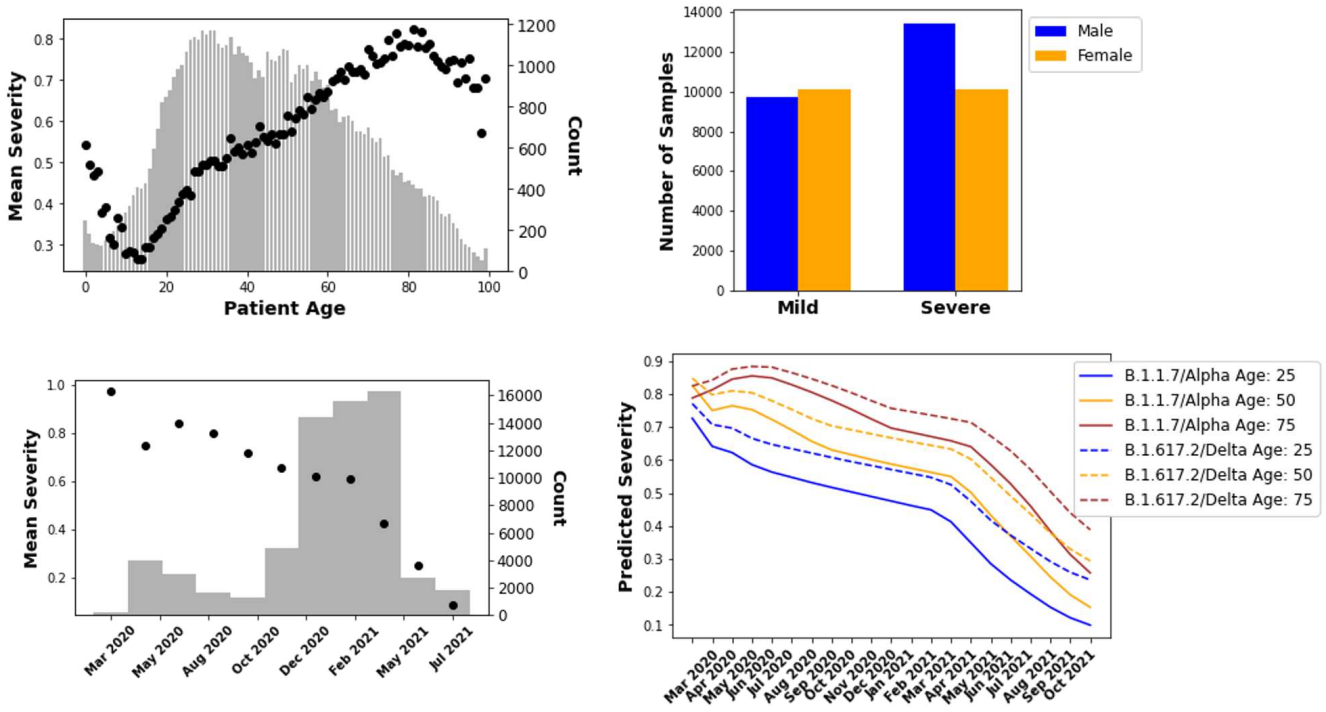

Figure 2. **(A – Upper Left)** Mean of clinical severity, where 0 is Mild and 1 is Severe (i.e., probability of a severe case) by patient age in the GISAID database, where age metadata are available. The bars show the count of samples for each age. **(B – Upper Right)** Number of mild and severe cases for each sex/gender as identified in the "gender" patient metadata field for GISAID samples. **(C – Lower Left)** Mean clinical severity (probability of severe case) by sample collection date recorded in the GISAID data. For clarity, data have been binned over time periods; the number of samples in each bin is shown by the bars. **(D – Lower Right)** Mean predicted case severity plotted for three different ages, with the gender variable set to male, are shown, over time including sequence, age, date, and gender information, of mean severity over time of patient samples with sequences from Alpha (B.1.1.7) and Delta (B.1.617.2) lineages.

vaccination effects, the sequence data are also impacted by artifacts arising from the context of where sequences are obtained. Because cases of confirmed COVID-19 were likelier to be hospitalized early in the pandemic, earlier cases are likelier to be categorized as Severe. Also, clinical data suggests that as vaccination has increased, more patients classified as "hospitalized" may have only tested positive on admission but have mild or no symptoms.[41]

Sample collection, therefore, can represent extrinsic factors with a substantial influence on observed case severity for which we lack sufficient data. Accordingly, we use sample collection date as an input variable to the model, along with age and gender.

Table 1 shows the classification metrics of the deep learning model on the validation data set, which was constructed to exclude any sequence found in the training data set. Overall, the deep learning model is able to correctly predict a substantial majority of both the Mild and Severe classes, i.e., perform better than chance on both classes. Including sequence and demographic variables together in the model led to better metrics than considering each alone. We also benchmarked the deep learning model with the random forests (RF) algorithm, which is a common machine learning method for sequence classification.[42] We expect RF to perform well since most of the differences between spike protein sequences, particularly at early time points in the pandemic, are at specific sequence positions. Despite RF's advantages, the error for the deep learning model approaches that of the RF classifier.

Fig. 2D shows predictions made by training the deep learning model. We simulated the trained model for dates starting in March 2020, well before either Alpha (early 2021) and then later Delta (mid-2021) were prevalent and their sequences had emerged (September 2020 and December 2020 respectively).[2,44] The model qualitatively predicts that Delta has a higher risk of causing severe disease, when controlling for age and sample collection date, in contrast to the empirical analysis of the source data summarized above.

Table 1. Classification accuracy for each class on validation data (class-specific and weighted average values shown). Demographic ("Demo") variables are age, date, and gender. RF was performed using the scikit-learn package in Python 3.8.[43] RF hyperparameters (after tuning) are 500 estimators, no maximum depth, 2 minimum number of samples required to split an internal node, and no limit on features considered for best split.

		Deep Learning			Random Forest
		Demo Only	Sequence Only	Sequence + Demo	Sequence + Demo
Precision	Mild	0.53	0.58	0.61	0.64
	Severe	0.75	0.69	0.74	0.76
	Average	0.66	0.65	0.69	0.71
Recall	Mild	0.72	0.52	0.63	0.66
	Severe	0.56	0.74	0.73	0.75
	Average	0.63	0.65	0.69	0.71
F_1 score	Mild	0.61	0.55	0.63	0.65
	Severe	0.64	0.72	0.73	0.75
	Average	0.63	0.65	0.69	0.71

Using Embeddings to Visualize the Effect of Sample Variables on Classification Results

As Fig. 1 shows, the N_H-dimensional embedding layer is the layer immediately before the final sigmoid-activated node that outputs the class probability (i.e., probability that a sample is from a severe case) which is rounded to the Mild or Severe (0 or 1) prediction. Therefore, at the N_H-dimensional layer, embedded samples will be separated in the embedding space in such a manner that they can be most accurately classified based on the trained model parameters. The t-Distributed Stochastic Neighbor Embedding (t-SNE) transformation[45] of the embeddings should their accentuate similarities and distances, thereby showing the separation and clustering of the embeddings according to how they will be classified.

Fig. 3 shows consistent patterns within the relative positions of sample embeddings according to age, sample date, and sequence distance. The exception is gender, which fails to show separation between male and female samples. The lack of a pattern for gender in Fig. 3 is consistent with Fig. 2, which shows that the effect of gender on disease severity is less significant than age or sample date (58% severe for male versus 53% for female patients). By contrast, age and sample date do play an important role. Fig. 3 further shows that the patterns in modeled embeddings for age and date are orthogonal, in accord with Figs. 2A and 2C showing that, in the data, severity trends with age and date in opposite directions. Fig. 3 also shows that sequence is playing a role in the model prediction. Fig. 3 shows a pattern of separation and clustering among samples by sequence similarity clockwise, similar to sample date, reflecting the empirical trends in Fig. 2.

Using Attention to Identify Sequence Features Important for the Trained Model

Fig. 4 shows a the mean attention obtained from running the validation samples through the model plotted across the spike protein sequence. Many high attention positions coincide with sites previously identified as potentially significant for virulence, infectivity, and/or immune evasion. For example, at the third-highest position, protein structure analysis suggests that G142D and deletions at 143-144, present in some Delta and Delta sub-lineage sequences, may be linked to higher infectivity and immune evasion.[47] The highly ranked sites also include locations known to be significant for differentiating SARS-CoV-2 variants, such as positions 69 (corresponding to the 68-69 deletion found in B.1.1.7/Alpha), 501 (where the N501Y mutation occurs, which is characteristic of B.1.1.7/Alpha and found in some Delta variants), and 681 (P681R and P681H found in Delta and Alpha respectively).[46] By contrast, another site with frequent mutations, site 484, where E484K, E484Q, and E484A mutations associated with antibody escape have been found[48], does not have high attention. The low attention at position 484 in the disease severity model may suggest that the site does not have a significant role in determining virulence rather than antibody evasion.[48–50]

The site with the highest mean attention, position 1258, is in the cytoplasmic tail domain of the S2 subunit, which anchors the spike protein in the plasma membrane.[51] It has not been as well-characterized. The most common mutation found in our training and validation data at that site was E1258D, although we also found E1258H. E1258D is associated with much higher disease severity. Among the 2.6% of samples with an E1258D mutation, 99% were categorized as severe cases. (Our model correctly predicted that 92% of them would be severe.) Another report indicated that E1258D had been found in Delta sublineages, but without additional analysis of its potential impact on patient outcomes or transmissibility. E1258D is not lineage specific though: 38.7% of the E1258 in the patient data set were from B.1.1.519, 18.1% were P.1 (Gamma), 14.8% B.1.1.7 (Alpha), and 11.9% were B.1.617.2 (Delta) and Delta sublineages (primarily AY.3 and AY.20).

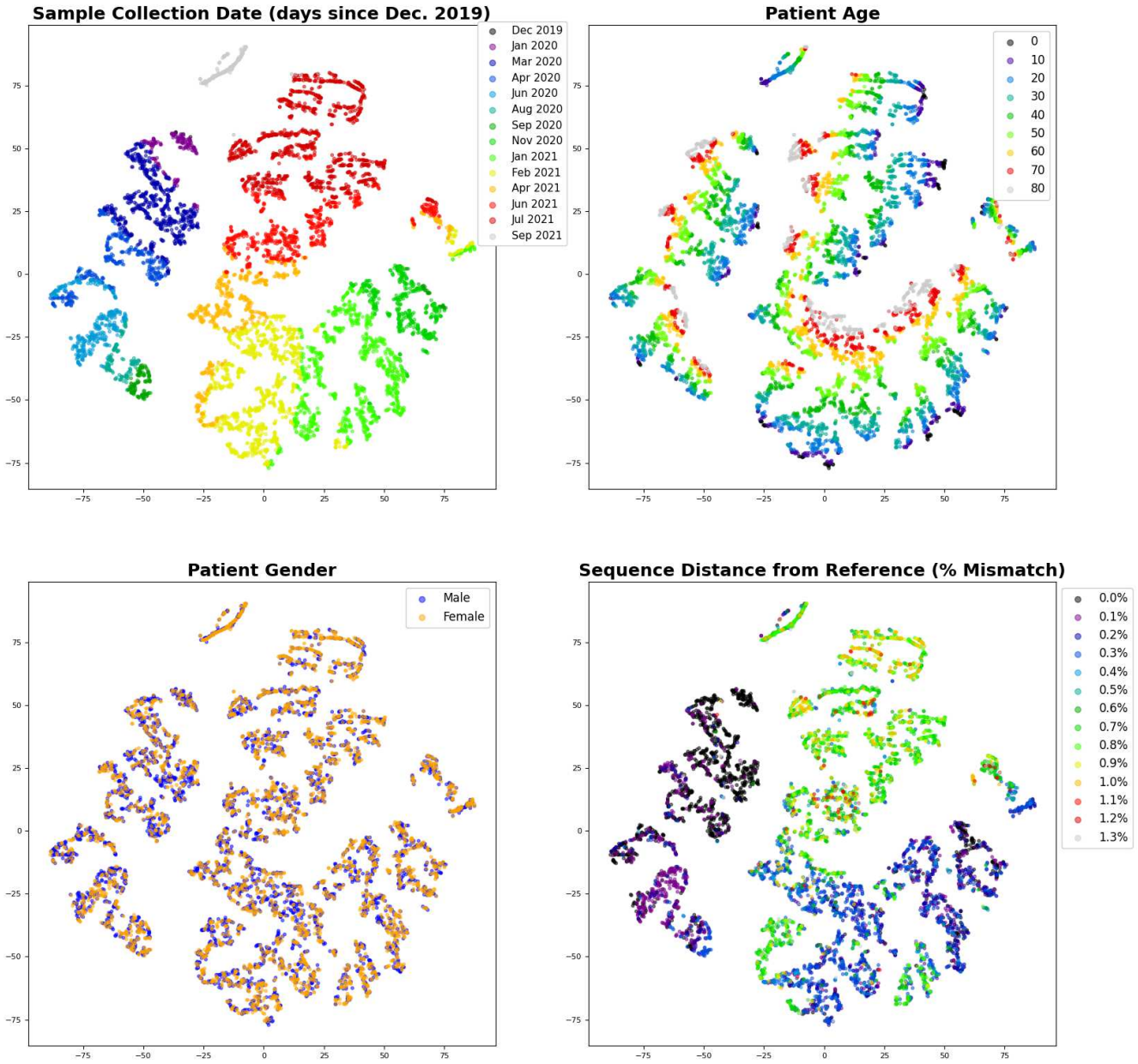

Figure 3. Sequence embeddings obtained from the trained deep learning model and visualized based on a 2-component t-SNE (perplexity of 30). The plots show the same random sampling of 150 samples each of the 45 most prevalent lineages to avoid overrepresenting particular sequences or sample variable values. Clockwise starting with the upper left, the graphs show the t-SNE distribution labeled by binned sample collection date, patient age, sequence distance, and gender respectively. Sequence distance is the number of mismatches between the sequence of the sample to the reference protein sequence divided by the reference protein sequence length, disregarding any padding or ambiguous amino acids in the sample sequence.

Predicting Omicron disease severity

As discussed above, since Omicron has emerged at a later date in the pandemic than other variants, we expect it to be less severe than if the same variant had emerged earlier in the pandemic. We must therefore control for age and date, as shown Fig. 2D, to compare Omicron's relative severity to other variants. We run the trained model described in the preceding with the same age, date, and gender inputs as predictions for other variants to compare predicted severity. Fig. 5A shows how the predicted disease severity of the Omicron sequences would vary over time through the pandemic. As Fig. 5A illustrates, Omicron results in a significantly lower severity than Delta (on average between 40 to 50% over time). Fig. 5A further shows a reduction in predicted Omicron severity as compared to the Alpha variant as well.

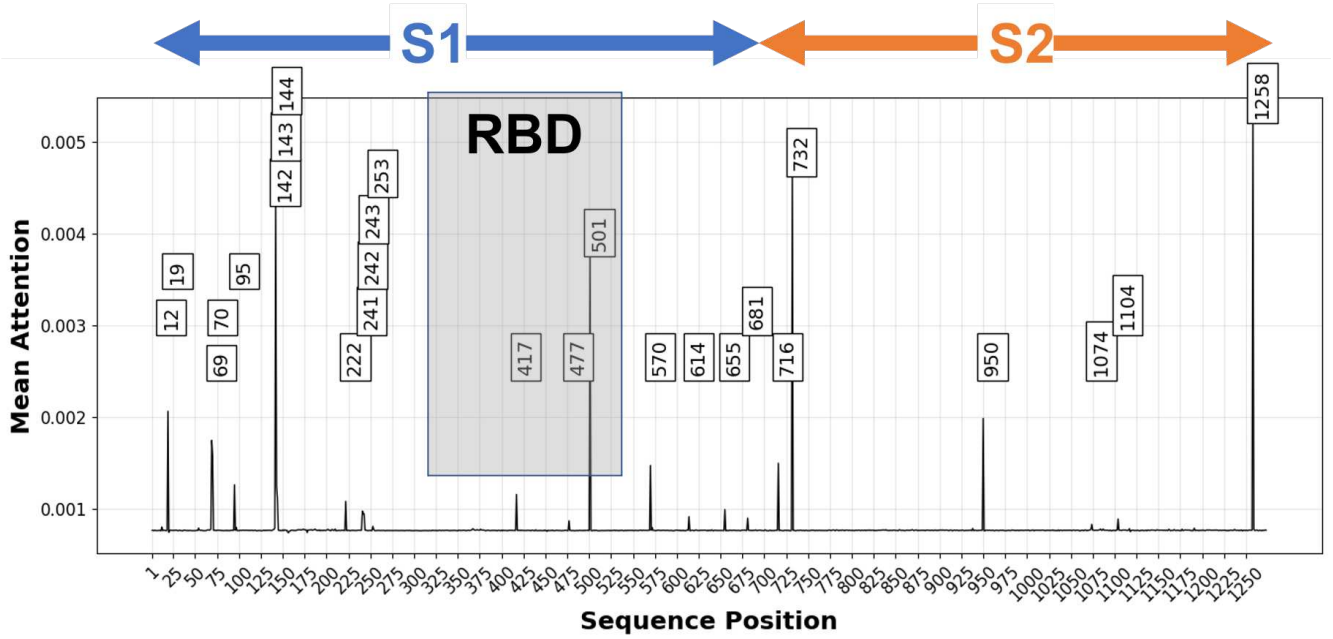

Figure 4. Mean attention at each spike protein sequence position of validation samples. Labeled positions have attention scores above 0.0008. The graph approximate locations for the S1 and S2 subunits, which is cleaved at a furin cleavage site (positions 680-689) by host proteases leading to membrane fusion and host cell entry.[15] The receptor binding domain (RBD) (391-541) is also shown.[46] Many high attention sites are clustered to the N-terminal domain (NTD) and C-terminal domain (CTD) of the S1 subunit.

Fig. 5B is a differential attention plot that shows the spike protein sequence locations that have the greatest difference in attention score between the most prevalent Delta and Omicron sequences. High differential attention at a position implies that the model found it to be particularly relevant in predicting that Omicron would likely be less severe than Delta (controlling for other sample variables). Sites with the greatest positive attention difference are 69-70 (where Omicron has a deletion also present in Alpha but not Delta), 95 (the most common Omicron sequence type, shown here, lacks Delta's T95I mutation), 143-144 (Omicron has a deletion), 211 (Omicron has a deletion), 417 (K417N is found in some Delta lineages), and 477 (Omicron lacks S477N). Sites with the greatest negative attention difference are 19 (T19R found in some but not all Delta sequences), 142 (Omicron lacks Delta's D142G), 501 (N501Y also found in Alpha) and 950 (Omicron lacks N950D). Further study is required to determine the potential effect of the spike protein sites and mutations identified in Fig. 5B on biological mechanisms that could affect infectivity or cell-cell fusion in ways that would impact the risk of severe outcomes.

Discussion

A key impediment to deep learning is the risk of overfitting due to insufficient training data. When models overfit, they can only explain training data—and unable to make predictions on sequences or patients that the model has not seen before. This is a particular challenge in biomedical problems, where it is difficult and costly to obtain clinical and experimental data.[52] To get as much data as possible, we utilize the GISAID database, a collection of millions of SARS-CoV-2 sequences. Even so, fewer than 45,000 samples (as of October 2021) include sufficient metadata for modeling. Sequence records also do not include additional metadata that could help account for case severity, such as comorbidities and racial disparities.[53] The severity of cases may be related to different countries and regions' enforcement of non-pharmaceutical interventions (NPI), social mobility, travel exposure, level of circulating virus, and differences in hospital capacity and standards of care.[54] Because sequences will be similar when they spread across a geographical region,[13] a sample's country of origin may be a confounder. However, there are too few samples with patient data originating in individual countries or regions to enable machine learning based on geographic variables. GISAID data also reflect regional inequities in sequencing. For example, vaccination has a disproportionate influence on GISAID data given worldwide vaccination levels, because countries in which vaccines were widely available among the elderly in early 2021, and general population through 2021, are highly overrepresented in GISAID. Of all sequences in GISAID through the October 2021 cutoff for this paper, 55% were in Europe and 34% were from North America, and of the subset with patient metadata, 69% were from Europe and North America combined.

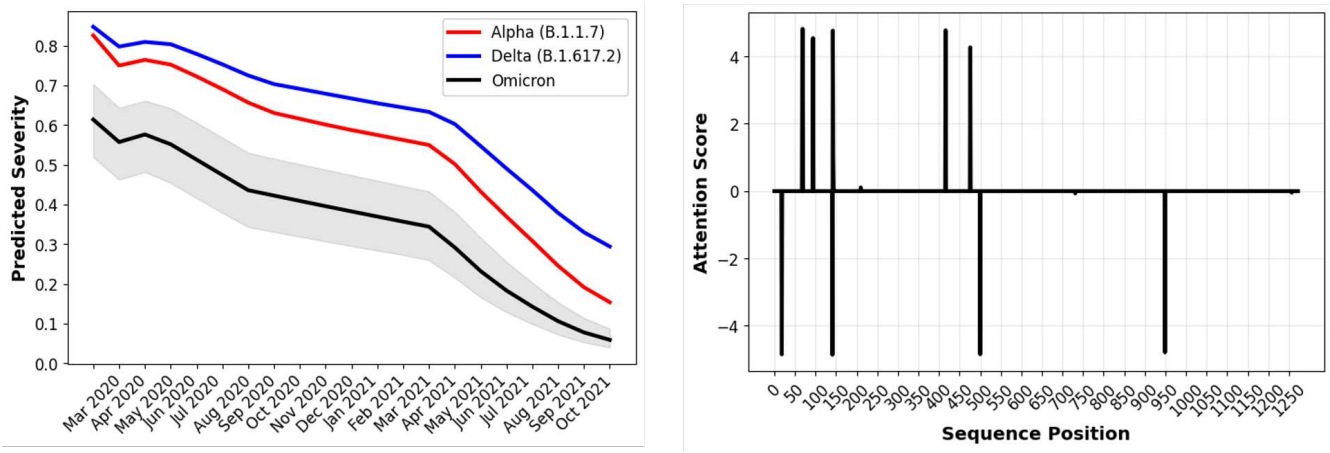

Figure 5. **(A – Left)** Model prediction of the average disease severity for Alpha, Delta, and Omicron over time, assuming a 50-year-old male patient. Omicron is predicted to have, on average, a 40-50% reduced risk of severe disease as compared to Delta over time. The gray shaded region represents the range of predictions for the various Omicron-designated sequences. **(B – Right)** Differential attention plot showing spike protein sequence positions where the greatest difference in attention score between the most prevalent Delta and Omicron sequences.

Despite data limitations, incorporating patient demographics to the model input along with sequence allows the model to predict patient severity from GISAID data at nearly 70% accuracy—in a validation data set with sequences distinct from training data. As this prediction accuracy is only slightly lower than for Random Forests, it is unlikely that the model could be more quantitatively accurate without including variables we do not have data for, such as patients' preexisting conditions, vaccination status, and variability in their immune response. The model can also predict that Delta is more severe than Alpha, in agreement with clinical observations.[36,38,39] Moreover, using attention to interpret the basis for the model's prediction, we found that mutations at site 1258 of spike may be highly correlated with increased disease severity. Conventional methods of studying specific lineages likely would fail to reveal this, as E1258D can appear or revert in sequences in the same lineage. Although not observed in our data, the E1258Q mutation at this site has been linked to spike protein accumulating in the plasma membrane resulting in syncytia, large multinuclear cellular masses associated with heightened virulence.[51] While it is possible that the E1258D mutation's apparent importance from our deep analysis of the GISAID patient data set may be an artifact of the limited data we analyzed, the strong signal suggests that it may warrant further investigation.

We demonstrate the ability of the model to avoid overfitting and generalize to novel variants by making predictions for Omicron. Using our trained model, which excludes any sequences collected later than October 2021—weeks before Omicron emerged—we predict that Omicron may cause substantially less risk of causing severe disease than the Delta variant, and other VOC as well (e.g., Alpha). Importantly, Fig. 5A shows that had Omicron emerged earlier in the pandemic, the variant's apparent severity would have been much greater, albeit less than other variants. Omicron is also highly transmissible, significantly more so than Delta.[55] Given the apparently dramatically increased infectivity of Omicron, it may still result in widespread severe disease, especially among vulnerable and unvaccinated populations.

While still capable of causing significant disease in the unvaccinated and particularly vulnerable, the reduced severity of Omicron as compared to Delta is consistent with early empirical data. Laboratory evidence shows a reduction in lower lung infectivity, deficient cell entry, and a reduction in syncytium formation due to reduced ability of the spike protein to mediate plasma membrane fusion.[56,57] Omicron has also been found to be less infective in the lung and cause less severe disease in a Syrian hamster model.[58,59] The decrease in severe disease is also consistent with preliminary clinical and epidemiological evidence that had emerged by the time of this manuscript's submission, suggesting that patients had less severe disease in South Africa.[60] Reduced severe disease in South Africa may be due to a particularly high degree of seroprevalence in South Africa as a result of prior infection and vaccination of the vulnerable population. However, early data from the United Kingdom also suggest a 20-25% reduced level of hospitalization of any kind and 40-45% reduction in hospitalizations of greater than one day.[61] A recent report of medical records in the United States showed that as Omicron replaced Delta as the predominant variant, there was a greater than 50% reduction in 3-day risks of various indicators of severity, including hospitalization, ICU admission, and mechanical ventilation.[62] The model's prediction that Omicron may be more severe than other VOC has also been supported by early reports. It is difficult to compare clinical and epidemiological evidence of severity for variants earlier in time, because of the effect of vaccination, previous infections, and improved therapies. However, the prediction of reduced severity is consistent with the preliminary data showing that Omicron may have deficient cell entry and less induction of

cell-cell fusion as compared to wild type (ancestral genome) as well as Delta.[56] Free energy calculations also suggest that the Omicron spike protein has deficient host cell receptor binding as compared to wild type or Delta.[63]

The ability of the model to make validated predictions for Omicron, despite Omicron's novelty and distance from previously observed SARS-CoV-2 sequences, illustrates the generalizability of our deep learning framework for protein sequence analysis. We provide the proof of concept for a computational modeling framework that can provide predictive insight on the properties of SARS-CoV-2 variants that may emerge in the future, before empirical data for such variants become available. Future work will likely need to account for variant sequences in other genes, such as nucleocapsid genes, which may have a significant impact on properties like virulence as well.[64] We also note that while there is an unprecedented amount of sequence data available for SARS-CoV-2, the tasks demonstrated here in large part only used a small fraction. This was due to the lack of metadata beyond geographic origin and age/gender for the vast majority of sequences. An important objective of sequencing work should be to collect and curate important information about the sample and to meet minimum information about a sequence standards.[65] However, even with data limitations, the deep protein sequence modeling framework described here can provide insight to SARS-CoV-2, demonstrating the potential to apply it in other biological contexts.

Methods

Sequence Collection and Pre-Processing

We download a FASTA file of protein sequences and corresponding metadata annotation file (including *inter alia* collection date, Pango designation, geographical information, patient status where available) from the GISAID database (`http://www.gisaid.org`). Aside from the Omicron sequences, the protein sequences we use for our analyses shown are those submitted and preprocessed by GISAID as of October 16, 2021. Spike protein sequences are preprocessed by GISAID by multiple sequencing alignment, identifying ORFs, and translating nucleotide sequences to obtain protein FASTA files. We parse the FASTA file to obtain a file with only the Spike protein sequences.

Aligned (with respect to reference) sequences are generated using the local pairwise Striped Smith-Waterman (SSW) method[43,66], with BLOSUM62, in the scikit-bio package in Python 3.8.[67] Following alignment, we obtain a protein sequence in which all insertions and deletions are positioned to correspond to the consensus Spike reference sequence (Wuhan-Hu-1 isolate) obtained by multiple sequence alignment of early genome sequences.[68] Aligned sequences shorter than the reference (1273 residues) are front and/or end padded with a "*", and otherwise all indels are at positions corresponding to the reference. To preserve as many samples as possible, we do not filter sequences with "*" (mask) or "X" (ambiguous amino acid).

To obtain Omicron sequences, we downloaded Spike sequences from a composite FASTA file of the whole genome nucleotide sequence of sequences available on GISAID and identified by lineage metadata as "B.1.1.529" or "Omicron" available as of December 9, 2021. We perform pairwise alignments between each individual Omicron nucleotide sequence with the spike nucleotide sequence from the annotated NCBI Reference Sequence (GenBank record NC-045512) available for download from NCBI at `https://www.ncbi.nlm.nih.gov/sars-cov-2/` using the local pairwise Striped Smith-Waterman (SSW) method as described above (but with nucleotide rather than amino acid sequence) and translate non-degenerate aligned nucleotide sequences to protein sequences.

Patient Status Metadata Pre-Processing

We found 155,545 sequence records with "patient metadata" in the GISAID database. Patient metadata on GISAID consists of a single field with text provided by the submitter of the sequence. We assign raw text in the metadata field of each record to one of the following categories, according to the scheme set forth in Supplementary Table S1: Alive, Asymptomatic, Dead, Hospitalized, Mild, Moderate, Released, Screening, Severe, Symptomatic, and Unknown (which includes any other categories identified in Table S1, such as "Vaccinated"). These categories follow the commonly used case classification such as those defined by the United States National Institutes for Health (NIH) COVID-19 guidelines.[69] For example, metadata indicating ICU admission is categorized as "Severe." We remove all data designated as Alive, Symptomatic, and Unknown, as they cannot be further classified as "Mild" or "Severe." For example, a "Symptomatic" or "Alive" patient may have severe symptoms or have been hospitalized. We also remove records not identified as human by the "host" metadata field. We designate the following categories as "Mild" (integer encoded as 0): Asymptomatic, Mild, Moderate, and Screening (assumed to be other than hospitalization for COVID or symptom reporting). The following categories are assigned to "Severe" (1): Dead, Severe, Hospitalized, and Released (i.e., after hospitalization). The last category indicates prior hospitalization. Finally, we also remove all samples not explicitly identified as being from a human patient in the "Host" metadata field of the database record.

We also integer encode patient age, date, gender for model input. We process each sample's age metadata to remove any unknowns or text entries which cannot be identified as an age. We remove any samples with a missing or unintelligible date metadata. We assign samples with incomplete dates to the latest possible date (e.g., "2021-01" to "2021-01-31"). We normalize the original sample collection dates by setting Day 0 to December 1, 2019, which shortly predates the first sequenced samples

from COVID-19 patients. For gender, we we remove all samples with unknown or ambiguous gender metadata and encode male and female as 0 and 1 respectively.

Metadata records are then merged with the aligned sequence data set obtained from GISAID, preprocessed as described above. Each sample will be represented by two vectors, the tokenized sequence and integer encoded demographic information. The training data set is made up of samples that were available as of September 12, 2021 that were found at least five times overall in the data to that point. (Note that we refer here to the date of availability on the database; there is often a lag between the sample collection date and when it is submitted to GISAID.) All remaining samples found up to October 1, 2021 dataset from GISAID with available patient metadata make up an independent validation set, which includes samples with sequences that were available on GISAID after September 12 or not found at least five times prior. Hence, there will be no exact sequence overlap between training and test samples.

Implementation

Model Parameters

The sequence length is set at 1273, which is the length (as shorter sequences are padded by masks). We tokenize each amino acid, and the deletion symbol "-", to a distinct nonzero integer. A position with padding mask "*" or ambiguous amino acids represented as X, B, J, or Z are assigned to zero. Positions with a token value of zero is masked at input. For the model as shown in Fig. 1, the embedding dimension N_E is set to 1500, the CNN has N_C of 300 filters, the Transformer block contains 8 attention heads, a feedforward network (FFN) of 64 dimensions, with dropout of 0.1; N_H is 64, and no other dropout is added. To obtain an optimal combination of hyperparameters, we performed cross-validation on a subset of the training data set and examined hyperparameters varying as follows: number of Transformer heads between 4, 6, and 8; number of FFN nodes between 32, 64, and 96 nodes; transformer dropout between 0.0, 0.1, and 0.2; dropout of 0.0, 0.1, and 0.2 after the embedding layer; N_C at 300, 400, and 500; N_H at 64 or 128; and N_E of 1000, 1200, 1500, and 1800.

Training

Model training was done using the standard Adam gradient descent fitting algorithm in Tensorflow 2, with binary cross-entropy and mean squared error as the loss functions for classification and regression tasks respectively. We set the learning rate parameter to 1×10^{-4}, after evaluating rates of 5×10^{-5}, 1×10^{-5}, 1×10^{-3}, 5×10^{-4}, 1×10^{-2}, and 1×10^{-1}. We train for 75 epochs, after evaluating epoch ranges from 25 to 200. We do not use early stopping with a validation set, as that reduces the training data and causes underfitting. Instead, early stopping is set at 20 epochs with no net change in the binary cross-entropy loss. We also used multiple runs to evaluate run-to-run consistency given the randomization of initial weights and other inherent randomization of training. Where not shown, standard deviations across runs were not greater than reported precision.

Training is performed using the standard model fitting algorithm provided by the Keras API for Tensorflow 2. The hardware for deep learning model training and evaluation, as well as the data pre-processing described above, consists of Nvidia Tesla P80 GPUs (primarily) and Google Cloud Tensor Processor Units (TPUs) the Google's Colab environment, running Tensorflow 2.70 and Python 3.7.12, and Nvidia Tesla V100-SXM2 GPUs, running Tensorflow 2.4 and Python 3.8. Training on 44,003 samples required 51 sec/epoch in the Google TPU environment and 480 sec/epoch on a GPU unit, representing a 9.2-fold TPU-speedup. We provide the source code for pre-processing and model implementation at `https://github.com/bahrad/Covid`.

Ethics Statement

This study involves only openly available human data, which can be obtained from the GISAID's EpiCoV database, which is available at `http://www.gisaid.org`. As this study involves deidentified data made available for public use, it does not require Institutional Review under 45 CFR 46.102.

References

1. Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance* **22**, 30494, DOI: 10.2807/1560-7917.ES.2017.22.13.30494 (2017).

2. Tao, K. *et al.* The biological and clinical significance of emerging SARS-CoV-2 variants. *Nat. reviews. Genet.* 1–17, DOI: 10.1038/s41576-021-00408-x (2021).

3. Bernasconi, A., Mari, L., Casagrandi, R. & Ceri, S. Analysis of amino acid change dynamics reveals SARS-CoV-2 variant emergence. *bioRxiv* DOI: 10.1101/2021.07.12.452076 (2021).

4. Hadfield, J. *et al.* Nextstrain: Real-time tracking of pathogen evolution. *Bioinforma. (Oxford, England)* **34**, 4121–4123, DOI: 10.1093/bioinformatics/bty407 (2018).

5. Martin, D. P. *et al.* The emergence and ongoing convergent evolution of the SARS-CoV-2 N501Y lineages. *Cell* **184**, 5189–5200.e7, DOI: 10.1016/j.cell.2021.09.003 (2021).

6. O'Toole, Á. *et al.* Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus evolution* **7**, veab064, DOI: 10.1093/ve/veab064 (2021).

7. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407, DOI: 10.1038/s41564-020-0770-5 (2020).

8. Parums, D. V. Editorial: Revised World Health Organization (WHO) Terminology for Variants of Concern and Variants of Interest of SARS-CoV-2. *Med. Sci. Monit. : Int. Med. J. Exp. Clin. Res.* **27**, e933622–1–e933622–2, DOI: 10.12659/MSM.933622 (2021).

9. Tasakis, R. N. *et al.* SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts. *PLOS ONE* **16**, e0255169, DOI: 10.1371/journal.pone.0255169 (2021).

10. Baj, A. *et al.* Spike protein evolution in the SARS-CoV-2 Delta variant of concern: A case series from Northern Lombardy. *Emerg. Microbes & Infect.* **10**, 2010–2015, DOI: 10.1080/22221751.2021.1994356 (2021).

11. Chen, L. *et al.* Emergence of Multiple SARS-CoV-2 Antibody Escape Variants in an Immunocompromised Host Undergoing Convalescent Plasma Treatment. *mSphere* **6**, e0048021, DOI: 10.1128/mSphere.00480-21 (2021).

12. Qin, L. *et al.* Co-mutation modules capture the evolution and transmission patterns of SARS-CoV-2. *Briefings bioinformatics* DOI: 10.1093/bib/bbab222 (2021).

13. Zhao, Z., Sokhansanj, B. A., Malhotra, C., Zheng, K. & Rosen, G. L. Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization. *PLoS computational biology* **16**, e1008269, DOI: 10.1371/journal.pcbi.1008269 (2020).

14. Pardo-Seco, J., Gómez-Carballa, A., Bello, X., Martinón-Torres, F. & Salas, A. Pitfalls of barcodes in the study of worldwide SARS-CoV-2 variation and phylodynamics. *Zool. research* **42**, 87–93, DOI: 10.24272/j.issn.2095-8137.2020.364 (2021).

15. Pillay, T. S. Gene of the month: The 2019-nCoV/SARS-CoV-2 novel coronavirus spike protein. *J. Clin. Pathol.* **73**, 366, DOI: 10.1136/jclinpath-2020-206658 (2020).

16. Vankadari, N. Overwhelming mutations or SNPs of SARS-CoV-2: A point of caution. *Gene* **752**, 144792, DOI: 10.1016/j.gene.2020.144792 (2020).

17. Ren, L. *et al.* Genetic drift of human coronavirus OC43 spike gene during adaptive evolution. *Sci. Reports* **5**, 11451, DOI: 10.1038/srep11451 (2015).

18. Cotten, M., Robertson, D. L. & Phan, M. V. T. Unique protein features of SARS-CoV-2 relative to other Sarbecoviruses. *Virus evolution* **7**, veab067, DOI: 10.1093/ve/veab067 (2021).

19. Liu, Y. & Rocklöv, J. The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *J. Travel. Medicine* **28**, taab124, DOI: 10.1093/jtm/taab124 (2021).

20. Liu, Y. *et al.* Delta spike P681R mutation enhances SARS-CoV-2 fitness over Alpha variant., DOI: 10.1101/2021.08.12.456173 (2021).

21. Mlcochova, P. *et al.* SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114–119, DOI: 10.1038/s41586-021-03944-y (2021).

22. Karim, S. S. A. & Karim, Q. A. Omicron SARS-CoV-2 variant: A new chapter in the COVID-19 pandemic. *The Lancet* **398**, 2126–2128, DOI: 10.1016/S0140-6736(21)02758-6 (2021).

23. Wang, L. & Cheng, G. Sequence analysis of the Emerging Sars-CoV-2 Variant Omicron in South Africa. *J. Med. Virol.* DOI: 10.1002/jmv.27516 (2021).

24. Luong, M.-T., Pham, H. & Manning, C. D. Effective Approaches to Attention-based Neural Machine Translation (2015). 1508.04025.

25. Watson, D. S. Interpretable machine learning for genomics. *Hum. genetics* 1–15, DOI: 10.1007/s00439-021-02387-9 (2021).

26. Raffel, C. & Ellis, D. P. W. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. *arXiv:1512.08756 [cs]* (2016). 1512.08756.

27. Zhou, P. *et al.* Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 207–212 (2016).

28. Lanchantin, J., Singh, R., Lin, Z. & Qi, Y. Deep Motif: Visualizing Genomic Sequence Classifications. *CoRR* **abs/1605.0** (2016).

29. Zhao, Z. *et al.* Learning, visualizing and exploring 16S rRNA structure using an attention-based deep neural network. *PLoS computational biology* **17**, e1009345, DOI: 10.1371/journal.pcbi.1009345 (2021).

30. Kobayashi, G., Kuribayashi, T., Yokoi, S. & Inui, K. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. *arXiv:2004.10102 [cs]* (2020). 2004.10102.

31. Clauwaert, J., Menschaert, G. & Waegeman, W. Explainability in transformer models for functional genomics. *Briefings bioinformatics* **22**, DOI: 10.1093/bib/bbab060 (2021).

32. Yang, K. K., Wu, Z., Bedbrook, C. N. & Arnold, F. H. Learned protein embeddings for machine learning. *Bioinforma. (Oxford, England)* **34**, 2642–2648, DOI: 10.1093/bioinformatics/bty178 (2018).

33. Gao, Y.-D. *et al.* Risk factors for severe and critically ill COVID-19 patients: A review. *Allergy* **76**, 428–455, DOI: 10.1111/all.14657 (2021).

34. Hamed, S. M., Elkhatib, W. F., Khairalla, A. S. & Noreddin, A. M. Global dynamics of SARS-CoV-2 clades and their relation to COVID-19 epidemiology. *Sci. reports* **11**, 8435, DOI: 10.1038/s41598-021-87713-x (2021).

35. Voss, J. D. *et al.* Variants in SARS-CoV-2 associated with mild or severe outcome. *Evol. medicine, public health* **9**, 267–275, DOI: 10.1093/emph/eoab019 (2021).

36. Bager, P., Wohlfahrt, J., Rasmussen, M., Albertsen, M. & Krause, T. G. Hospitalisation associated with SARS-CoV-2 delta variant in Denmark. *The Lancet. Infect. diseases* **21**, 1351, DOI: 10.1016/S1473-3099(21)00580-6 (2021).

37. Fisman, D. N. & Tuite, A. R. Evaluation of the relative virulence of novel SARS-CoV-2 variants: A retrospective cohort study in Ontario, Canada. *CMAJ : Can. Med. Assoc. journal = journal de l'Association medicale canadienne* **193**, E1619–E1625, DOI: 10.1503/cmaj.211248 (2021).

38. Paredes, M. I. *et al.* Associations between SARS-CoV-2 variants and risk of COVID-19 hospitalization among confirmed cases in Washington State: A retrospective cohort study., DOI: 10.1101/2021.09.29.21264272 (2021).

39. Twohig, K. A. *et al.* Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: A cohort study. *The Lancet. Infect. diseases* DOI: 10.1016/S1473-3099(21)00475-8 (2021).

40. Hsu, S. H., Chang, S.-H., Gross, C. P. & Wang, S.-Y. Relative risks of COVID-19 fatality between the first and second waves of the pandemic in Ontario, Canada. *Int. journal infectious diseases : IJID : official publication Int. Soc. for Infect. Dis.* **109**, 189–191, DOI: 10.1016/j.ijid.2021.06.059 (2021).

41. Calderwood, M. S. *et al.* Policies and practices of SHEA Research Network hospitals during the COVID-19 pandemic. *Infect. Control. Hosp. Epidemiol.* **41**, 1127–1135, DOI: 10.1017/ice.2020.303 (2020).

42. Goldstein, B. A., Polley, E. C. & Briggs, F. B. S. Random forests for genetic association studies. *Stat. Appl. Genet. Mol. Biol.* **10**, 32, DOI: 10.2202/1544-6115.1691 (2011).

43. The scikit-bio development team. Scikit-bio: A Bioinformatics Library for Data Scientists, Students, and Developers (2020).

44. Duong, D. Alpha, Beta, Delta, Gamma: What's important to know about SARS-CoV-2 variants of concern? *CMAJ : Can. Med. Assoc. J.* **193**, E1059–E1060, DOI: 10.1503/cmaj.1095949 (2021).

45. van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).

46. Joshi, N., Tyagi, A. & Nigam, S. Molecular Level Dissection of Critical Spike Mutations in SARS-CoV-2 Variants of Concern (VOCs): A Simplified Review. *ChemistrySelect* **6**, 7981–7998, DOI: 10.1002/slct.202102074 (2021).

47. Shen, L. *et al.* Spike Protein NTD mutation G142D in SARS-CoV-2 Delta VOC lineages is associated with frequent back mutations, increased viral loads, and immune evasion, DOI: 10.1101/2021.09.12.21263475 (2021).

48. Saha, I., Ghosh, N., Sharma, N. & Nandi, S. Hotspot Mutations in SARS-CoV-2. *Front. Genet.* **12**, 753440, DOI: 10.3389/fgene.2021.753440 (2021).

49. Adiga, R. & Nayak, V. Emergence of Novel SARS-CoV-2 variants in India: Second wave. *J. Infect. Dev. Ctries.* **15**, 1578–1583, DOI: 10.3855/jidc.15484 (2021).

50. Jangra, S. *et al.* SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. *The Lancet. Microbe* **2**, e283–e284, DOI: 10.1016/S2666-5247(21)00068-9 (2021).

51. Rocheleau, L. *et al.* Identification of a High-Frequency Intrahost SARS-CoV-2 Spike Variant with Enhanced Cytopathic and Fusogenic Effects. *mBio* DOI: 10.1128/mBio.00788-21 (2021).

52. Jones, D. T. Setting the standards for machine learning in biology. *Nat. Rev. Mol. Cell Biol.* **20**, 659–660, DOI: 10.1038/s41580-019-0176-5 (2019).

53. Wang, Z. *et al.* Hospitalised COVID-19 patients of the Mount Sinai Health System: A retrospective observational study using the electronic medical records. *BMJ open* **10**, e040441, DOI: 10.1136/bmjopen-2020-040441 (2020).

54. Chen, Z. *et al.* A global analysis of replacement of genetic variants of SARS-CoV-2 in association with containment capacity and changes in disease severity. *Clin. microbiology infection : official publication Eur. Soc. Clin. Microbiol. Infect. Dis.* **27**, 750–757, DOI: 10.1016/j.cmi.2021.01.018 (2021).

55. Brandal, L. T. *et al.* Outbreak caused by the SARS-CoV-2 Omicron variant in Norway, November to December 2021. *Euro Surveillance: Bull. Eur. Sur Les Maladies Transm. = Eur. Commun. Dis. Bull.* **26**, DOI: 10.2807/1560-7917.ES.2021.26.50.2101147 (2021).

56. Meng, B. *et al.* SARS-CoV-2 Omicron spike mediated immune escape, infectivity and cell-cell fusion, DOI: 10.1101/2021.12.17.473248 (2021).

57. Zhao, H. *et al.* SARS-CoV-2 Omicron variant shows less efficient replication and fusion activity when compared with delta variant in TMPRSS2-expressed cells. *Emerg. Microbes & Infect.* **0**, 1–18, DOI: 10.1080/22221751.2021.2023329 (2021).

58. Abdelnabi, R. *et al.* The omicron (B.1.1.529) SARS-CoV-2 variant of concern does not readily infect Syrian hamsters, DOI: 10.1101/2021.12.24.474086 (2021).

59. Ryan, K. A. *et al.* Convalescence from prototype SARS-CoV-2 protects Syrian hamsters from disease caused by the Omicron variant, DOI: 10.1101/2021.12.24.474081 (2021).

60. Wolter, N. *et al.* Early assessment of the clinical severity of the SARS-CoV-2 Omicron variant in South Africa, DOI: 10.1101/2021.12.21.21268116 (2021).

61. Ferguson, N., Ghani, A., Hinsley, W. & Erik Volz. Report 50 - Hospitalisation risk for Omicron cases in England. http://www.imperial.ac.uk/medicine/departments/school-public-health/infectious-disease-epidemiology/mrc-global-infectious-disease-analysis/covid-19/report-50-severity-omicron/ (2021).

62. Wang, L. *et al.* Comparison of outcomes from COVID infection in pediatric and adult patients before and after the emergence of Omicron, DOI: 10.1101/2021.12.30.21268495 (2022).

63. Glocker, M. O., Opuni, K. F. M. & Thiesen, H.-J. Compared with SARS-CoV2 wild type's spike protein, the SARS-CoV2 omicron's receptor binding motif has adopted a more SARS-CoV1 and/or bat/civet-like structure, DOI: 10.1101/2021.12.14.472585 (2021).

64. Syed, A. M. *et al.* Rapid assessment of SARS-CoV-2–evolved variants using virus-like particles. *Science* DOI: 10.1126/science.abl6184 (2021).

65. Schriml, L. M. *et al.* COVID-19 pandemic reveals the peril of ignoring metadata standards. *Sci. Data* **7**, 188, DOI: 10.1038/s41597-020-0524-5 (2020).

66. Zhao, M., Lee, W.-P., Garrison, E. P. & Marth, G. T. SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. *PLOS ONE* **8**, e82138, DOI: 10.1371/journal.pone.0082138 (2013).

67. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

68. Wang, C. *et al.* The establishment of reference sequence for SARS-CoV-2 and variation analysis. *J. Med. Virol.* **92**, 667–674, DOI: 10.1002/jmv.25762 (2020).

69. National Institutes of Health. Clinical Spectrum of SARS-CoV-2 Infection (2021).

Acknowledgements

We acknowledge the the Global Initiative on Sharing All Influenza Data (GISAID) EpiCoV database, which has made accessible SARS-CoV-2 sequence data, including from the NIH's Genbank database, as well as the authors and laboratories originating and submitting the GISAID sequences on which this research is based. A list of contributors to the data used in this paper is available at this paper's Github repository. Work reported here was run on hardware supported by Drexel's University Research Computing Facility. GLR received National Science Foundation (NSF) grants #1919691 and #2107108. The funders had no role in study design, deciding to publish, collecting or analyzing data, or preparing the manuscript.

Author contributions statement

BAS conceived and developed the model, analyzed the results, and prepared the manuscript; ZZ. contributed to model development and analysis and assisted in preparing the manuscript; and GLR contributed to model development and analysis and assisted in preparing the manuscript. All authors reviewed the manuscript.

Additional information

Competing interests The authors declare no competing interests.

Supplementary Files

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

- SupplementaryTableS1.pdf