Predicting Variant Fitness of SARS-CoV-2 from Full Viral Genome Sequences

Richard Annan[1], Ursula Nkonu[2], Parisa Hatami[2], Md Jubair Pantho[2], Letu Qingge[1], Hong Qin[3]

[1]Department of Computer Science, North Carolina A&T State University, Greensboro, NC, U.S.A.,
[2]Department of Computer Science & Engineering, University of Tennessee at Chattanooga, Chattanooga, TN, U.S.A.,
[3]School of Data Science, Department of Computer Science, Old Dominion University, Norfolk, VA, U.S.A.,
rkannan@aggies.ncat.edu, unkon001@odu.edu, qxy699@mocs.utc.edu, wzs444@mocs.utc.edu, lqingge@ncat.edu, hqin@odu.edu

Abstract

Accurate prediction of the transmission fitness of emerging SARS-CoV-2 variants is vital for timely public health responses. In this study, we present a deep learning framework that predicts variant fitness from raw genomic sequences using a convolutional neural network (CNN) trained to regress Differential Population Growth Rate (DPGR) values. Our approach achieves high predictive accuracy ($R^2 = 0.9168$, MSE $\approx 1.94 \times 10^{-4}$) on genomic sequences sampled from the USA and Europe. To interpret the model's predictions, we apply SHapley Additive exPlanations (SHAP) to identify nucleotide-level contributions to predicted fitness. Our analysis highlights key mutations in ORF9 (nucleocapsid), ORF2 (spike), ORF5 (membrane), and ORF8 that either enhance or reduce predicted DPGR. Notably, we identify amino acid–altering mutations such as D3L, E484K, N501Y, and V97I as strong positive contributors to fitness, while synonymous or non-coding mutations had more subtle or regulatory effects. These findings validate the potential of sequence-based modeling and interpretable AI to support early detection and prioritization of high-risk variants.

Introduction

The Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2) is responsible for the coronavirus disease (COVID-19) pandemic which resulted in millions of deaths globally (Jha, Brown, and Ansumana 2022). SARS-CoV-2 belongs to the Betacoronavirus group; the same phylogenetic group as the SARS and MERS viruses, which have previously resulted in widespread disease outbreaks (Singh and Yi 2021). As an RNA virus, SARS-CoV-2 evolves rapidly, and this rapid evolution is as a result of the rate at which mutations occur within its genome and spread through various populations (Markov et al. 2023). In late 2020, there was an emergence of new and more severely mutated SARS-CoV-2 variants (Kung et al. 2025). They were characterized by an increase of non-synonymous mutations found mainly in the spike protein (Carabelli et al. 2023). These variants also expressed distinctive phenotypes, particularly in the way they bind to and interact with the immune system and in their transmissibility (Carabelli et al. 2023). As of September 2022, the World Health Organization had declared five of these variants; Alpha, Beta, Gamma, Delta, and Omicron, as variants of concern (VOC) (Ahmad, Fawaz, and Aisha 2022). These VOCs are distinguished by their heightened fitness (i.e., spreading potential in the host population (Ito et al. 2025). Viral fitness is influenced by various factors such as its ability to elude innate immunity, its efficiency in replicating within the host cells and how well it can elude population-level immunity (Ito et al. 2025). Genotype-fitness relationships are important to assess because they can highlight the specific mutations that promote viral fitness (Ito et al. 2025).

We operationalize transmission fitness as a population-level *growth advantage* estimated from surveillance trajectories via the Differential Population Growth Rate (DPGR); see *Methodology* section for the formal definition. The ultimate goal of assessing this relationship is to develop computational models that can make accurate predictions by identifying high-risk variants shortly after they are detected, to control the spread of disease.

The remainder of this paper is structured as follows. Section reviews existing literature on viral fitness estimation and sequence-based predictive models. Section presents the derivation of Differential Population Growth Rate (DPGR), dataset construction from GISAID sequences, and the design of the regression model. Section reports the model's performance and interpretability findings using SHAP. Section analyzes key genomic features contributing to transmission fitness across variants. Finally, Section presents the conclusions and outlines potential future directions.

Related Work

The emergence and rapid evolution of SARS-CoV-2 VOCs has resulted in the need for more complex computational approaches to model viral fitness. Accurately modeling the viral fitness of VOCs is critical for epidemiological forecasting and pandemic response as it enables public health authorities to anticipate variant spread patterns and implement targeted interventions. However, traditional parametric modeling approaches face significant challenges when applied to rapidly evolving viral populations, particularly in accounting for sampling biases and inability to align with real-world viral dynamics (Pantho et al. 2025). To address these limitations, the differential population growth rate (DPGR) model has been proposed as a data-driven approach that uses viral strains as internal controls to reduce sampling bias er-

rors (Pantho et al. 2025). DPGR represents a methodological advancement through its pairwise comparative framework, employing straightforward log-linear regression to estimate relative fitness between two viral strains (Pantho et al. 2025). Its additive distance framework allows it to infer fitness landscapes and construct distance matrices that capture the rapid evolutionary shifts among SARS-CoV-2 variants (Pantho et al. 2025). The increase in artificial intelligence (AI) methods, particularly deep learning, to advance global health prediction and prevention strategies in current times can be attributed to the COVID-19 pandemic (Lytras et al. 2025). AI offers the potential to capture complex, nonlinear relationships (Min, Lee, and Yoon 2017) between genotypes and phenotypes, which traditional models often fail to detect, (Sehrawat, Najafian, and Jin 2023) which is crucial for pandemic prediction. Specifically, convolutional neural networks (CNNs) have immense potential in this area. They consist of convolutional layers that can automatically extract features and identify both local and global features, and can handle 1D input data such as sequences (LeCun, Bengio, and Hinton 2015; Washburn et al. 2021). They also highly adaptable and can be applied on a wide range of input data, and with sufficient data and good design they can perform robustly even on lower-quality datasets (Qin et al. 2019). Hatami et al. developed a Convolutional Neural Network (CNN) model that accurately pinpoints specific nucleotide alterations and their corresponding phenotypic changes compared to traditional analytical methods (Hatami et al. 2024). Unlike traditional genome-wide association studies (GWAS), the CNN model successfully identified common spike gene mutations that GWAS failed to detect (Hatami et al. 2024). This advantage likely stems from CNNs' ability to capture complex, non-linear relationships between sequence features and phenotypic outcomes, whereas GWAS approaches are limited by their assumption of linear associations and independence between genetic variants (Hu, Darabos, and Urbanowicz 2020). CoVFit, a protein language model introduced by Ito et al., uses multitask learning in combination with deep mutational scanning data to predict the fitness of novel SARS-CoV-2 variants, including highly divergent lineages like XBB (Ito et al. 2025). Despite the XBB lineage differing by approximately 15 amino acids from the training data, CoVFit maintained accurate predictions. According to ablation studies, CoVFit's success likely results from its innovative combination of multitask learning with deep mutational scanning data (Ito et al. 2025). This approach suggests that protein language models can capture generalizable evolutionary principles that extend beyond their training data. Donker et al. (Donker et al. 2024) presented an alternative approach that uses only nucleotide information from genomic surveillance to directly estimate SARS-CoV-2 fitness gains based solely on the abundance of single nucleotide polymorphisms. Their method sidesteps lineage-level models and instead derives fitness from the changing prevalence of individual mutations over time. Their method addresses a fundamental challenge in viral fitness estimation: the difficulty of tracking changing linear proportions of particular lineages over time, which traditional fitness estimation methods typically rely upon (Donker et al. 2024). Elkin et al. developed Deep Novel Mutation Search (DNMS), a deep learning approach that uses neural networks to model protein sequences for predicting mutations (Elkin and Zhu 2025). Using the SARS-CoV-2 spike protein as the target, they applied a protein language model to detect novel mutations with a methodological innovation that distinguishes their work from previous studies (Elkin and Zhu 2025). Unlike previous approaches where predictions typically depend on mutations of reference sequences, DNMS implements a 'parent-child mutation' framework where predictions are generated from parent sequences serving as templates for mutation modeling (Elkin and Zhu 2025). Advances in artificial intelligence coupled with growing availability of genomic datasets, will continue to enhance our ability to estimate viral fitness directly from genomic sequences. Sequence-based predictive models offer robust alternatives to traditional epidemiological frameworks by capturing complex, nonlinear relationships between mutations and viral phenotypes.

Methodology

Differential Population Growth Rate (DPGR): Definition and Derivation

Transmission fitness is operationalized as the relative exponential growth advantage between two variants within a fixed region and short time window where log-linear dynamics hold. Let $N_1(t)$ and $N_2(t)$ denote weekly counts (or frequencies). When the log-ratio is approximately linear in time,

$$\log\left(\frac{N_1(t)}{N_2(t)}\right) \;=\; (g_1 - g_2)\,t + C \;=\; \mathrm{DPGR}_{1,2}\,t + C, \quad (1)$$

the slope $\mathrm{DPGR}_{1,2} := g_1 - g_2$ serves as a population-level proxy for relative transmission fitness (positive values indicate variant 1 grows faster than variant 2 in that context).

Following Pantho *et al.* (Pantho et al. 2025), exponential growth within selected observation windows is assumed, characterized by clear log-linear trends. Variant populations (N_1 and N_2) with respective growth rates g_1 and g_2, and initial populations $N_{1,0}$ and $N_{2,0}$, are modeled as

$$N_1(t) = N_{1,0}\, e^{g_1(t-T_1)}, \qquad N_2(t) = N_{2,0}\, e^{g_2(t-T_2)}, \quad (2)$$

where t is observational time and T_1, T_2 allow temporal phase offsets. The ratio becomes

$$\frac{N_1}{N_2} = \frac{N_{1,0}}{N_{2,0}}\, e^{(g_1-g_2)t+(-g_1T_1+g_2T_2)} = C_1\, e^{(g_1-g_2)t+C_2}, \quad (3)$$

with constants C_1, C_2 absorbing initial conditions and offsets. Taking logarithms yields

$$\log\left(\frac{N_1}{N_2}\right) \;=\; (g_1 - g_2)\,t + C, \quad (4)$$

so the DPGR equals the slope $(g_1 - g_2)$ of the linear fit to the log-ratio.

Practical estimation and sign convention $\mathrm{DPGR}_{1,2}$ is estimated via ordinary least squares on $\log(N_1/N_2)$ within sliding windows of W weeks (typically $W \in [4, 12]$), retaining windows that satisfy $R^2 > 0.9$ and $p < 0.05$ for the slope. Throughout, the supervised target is $y = \mathrm{DPGR}_{\text{ref, target}}$ computed per region and window; by convention $y > 0$ indicates the *target* variant grows faster than the *reference*.

Sampling and population context. Because DPGR is defined on *ratios*, many *nondiscriminatory* sampling effects approximately cancel, offering robustness to common genomic-surveillance biases. DPGR remains a *population-level* proxy that implicitly aggregates contemporaneous immunity, interventions, and testing practices.

Indirect Estimation of DPGR Due to sampling constraints, direct comparative observations between certain variant pairs were not always feasible. To address this, indirect estimation of DPGR was employed by leveraging intermediate variants. Utilizing the additive property of logarithmic transformations, indirect DPGR estimation is mathematically represented as:

$$\mathrm{DPGR}_{A,C} = \mathrm{DPGR}_{A,B} + \mathrm{DPGR}_{B,C} \tag{5}$$

This relationship was essential in estimating DPGR for variants without direct observational overlaps. For instance, in cases where direct co-sampling of the Alpha and Delta variants was limited, an intermediate Beta variant allowed indirect DPGR estimation through:

$$\mathrm{DPGR}_{\text{Alpha, Delta}} = \mathrm{DPGR}_{\text{Alpha, Beta}} + \mathrm{DPGR}_{\text{Beta, Delta}} \tag{6}$$

This approach significantly enhanced the method's robustness and expanded its applicability, particularly in real-world surveillance contexts marked by incomplete or asynchronous data. The detailed derivation and indirect estimation methods, as proposed by Pantho *et al.* in (Pantho et al. 2025), provide a comprehensive and statistically robust framework for accurately estimating comparative transmission fitness among SARS-CoV-2 variants.

Application of indirect DPGR estimation The indirect estimation method described above was practically implemented to derive DPGR values presented in Table 1 and Table 2. This method was particularly critical in cases where direct co-sampling was limited, such as for the Alpha and Delta variants. Leveraging the Beta variant as an intermediate, the DPGR value for Delta versus Alpha was calculated using the additive property of DPGR:

$$\mathrm{DPGR}_{\text{Delta, Alpha}} = \mathrm{DPGR}_{\text{Delta, Beta}} + \mathrm{DPGR}_{\text{Beta, Alpha}} \tag{7}$$

From the USA dataset provided (Table 1), the following calculation explicitly illustrates this indirect estimation process:

- DPGR (Beta vs. Alpha): 0.002825 (direct regression)
- DPGR (Delta vs. Beta): 0.006088 (direct regression)
- DPGR (Delta vs. Alpha): $0.002825 + 0.006088 = 0.008913$ (indirect estimation)

Similarly, for Omicron relative to Alpha, the DPGR was estimated indirectly through sequential intermediate variants:

- DPGR (Omicron vs. Delta): 0.009315 (direct regression)
- DPGR (Delta vs. Alpha): 0.008913 (previously computed indirectly)
- DPGR (Omicron vs. Alpha): $0.009315 + 0.008913 = 0.018228$ (indirect estimation)

The neighbor-joining method proposed by Saitou and Nei (Saitou and Nei 1987) was utilized by Pantho e*t al.* in (Pantho et al. 2025) to construct phylogenetic relationships among these variants based on computed DPGR distances, effectively visualizing evolutionary trajectories and relationships among variants as shown in Figure 1 for USA. Thus, the indirect estimation technique significantly augmented the robustness and interpretability of DPGR measures, enabling comprehensive analysis even in scenarios with incomplete direct co-sampling data.

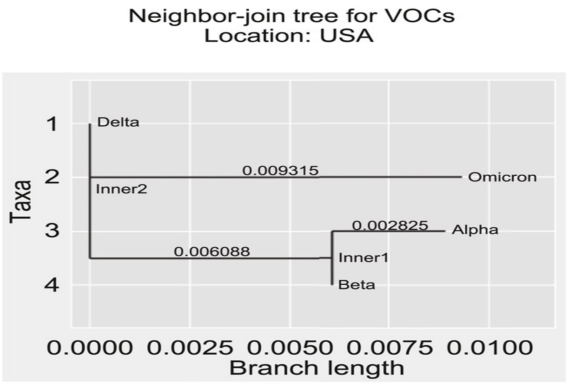

Figure 1: Neighbor-join Tree for VOC. Location is USA

Variant	DPGR vs. Alpha
Alpha	0.000000
Beta	0.002825
Delta	0.008913
Omicron	0.018228

Table 1: DPGR Values for SARS-CoV-2 Variants Relative to Alpha in the USA (Derived via Direct and Indirect Methods)

Variant	DPGR vs. Alpha
Alpha	0.000000
Beta	0.012587
Delta	0.044362
Omicron	0.117089

Table 2: DPGR Values for SARS-CoV-2 Variants Relative to Alpha in Europe (Derived via Direct and Indirect Methods)

Dataset Construction from GISAID Sequences

To enable predictive modeling of relative transmission fitness across SARS-CoV-2 variants, we constructed a labeled dataset using publicly available viral genome sequences from the GISAID database (Shu and McCauley 2017). This dataset combines raw sequence information, curated metadata, and fitness labels expressed as Differential Population Growth Rate (DPGR) values.

Data Acquisition and Cleaning SARS-CoV-2 genome sequences were downloaded from the GISAID database (Shu and McCauley 2017), along with associated metadata including variant lineage, geographic location, and collection date. Only entries with complete metadata and valid timestamp annotations were retained. Pango lineages were mapped to standardized WHO variant labels (e.g., Alpha, Beta, Delta, Omicron), and collection dates were grouped into epidemiological weeks to align with temporal splits and with the regional DPGR windows reported in Tables 1 and 2. We focused on two well-sampled regions:(1) USA and (2) Europe over the period January 2020 to May 2022, during which these variants exhibited substantial co-circulation (Robles-Escajeda et al. 2023; Pascall et al. 2023). A total of 6,950 sequences were collected from Europe and 5,690 from the USA. The variant distribution within each region was as follows:

- **USA:** Delta (2,122), Omicron (1,828), Alpha (1,304), Beta (436)
- **Europe:** Alpha (2,192), Omicron (1,815), Delta (1,686), Beta (1,257)

To ensure fair learning and unbiased prediction across different variant types and regions, the dataset used for regression training and testing was balanced with respect to both variant class and geographic origin. This careful curation enabled robust generalization in downstream modeling and minimized over-representation bias.

DPGR label assignment For this study, transmission-fitness labels are not recomputed from surveillance windows. Instead, fixed DPGR values reported in Tables 1 and 2 from prior work (Pantho et al. 2025) are used as external targets. Extension to additional regions is planned to broaden geographic coverage. Each sequence s with lineage/variant $v(s)$ and region $r(s) \in \{\text{USA}, \text{Europe}\}$ is annotated with

$$y(s) \;=\; \text{DPGR}_{\text{ref}\to v(s)}^{\,r(s)} \in \mathbb{R},$$

looked up from the corresponding regional table. The sign convention follows the tables: $y > 0$ indicates that $v(s)$ exhibits higher population-level transmission fitness than the reference in region $r(s)$ over the reported window(s). No recomputation of DPGR uses the sequences in this dataset.

These values are based on region-specific comparisons of weekly variant prevalence and are assigned according to a reference variant (e.g., Alpha). In that prior work (Pantho et al. 2025), DPGR values were obtained either by direct regression on co-circulating variant pairs or, when pairs were not concurrently observed, by indirect inference using the additive relationships between variant fitness values, as described earlier. This process yields a fitness score derived from empirical frequency dynamics, enabling a consistent and interpretable regression target across all variants and regions.

Label sensitivity and provenance. The DPGR labels used in this study are fixed external estimates sourced from prior work (Pantho et al. 2025), where sensitivity to window start/length and linearity criteria was systematically analyzed (e.g., retaining windows with $R^2 > 0.9$ and $p < 0.05$; see Supplementary Fig. S14 in (Pantho et al. 2025)). The values reported in Tables 1 and 2 inherit these criteria; only windows meeting the designated thresholds were included. No additional DPGR re-estimation or sensitivity sweep is performed here; the present model evaluates sequence→DPGR mapping conditional on those vetted labels.

Sequence Encoding and Feature Engineering Genomic sequences were preprocessed by replacing any ambiguous or non-standard nucleotide symbols with a placeholder token. All sequences were one-hot encoded using seven nucleotide channels: `a`, `t`, `c`, `g`, `n`, `-`, and `i`, where the last two represent gaps and ambiguities, respectively. Each sequence was padded or truncated to a fixed length of 29,891 positions, resulting in input tensors of shape $(29891, 7)$. This encoding ensured uniform input structure and preserved mutation patterns and sequence variability relevant to transmission fitness.

Final Dataset Schema The resulting dataset consisted of the following elements:

- **Input features:** One-hot encoded sequence matrices of size (29891×7).
- **Target label:** A scalar DPGR value corresponding to the variant's inferred transmission advantage.
- **Metadata:** Variant label, geographic region, and week of collection.

This structured dataset formed the basis for training machine learning models to predict DPGR from raw genomic sequences, thereby enabling indirect estimation of variant fitness from emerging or unsampled genomes. For model development, an i.i.d. (Independent and Identically Distributed), stratified random 10:20:70 split (train:validation:test = 70%:10%:20%) was applied at the sequence level, stratified by region and variant to preserve class balance.

Regression Model Development

To predict the Differential Population Growth Rate (DPGR) from genomic sequences, we trained a deep convolutional neural network (CNN) using one-hot encoded viral genome inputs and scalar DPGR values as regression targets. The model was implemented using TensorFlow and Keras, with hyperparameters optimized through automated tuning.

Model Architecture The model as shown in Figure 2 follows a sequential architecture beginning with three 1D convolutional blocks, each followed by max pooling layers for

dimensionality reduction. The convolutional layers progressively extract local nucleotide patterns at different resolutions.

- **Input shape:** $(29891, 7)$ — one-hot encoded genome sequences with 7 nucleotide channels.
- **Conv1D layer 1:** 76 filters, kernel size 3, stride 2.
- **Conv1D layer 2:** 54 filters, kernel size 5, stride 2.
- **Conv1D layer 3:** 18 filters, kernel size 6, stride 2.
- **MaxPooling:** Applied after each convolutional layer, with pool size 2 and stride size 3.
- **Flatten:** Converts final feature map to a flat vector.
- **Fully connected layers:** Four dense layers with [4536, 1500, 628, 66] neurons respectively.
- **Output:** A single neuron with linear activation to output the predicted DPGR value.

Layer normalization is applied after each dense layer to stabilize training. No dropout was used, and the final model has 19,103,225 trainable parameters.

Regularization and Optimization To mitigate overfitting, L1 regularization was applied with a weight of 0.0001. No L2 regularization was used. The model was optimized using Stochastic Gradient Descent (SGD) with the following configuration:

- **Optimizer:** SGD with momentum = 0.8 and Nesterov acceleration enabled.
- **Learning rate:** 0.0023 (tuned via hyperparameter search).
- **Loss function:** Mean Squared Error (MSE).

Hyperparameter Tuning and Evaluation Hyperparameters such as filter sizes, kernel sizes, dense layer widths, and learning rate were selected using automated hyperparameter tuning with KerasTuner. The best-performing configuration yielded a final validation loss (MSE) of approximately 1.94×10^{-4}, indicating strong agreement between predicted and actual DPGR values.

This CNN-based regression model serves as the backbone of our framework, learning to map complex nucleotide patterns to relative transmission fitness encoded as DPGR values.

Results

Model Performance

The regression model exhibited robust predictive performance in estimating the Differential Population Growth Rate (DPGR) from genomic sequence data. Initially, the training and validation loss metrics displayed a rapid decrease within the first 100 epochs, indicating efficient learning and rapid convergence toward an optimal solution, as illustrated in Figure 3. After this initial phase, both metrics stabilized and maintained consistent low values, signifying that the model reached and maintained a global minimum effectively. More specifically, the training and validation Mean Squared Error (MSE) values achieved convergence at approximately 1.94×10^{-4}, reflecting a remarkably low prediction error across the dataset. This alignment between the training and validation error rates further highlights the strong generalization capabilities of the model, suggesting minimal susceptibility to overfitting despite the complexity of genomic feature patterns. To quantitatively evaluate predictive accuracy, a scatter plot comparing predicted versus actual DPGR values (Figure 4) was analyzed. This visual assessment revealed a very tight clustering of predictions around the ideal prediction line, demonstrating the model's accuracy and consistency. The statistical evaluation of this relationship yielded a high coefficient of determination (R^2) value of 0.9168, indicating that the model effectively accounted for approximately 91.68% of the variance present in the actual DPGR measurements. This result underscores the model's exceptional ability to capture intricate genomic features associated with viral transmission fitness.

Overall, these detailed performance metrics and evaluations confirm the robustness and reliability of the convolutional neural network (CNN) architecture employed, validating its capacity to accurately discern complex genomic determinants of variant transmissibility from raw nucleotide sequence data.

SHAP-Based Interpretability of Genomic Features

To rigorously quantify the influence of specific genomic mutations on the predicted Differential Population Growth Rate (DPGR), SHapley Additive exPlanations (SHAP) analysis was employed. SHAP values provide a method rooted in cooperative game theory to interpret individual contributions of each feature (here, specific nucleotide mutations) to the final prediction of the regression model (Lundberg and Lee 2017). The waterfall plots depicted in Figure 5 illustrate how these genomic mutations cumulatively influence the predicted DPGR for each SARS-CoV-2 variant. In these SHAP plots, each mutation is represented by a bar that contributes positively (red) or negatively (blue) to the final model prediction. The length and direction of each bar indicate the magnitude and type of contribution, respectively. The base value E[f(x)] represents the average predicted DPGR across the entire dataset, while f(x) represents the specific prediction for the genomic sequence in question. The bars thus depict how individual mutations either increase or decrease the predicted DPGR relative to this baseline.

Detailed interpretation per variant is as follows:

- **Alpha Variant**: The mutations in `ORF9 (Nucleocapsid)` and `ORF5 (Membrane)` proteins exhibited predominantly negative contributions (blue), slightly lowering predicted fitness relative to the baseline. These mutations reflect moderate, yet biologically meaningful, impacts on the variant's transmissibility.
- **Beta Variant**: Highlighted contributions from `ORF2 (Spike)` and `ORF9 (Nucleocapsid)` mutations were largely negative (blue), indicating their limited enhancement of fitness in comparison to the baseline prediction. This aligns with known epidemiological observations of the Beta variant's moderate transmissibility.

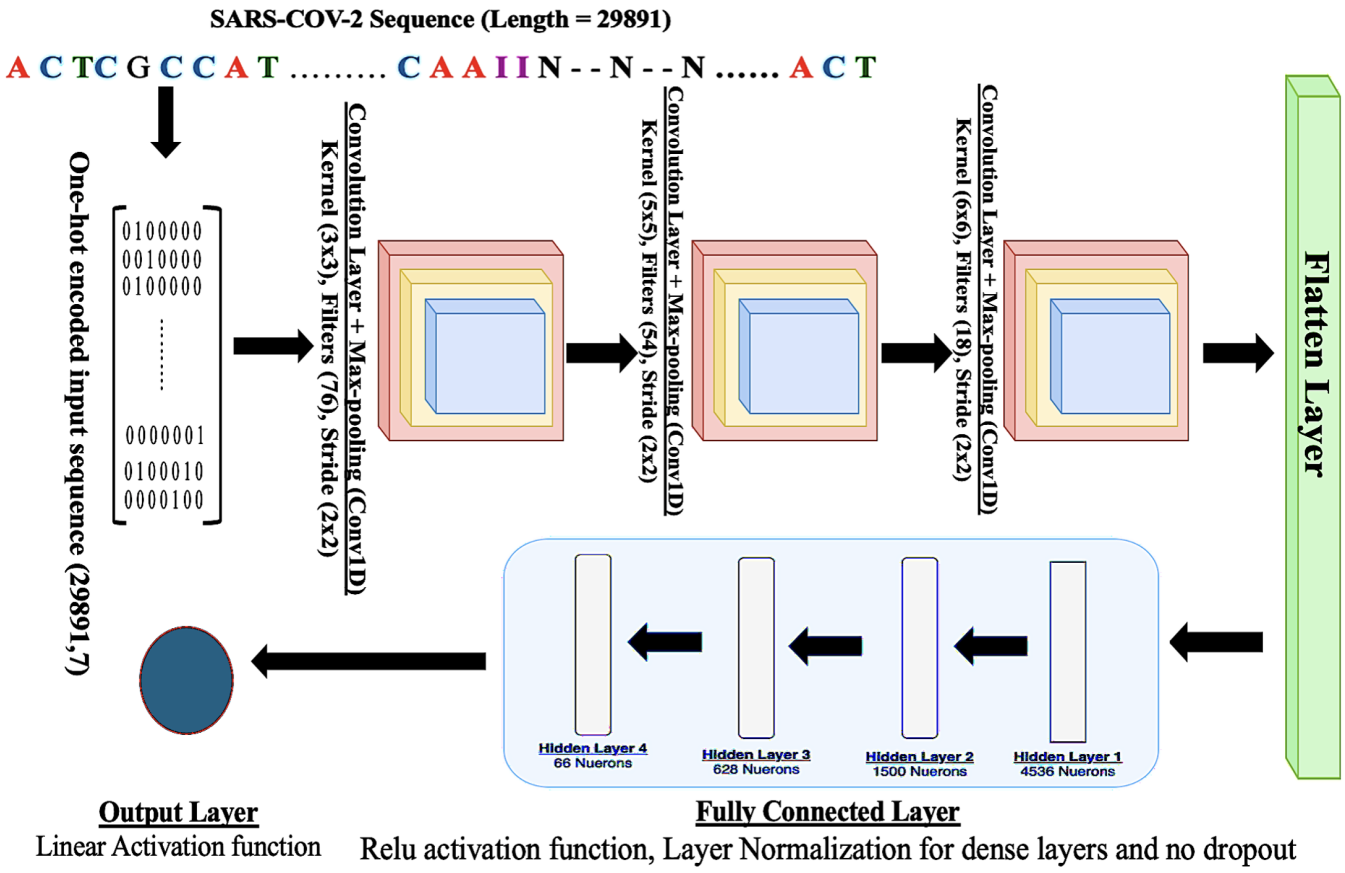

Figure 2: Model Architecture

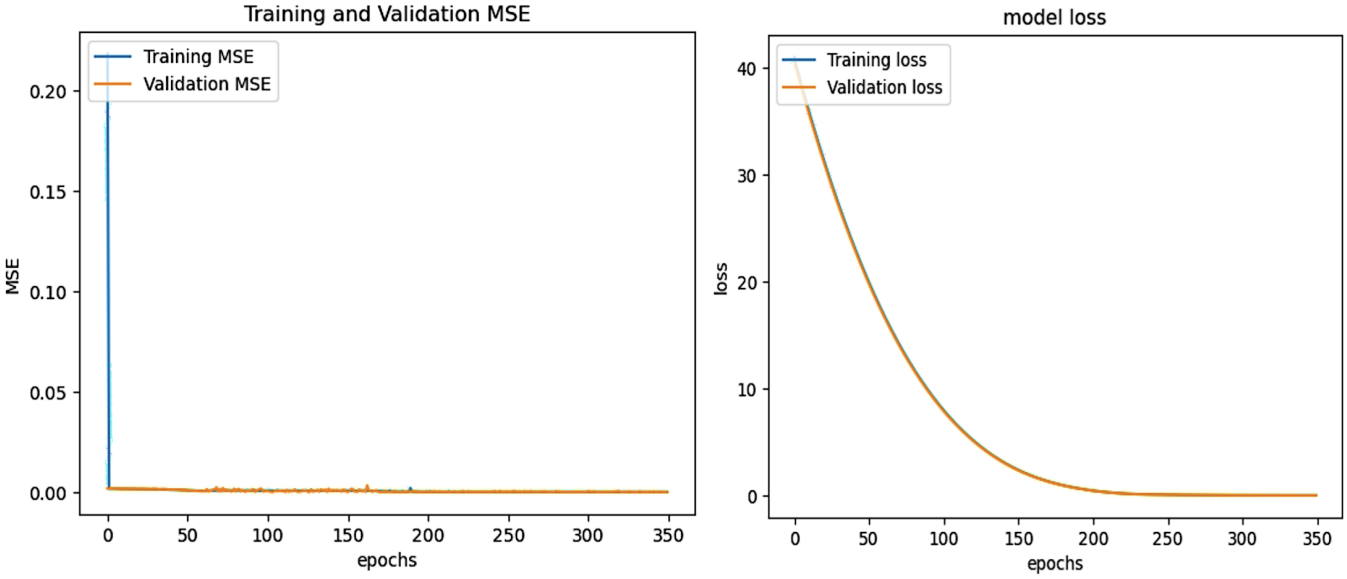

Figure 3: CNN Model Training and Validation Loss and MSE

- **Delta Variant**: Mutations within `ORF8`, `ORF9 (Nucleocapsid)`, and `ORF5 (Membrane)` proteins had mixed contributions, with positive (red) and negative (blue) SHAP values indicating complex interactions among mutations. Notably, key positive mutations underscore the Delta variant's elevated fitness observed epidemiologically.
- **Omicron Variant**: Mutations in `ORF5 (Membrane)`,

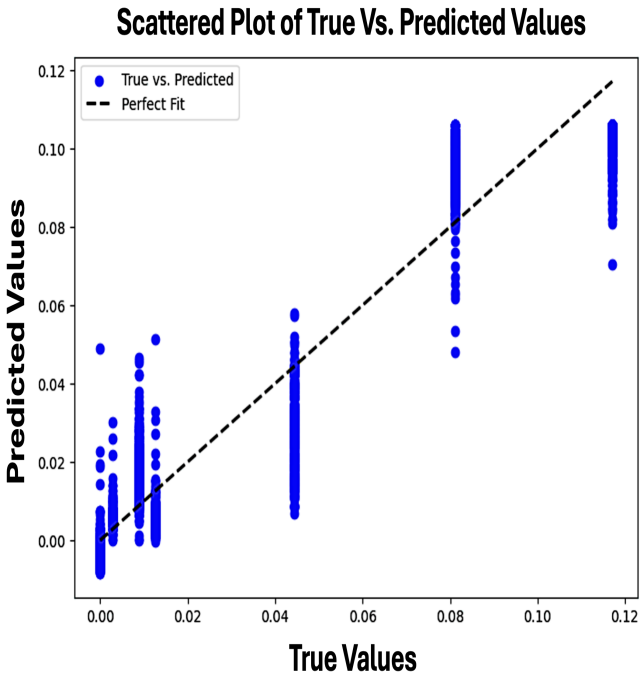

Figure 4: Test data Coefficient of Determination ($R^2 = 0.9168$) value

`ORF2 (Spike)`, and `ORF8` proteins showed strongly positive contributions (red bars), significantly elevating the predicted fitness. This robust positive contribution aligns closely with Omicron's observed higher transmissibility and immune evasion properties.

Overall, the SHAP analysis provides a scientifically robust method for interpreting the regression model's predictions at the mutation level, clarifying how individual genomic features collectively determine the DPGR. These insights are invaluable for understanding variant-specific evolutionary trajectories and guiding public health strategies.

Discussion

This study introduces a rigorous CNN-based approach. to predicting the transmission fitness of SARS-CoV-2 variants directly from genomic sequences, using the Differential Population Growth Rate (DPGR) as a proxy. The model demonstrated high predictive accuracy, with a mean squared error (MSE) of approximately 1.94×10^{-4} and a strong coefficient of determination ($R^2 = 0.9168$), indicating that it effectively captures sequence-to-fitness relationships. However, the true strength of this framework lies in its interpretability, made possible through SHapley Additive exPlanations (SHAP) which enables a mechanistic understanding of how specific nucleotide mutations contribute to transmission fitness.

The SHAP analysis provided mutation-level insights for each major variant. For the Alpha variant, the most influential features were mutations such as G28280C and T28282A, both located in ORF9 (nucleocapsid), corresponding to the amino acid change D3L. These mutations contributed negatively to the predicted fitness, consistent with the Alpha variant's moderate transmission advantage (Liu et al. 2022). Other contributing mutations like A28363A and A28365A also mapped to ORF9, though their exact protein-level implications are less well characterized. C26577G (ORF5) was observed but does not result in an amino acid change; it corresponds to S83S, a synonymous substitution in the membrane protein that is unlikely to alter protein function but may still impact regulatory or structural genome contexts (Bailey, Morales, and Kassen 2021) .

For the Delta variant, the top features included deletions such as A28249- and T28252- in ORF8, and a non-coding region deletion A68-. These mutations had mixed SHAP values, with both positive and negative contributions. The ORF8 deletions may suggest structural disruption or altered immune modulation, which have been documented in previous studies. Mutations A28365A (ORF9, N protein) and C26577G (ORF5) were also observed, again suggesting that non-spike regions significantly influence model predictions.

In the case of the Beta variant, prominent mutations included G23012A and C22995C in ORF2 (spike protein), likely corresponding to amino acid mutations E484K and N501Y, both associated with immune escape (Harvey et al. 2021; Tian et al. 2021). Additionally, A28365A and A28363A in ORF9 were again detected. C26577G (ORF5, synonymous S83S) was also noted, consistent with its recurrence in other variants.

Omicron, the most transmissible of the variants examined, had the highest overall prediction score ($f(x) = 0.09413$). Key features driving this score included C26577G and G26709A in ORF5 (membrane protein), corresponding to S83S and V97I, respectively (Lee et al. 2022). T21992- in ORF2 (spike protein) likely indicates a structural change near the N-terminal domain. Additionally, A28365– (ORF9) and G28242N (ORF8) were strong positive contributors.

Taken together, these SHAP-derived interpretations reveal consistent themes: mutations in ORF9 (nucleocapsid), ORF2 (spike), ORF5 (membrane), and ORF8 are repeatedly implicated in fitness predictions across all variants. This suggests that while spike mutations are central to entry and immune evasion, changes in structural and accessory genes also play important roles. Importantly, the model successfully surfaced these relationships without being explicitly trained to prioritize them, underscoring its potential utility in genomic surveillance. Extending interpretability to additional samples per variant and comparing mutation importance across temporal windows could reveal evolutionary pressures or shifts in fitness determinants. Furthermore, integrating functional annotations or protein-structure data may enable deeper biological validation of impactful mutations identified through SHAP.

Conclusion

This work demonstrates that convolutional neural networks, trained on one-hot encoded SARS-CoV-2 genomic sequences, can effectively predict the transmission fitness of viral variants using Differential Population Growth Rate (DPGR) as a target metric. The high correlation between predicted and actual DPGR values confirmed the model's

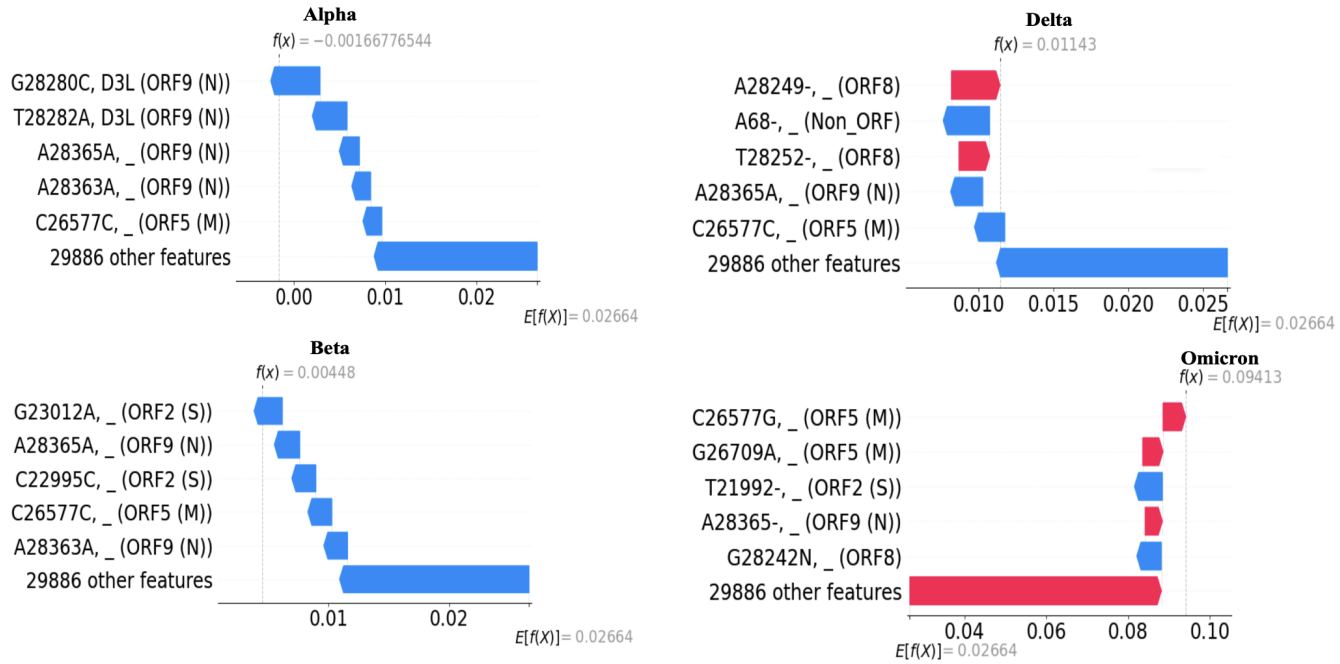

Figure 5: Figure: SHAP waterfall plots illustrating the mutation-level contributions to the predicted Differential Population Growth Rate (DPGR) for different SARS-CoV-2 variants. Top-left plot represents the Alpha variant, highlighting positive fitness impacts primarily from mutations in ORF9 (Nucleocapsid protein) and ORF5 (Membrane protein). Top-right plot depicts the Delta variant, emphasizing substantial fitness contributions from mutations in ORF8, ORF5 (M), and ORF9 (N). Bottom-left plot corresponds to the Beta variant, showing influential mutations predominantly within ORF2 (Spike protein), ORF9 (N), and ORF5 (M). Bottom-right plot illustrates the Omicron variant, demonstrating pronounced fitness enhancements driven by significant mutations in ORF5 (M), ORF2 (S), and ORF8, consistent with its high transmissibility and immune evasion characteristics.

capacity to learn biologically relevant sequence patterns. Importantly, through SHapley Additive exPlanations (SHAP), we uncover mutation-level insights into how individual nucleotide changes influence model predictions. Several mutations that produce amino acid changes; such as D3L in ORF9, E484K and N501Y in ORF2, and V97I in ORF5 were found to strongly elevate predicted fitness, consistent with known roles in immune escape and transmissibility. Meanwhile, synonymous substitutions such as C26577G (S83S) demonstrated minimal effect on model output but suggest possible regulatory importance. The alignment between SHAP-derived attributions and biologically validated fitness-enhancing mutations supports the utility of our model not only for prediction, but also for mechanistic insight. As SARS-CoV-2 continues to evolve, such interpretable deep learning frameworks can offer valuable tools for variant prioritization and genomic surveillance. Future extensions will involve integrating protein-structure data and larger variant pools to further enrich interpretability and generalizability. This could support the real-time evaluation of novel variants as they emerge in global surveillance systems.

Limitations and future work. While the 1D CNN captures local sequence motifs, it may under-represent long-range and combinatorial (epistatic) dependencies. Future work includes: (i) introducing dilated convolutions and self-attention layers to expand the receptive field and capture cross-genome dependencies; (ii) incorporating interaction-aware attribution (e.g., SHAP interaction values and pairwise ablations) to characterize mutational synergy/antagonism; and (iii) exploring hybrid encoders that fuse sequence context with population-level covariates (e.g., region, collection week, prior exposure proxies) to improve robustness and interpretability.

Population context and sampling considerations The labels used here quantify *population-level* transmission fitness within specific regions and periods and therefore implicitly reflect contemporaneous community immunity, interventions, and testing practices. Although training focused on USA and Europe, extension to additional regions (e.g., Asia, Africa, and South America) is planned. Notably, DPGR's ratio-based design can help cancel *nondiscriminatory* sampling errors in genomic surveillance (Pantho et al. 2025), yet region-specific biases may persist, motivating broader geographic coverage and covariate adjustment in future evaluations.

Acknowledgments

We thank the support of USA NSF 2525493 and 2200138 and internal support of the Old Dominion University. A patent is pending based on portion of this work.

References

Ahmad, A.; Fawaz, M. A. M.; and Aisha, A. 2022. A comparative overview of SARS-CoV-2 and its variants of concern. *Le Infezioni in Medicina*, 30(3): 328–343.

Bailey, S. F.; Morales, L. A. A.; and Kassen, R. 2021. Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments. *Genome Biology and Evolution*, 13(9): evab141.

Carabelli, A. M.; Peacock, T. P.; Thorne, L. G.; Harvey, W. T.; Hughes, J.; de Silva, T. I.; Peacock, S. J.; Barclay, W. S.; Towers, G. J.; and Robertson, D. L. 2023. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology*, 21: 162–177. Publisher: Nature Publishing Group.

Donker, T.; Papathanassopoulos, A.; Ghosh, H.; Kociurzynski, R.; Felder, M.; Grundmann, H.; and Reuter, S. 2024. Estimation of SARS-CoV-2 fitness gains from genomic surveillance data without prior lineage classification. *Proceedings of the National Academy of Sciences*, 121(25): e2314262121.

Elkin, M. E.; and Zhu, X. 2025. Paying attention to the SARS-CoV-2 dialect : a deep neural network approach to predicting novel protein mutations. *Communications Biology*, 8(1): 98.

Harvey, W. T.; Carabelli, A. M.; Jackson, B.; et al. 2021. SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19(7): 409–424.

Hatami, P.; Annan, R.; Miranda, L. U.; Gorman, J.; Xie, M.; Qingge, L.; and Qin, H. 2024. Explainable convolutional neural network model provides an alternative genome-wide association perspective on mutations in SARS-CoV-2. arXiv:2410.22452.

Hu, T.; Darabos, C.; and Urbanowicz, R. 2020. Editorial: Machine Learning in Genome-Wide Association Studies. *Frontiers in Genetics*, Volume 11 - 2020.

Ito, J.; Strange, A.; Liu, W.; Joas, G.; Lytras, S.; and Sato, K. 2025. A protein language model for exploring viral fitness landscapes. *Nat Commun*, 16: 4236. Publisher: Nature Publishing Group.

Jha, P.; Brown, P. E.; and Ansumana, R. 2022. Counting the global COVID-19 dead. *The Lancet*, 399(10339): 1937–1938.

Kung, Y.-A.; Chuang, C.-H.; Chen, Y.-C.; Yang, H.-P.; Li, H.-C.; Chen, C.-L.; Janapatla, R. P.; Chen, C.-J.; Shih, S.-R.; and Chiu, C.-H. 2025. Worldwide SARS-CoV-2 Omicron variant infection: Emerging sub-variants and future vaccination perspectives. *Journal of the Formosan Medical Association*, 124(7): 592–599.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature*, 521(7553): 436–444.

Lee, J.-M.; Jung, J.; Lee, K. M.; et al. 2022. Membrane protein mutations (e.g., V97I) modulate SARS-CoV-2 pathogenicity. *Frontiers in Medicine*, 9: 815389.

Liu, Y.; Liu, J.; Plante, K. S.; et al. 2022. The Nucleocapsid Protein of SARS-CoV-2: a Target for Vaccine and Therapeutic Development. *Journal of Virology*, 96(4): e01991–21.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lytras, S.; Lamb, K. D.; Ito, J.; Grove, J.; Yuan, K.; Sato, K.; Hughes, J.; and Robertson, D. L. 2025. Pathogen genomic surveillance and the AI revolution. *Journal of Virology*, 99(2): e01601–24.

Markov, P. V.; Ghafari, M.; Beer, M.; Lythgoe, K.; Simmonds, P.; Stilianakis, N. I.; and Katzourakis, A. 2023. The evolution of SARS-CoV-2. *Nature Reviews Microbiology*, 21(6): 361–379.

Min, S.; Lee, B.; and Yoon, S. 2017. Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5): 851–869.

Pantho, M. J.; Annan, R.; Bauder, L. A.; Huang, S.; Qingge, L.; and Qin, H. 2025. A data-driven sliding-window pairwise comparative approach for the estimation of transmission fitness of SARS-CoV-2 variants and construction of the evolution fitness landscape. *Quantitative Biology*, 13(4): e70003.

Pascall, D. J.; Vink, E.; Blacow, R.; Bulteel, N.; Campbell, A.; Campbell, R.; Clifford, S.; Davis, C.; da Silva Filipe, A.; El Sakka, N.; Fjodorova, L.; Forrest, R.; Goldstein, E.; Gunson, R.; Haughney, J.; Holden, M. T.; Honour, P.; Hughes, J.; James, E.; Lewis, T.; MacLean, O.; McHugh, M.; Mollett, G.; Nyberg, T.; Onishi, Y.; Parcell, B.; Ray, S.; Robertson, D. L.; Seaman, S. R.; Shabaan, S.; Shepherd, J. G.; Smollett, K.; Templeton, K.; Wastnedge, E.; Wilkie, C.; Williams, T.; and Thomson, E. C. 2023. Directions of change in intrinsic case severity across successive SARS-CoV-2 variant waves have been inconsistent. *Journal of Infection*, 87(2): 128–135.

Qin, Z.; Zhang, Z.; Li, Y.; and Guo, J. 2019. Making Deep Neural Networks Robust to Label Noise: Cross-Training With a Novel Loss Function. *IEEE Access*, 7: 130893–130902.

Robles-Escajeda, E.; Mohl, J. E.; Contreras, L.; Betancourt, A. P.; Mancera, B. M.; Kirken, R. A.; and Rodriguez, G. 2023. Rapid Shift from SARS-CoV-2 Delta to Omicron Sub-Variants within a Dynamic Southern U.S. Borderplex. *Viruses*, 15(3): 658.

Saitou, N.; and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4): 406–425.

Sehrawat, S.; Najafian, K.; and Jin, L. 2023. Predicting phenotypes from novel genomic markers using deep learning. *Bioinformatics Advances*, 3(1): vbad028.

Shu, Y.; and McCauley, J. 2017. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13): 30494.

Singh, D.; and Yi, S. V. 2021. On the origin and evolution of SARS-CoV-2. *Experimental & Molecular Medicine*, 53: 537–547. Publisher: Nature Publishing Group.

Tian, F.; Tong, B.; Sun, L.; et al. 2021. N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2. *eLife*, 10: e69091.

Washburn, J. D.; Cimen, E.; Ramstein, G.; Reeves, T.; O'Briant, P.; McLean, G.; Cooper, M.; Hammer, G.; and Buckler, E. S. 2021. Predicting phenotypes from genetic, environment, management, and historical data using CNNs. *Theoretical and Applied Genetics*, 134(12): 3997–4011.