AI-based model for SARS-CoV-2 ACE2 binding prediction and lineage fitness forecasts

Zhaojun Ding [1], Hongsen Liang [1], Hsiang-Yu Yuan [1,2 *]

[1] Department of Biomedical Sciences, College of Biomedicine, City University of Hong Kong, Hong Kong

[2] Centre for Applied One Health Research and Policy Advice, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong

*Correspondence author: sean.yuan@cityu.edu.hk

Summary

We train sequence-to-trait models that map SARS-CoV-2 RBD sequences to ACE2 binding using ESM-1b embeddings and four deep regressors: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), and Recurrent Neural Network (RNN). The MLP achieves the best test accuracy (lowest MSE; Pearson r ≈ 0.84). When AI-predicted ACE2 binding and antibody-retained scores replace additive trait sums in the Multinomial Logistic Regression (MLR) fitness framework (see supplementary Figure S1), predictive accuracy and robustness improve at the country level. This forms an end-to-end pipeline: genome → traits → lineage growth.

Background

Deep Mutational Scanning (DMS) offers precise trait measurements but lags behind the pace of viral evolution, necessitating a learned mapping from sequence to phenotype for real-time scoring. Prior models used regression model with additive trait scores, such as ACE2 binding, immune escape, and cell entry, combined with mobility and social distancing indices. However, additive frameworks struggle to capture epistatic effects. Here, we leverage protein language model embeddings and deep neural networks (DNNs) to model non-additive interactions and integrate AI-derived trait scores into the MLR fitness framework.

Motivation

The primary motivation is timeliness and accuracy. Models must generalize to unseen mutations without waiting for laboratory assays, and they must capture epistasis that additive trait sums cannot represent. From a public-health perspective, more accurate trait estimates translate into more reliable early warnings about high-risk lineages, enabling faster and better-targeted responses.

Data & Preprocessing

We curate SARS-CoV-2 RBD sequences spanning residues N331 to T531 in the SARS-CoV-2 spike protein [1]. The target labels are experimentally measured ACE2 binding values for more than four thousand variants. Data are partitioned into training, validation, and test sets with an additional independent hold-out for external verification. For representation, we generate per-token ESM-1b embeddings and pool them into a fixed-length vector for each sequence [2]. All features are standardized using statistics computed on the training split only and then applied to the validation and test sets. Performance is evaluated primarily by mean squared error and Pearson correlation, supplemented by calibration analyses comparing predictions against observations.

Methods

Embeddings of SARS-CoV-2 RBD region

We generate ESM-1b embeddings for each sequence and train four regressors with MSE loss and early stopping. Hyperparameters are tuned on validation performance. The best trait predictor is plugged into the MLR lineage-fitness model by substituting additive inputs with AI-predicted ACE2 and antibody-retained scores. ESM-1b, a transformer-based protein language model, provides contextual embeddings for each RBD sequence (Figure 1). We apply a pooling rule to obtain a fixed-length vector that serves as input to the regressors.

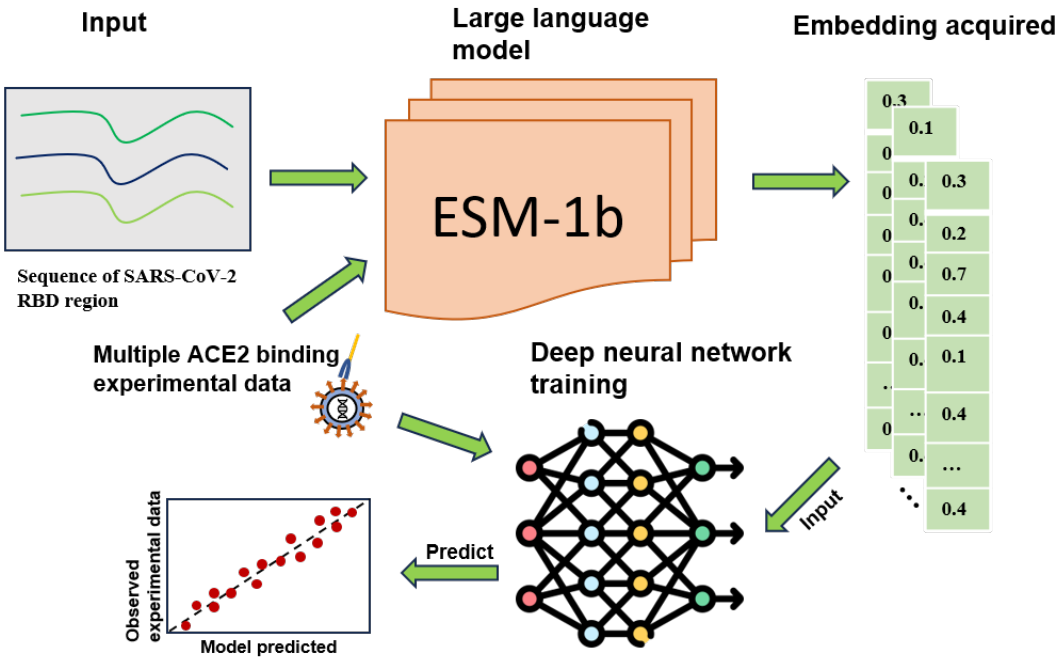

Figure 1. Workflow for predicting ACE2 binding affinity using ESM-1b and deep neural networks. The input consists of the SARS-CoV-2 RBD sequence. This sequence is processed by the ESM-1b, which generates an embedding representing the sequence. The model is trained using multiple ACE2 binding experimental data to optimize the prediction. The embedding acquired from ESM-1b is fed into a deep neural network for training, which then predicts the binding affinity between the SARS-CoV-2 RBD and the human ACE2 receptor.

The input of the ESM-1b model was a protein sequence, and the output was the vector representation (embedding) of this sequence (Fig. 4.1). This embedding captured the evolutionary information of protein sequences and was crucial for helping deep neural network models predict the effects of unseen mutation combinations. Specifically, ESM-1b produces final-layer hidden representations via forward propagation [3]. In this study, the length of the SARS-CoV-2 RBD sequence was 201. In order to generate the comprehensive representation of the entire sequence, we applied mean pooling to these vectors to obtain a single sequence representation [2]. This representation was used as input to train various DNN models.

Evaluation & Model Selection

After obtaining the embeddings generated by ESM-1b, this study trained four different deep neural network models: CNN, LSTM, MLP, and RNN to predict viral ACE2 binding [4]. These models took the embeddings as input and learned the relationship

between the binding affinity of RBD and the ACE2 receptor through training. Each model continuously optimized its parameters by minimizing the error between the predicted values and the experimental data. For model evaluation, we randomly split the data into 80% for training and 20% for testing. Ultimately, the trained models were capable of predicting the ACE2 binding affinity, accounting for the epistasis of RBD mutations and providing an efficient tool for the downstream analysis.

We compare CNN, LSTM, MLP, and simple RNN architectures under identical training protocols. Regularization through dropout and early stopping is used to avoid overfitting. Model selection is based on the lowest test mean squared error and the strongest Pearson correlation with ground-truth measurements.

Viral fitness prediction

The prediction of SARS-CoV-2 lineage fitness in this chapter follows the same analytic framework as established in MLR fitness framework (see Figure S1). Specifically, viral fitness was defined as the lineage growth rate, with growth rate estimates directly adopted from the results of in MLR fitness framework and not recalculated in this study.

In this study, two viral trait predictors, ACE2 binding and antibody binding retained, were updated by replacing the additive scores from in MLR fitness framework. The cell entry score and Social Demographic Index (SDI) for each country remained unchanged and were sourced from MLR fitness framework (see Figure S1 and S2). These four predictors (updated ACE2 binding, updated antibody binding retained, cell entry score, and SDI) were used as input features for the same machine learning models (XGBoost, stepwise-GAM, Random Forest, and Ridge regression) mentioned in in MLR fitness framework (see Figure S1) to predict viral lineage fitness, allowing direct comparison with the previous additive modeling approach.

Results

Trait Prediction

The performance of four deep neural network models (i.e., CNN, LSTM, MLP, and RNN) in the regression prediction task was compared. The results showed that the MLP model performs best among all the evaluation indicators, demonstrating excellent training stability and prediction accuracy (Fig. 4.2). First of all, in the training of the MLP model

in the process, the loss value continues to decline, and the loss curve shows the smallest fluctuation; this indicates that the training process was very stable, with quick convergence in the late training (Figure 2A). In contrast, the loss curves of the LSTM and RNN models show significant fluctuations and fail to achieve smooth convergence throughout the training process. The loss curve of the CNN model lies between MLP, LSTM, and RNN.

In terms of mean square error (MSE), the MLP model is significantly superior to other models. The minimum MSE values showed that the prediction accuracy was the highest on the test set. CNN follows closely behind. Its MSE is slightly higher than that of MLP, but it is still significantly superior to the LSTM and RNN models (Figure 2B). The MSE values of the LSTM and RNN models are significantly higher, especially LSTM, whose performance is much lower than that of CNN and MLP. The MLP model prediction accuracy also got further verification through a scatter plot (see Figure 2C). In the comparison between the predicted values and the true values, the prediction results of the MLP model show a strong linear relationship with the real data, and the Pearson correlation coefficient is 0.84, indicating that it has high accuracy and reliability in the prediction task (Figure 2C).

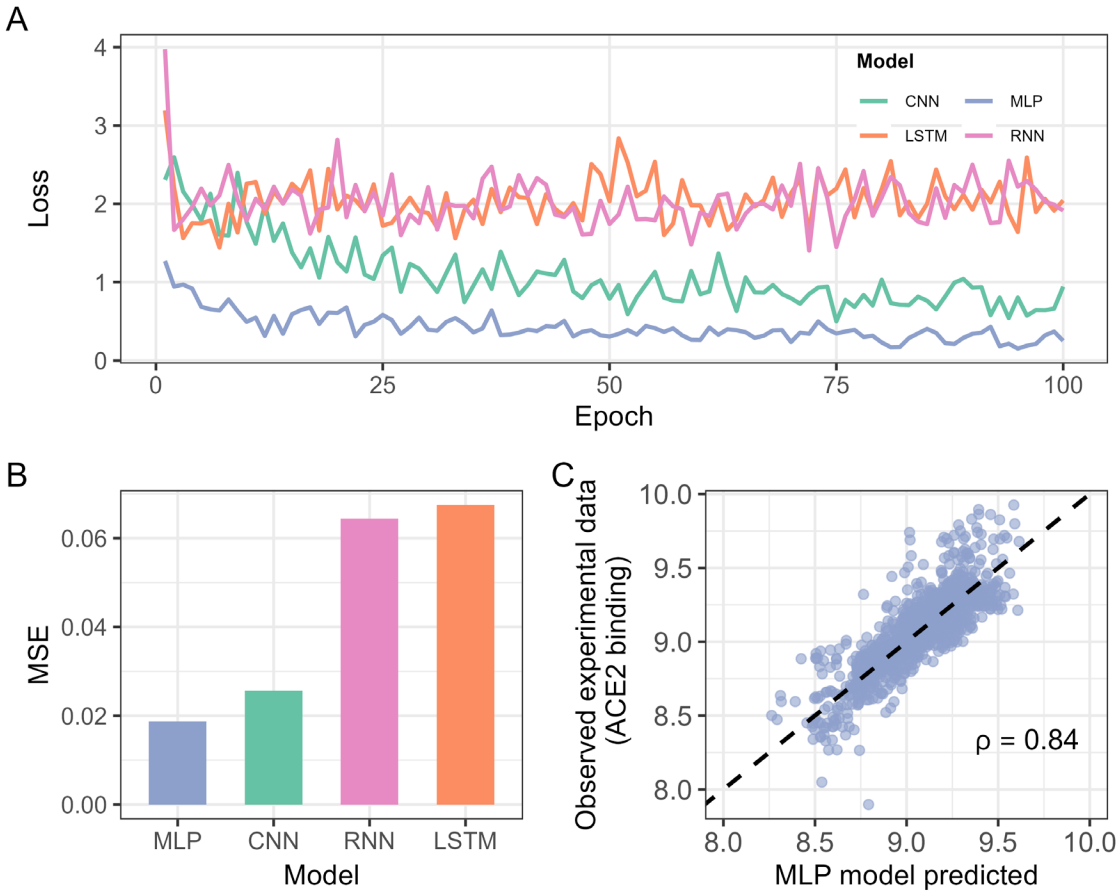

Figure 2. Performance comparison of four deep neural network models in regression prediction tasks. (A) Training loss curves of four models: Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Multilayer Perceptron (MLP), and Recurrent Neural Network (RNN) across training epochs. Among them, MLP exhibits the lowest and most stable loss throughout training, while LSTM and RNN show higher and more fluctuating loss trends. (B) Bar plot comparing the mean squared error (MSE) of the four models on the test set. MLP achieves the lowest MSE, indicating superior predictive accuracy. (C) Scatter plot of predicted values from the MLP model versus observed experimental data. The dashed diagonal line represents the ideal prediction, and the close clustering around it suggests high predictive accuracy. ρ is the Pearson correlation coefficient.

Downstream Fitness Prediction Gains

Compared with the baseline MLR fitness framework, replacing the trait input with AI-predicted ACE2 makes the predicted–observed clouds adhere more tightly to the 45° line, especially in the high-fitness range (e.g.,100–170) (see Figure 3 and supplementary Figure S2). For nonlinear tree models, this translates into equal or better accuracy and visibly stronger robustness across countries: XGBoost improves from ρ = 0.887 to ρ = 0.895, while Random Forest remains effectively unchanged (0.879 → 0.878) but shows tighter clustering around the diagonal at both tails, fewer outliers, and a more uniform

band across national subsets. Error "fan" patterns shrink (reduced heteroscedasticity), and the dashed trend line slope moves closer to 1, indicating reduced systematic bias and better tail calibration.

For Stepwise GAM and Ridge, the aggregate ρ decreases (0.835 → 0.683; 0.728 → 0.535) when using AI-predicted ACE2. This contrast is informative: AI-predicted ACE2 carries nonlinear/epistatic structure that nonlinear learners (XGBoost/RF) can exploit, whereas lower-capacity linear/semi-parametric models cannot fully absorb the added signal. Even so, in the mid-to-upper range these models show less systematic underestimation and improved responsiveness to extremes.

Overall, plugging AI-predicted ACE2 into the MLR framework yields (i) higher accuracy for the best-performing nonlinear model (XGBoost) and at least no loss for Random Forest, and (ii) clear robustness gains across countries, seen as diminished country-wise stratification, fewer high-end outliers, and a trend line closer to identity. These patterns substantiate the claim that AI-predicted ACE2 improves both accuracy and robustness across all countries, with the largest benefits realized when paired with models capable of capturing nonlinear interactions.

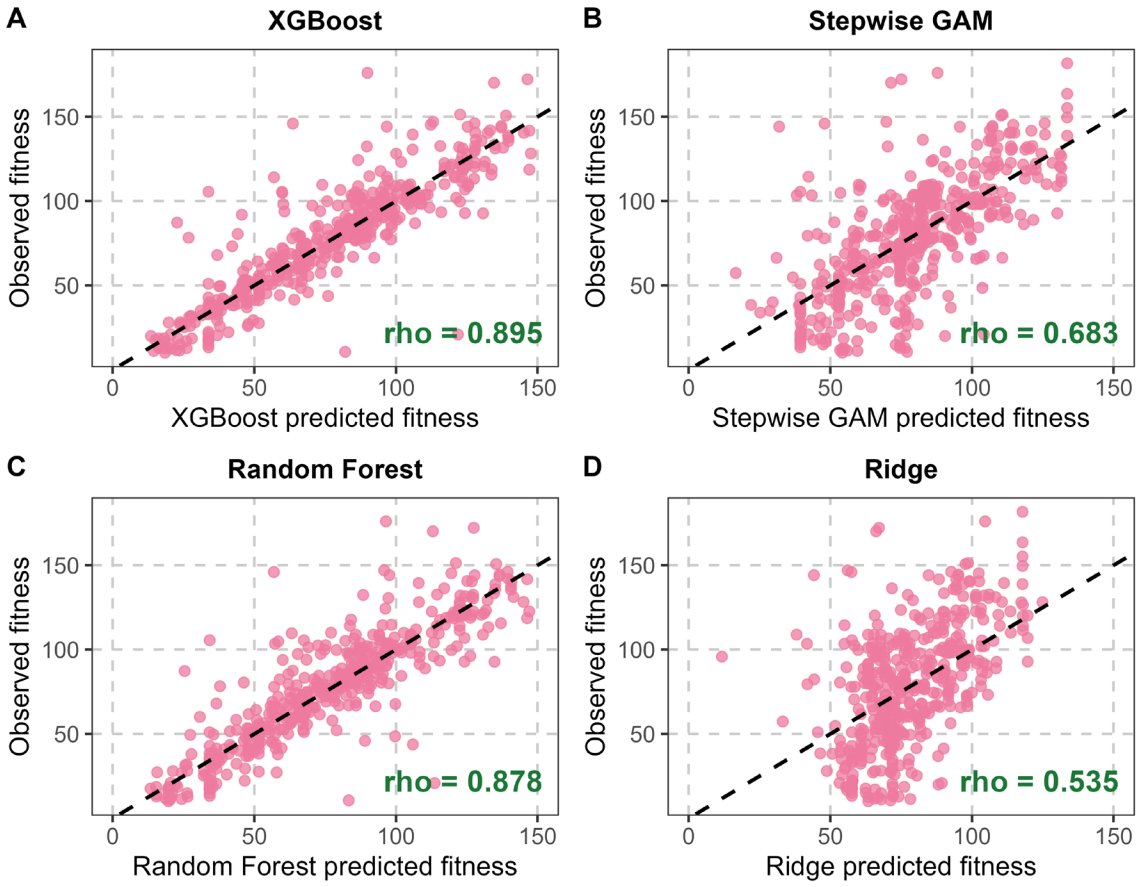

Figure 3. Performance of different machine learning models in predicting SARS-CoV-2 lineage fitness. Each panel shows the scatter plot of predicted versus observed lineage fitness (growth rate),

with the black dashed line indicating the ideal prediction. The Pearson correlation coefficient (rho/Pearson r) is shown in green for each model. AI-based predicted ACE2 binding and antibody binding retained scores, along with cell entry score and SDI, were used as input features.

The decreased trend of MLP-predicted ACE2 binding from Alpha (B.1.1.7) to Omicron (BA.3) was observed (Figure 4), which was consistent with the trend observed in both Calvaresi et al. (2023) and Han et al. (2022) data [5, 6]. The predicted values closely follow the trend observed in the experimental data, validating the model's predictive ability for ACE2 binding across different VOCs.

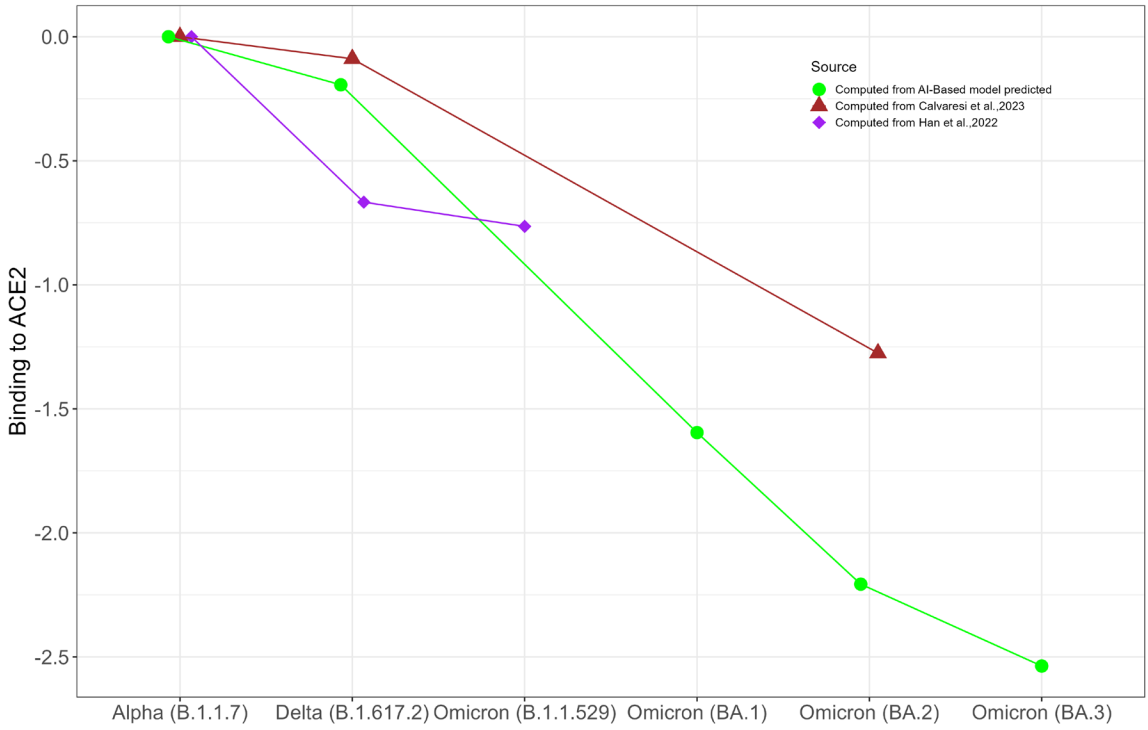

Figure 4. MLP model prediction of ACE2 binding for each VOC. The trend of binding to ACE2 among each VOC for the Alpha (B.1.1.7), Delta (B.1.617.2), and Omicron variants (B.1.1.529, BA.1, BA.2, BA.3), with values adjusted by subtracting Alpha's binding. Data are derived from three sources: Green circles represent values obtained from AI-based model predictions. Red triangles indicate data sourced from Calvaresi et al. (2023) [5]. Purple diamonds correspond to data from Han et al. (2022) [6].

Limitations & Risk Mitigations

The label set may be biased toward specific sequence backgrounds, which can be mitigated through re-weighting or diversification of training data. Out-of-distribution mutations remain a challenge; uncertainty estimates via Monte-Carlo dropout or ensembles can flag low-confidence predictions. Finally, the present work focuses on ACE2 binding and antibody-binding retention; extending to multi-trait joint modeling that also includes cell-entry efficiency constitutes an important direction for future research.

Conclusion

Protein-LM embeddings + DNNs yield accurate, scalable trait predictions. When integrated into the MLR fitness frame work, AI-predicted traits outperform additive trait sums, enabling faster and more reliable surveillance of emerging lineages.

Supplementary material

In this study, a statistical framework to elucidate the differences in SARS-CoV-2 growth rates across countries with varying development status was conducted firstly. This was achieved by fitting the MLR model, which effectively captures the transmission advantages of multiple virus lineages and quantifies their growth rates. For the first time, the association between SARS-CoV-2 growth rates and three key viral traits (i.e., ACE2 binding, immune escape, and cell entry) was systematically quantified among different socio-economic contexts. Previous comparative analyses of COVID-19 were largely limited to high- or middle-income countries, or examined global patterns without disaggregating by economic context. This work extends this literature by incorporating data from low-income countries, which have been under-represented in epidemiological modeling due to data scarcity. The findings suggest that socio-economic context is significantly associated with viral lineage growth rate, with notable differences in how viral traits influence growth rates across countries with varying levels of economic development. The relationships between SARS-CoV-2 viral traits, socio-demographic factors, and lineage growth rates were further validated using multiple machine-learning models. The results demonstrate that a combination of viral ACE2 binding, immune escape, cell entry scores, and SDI enables accurate prediction of viral lineage growth rate. Notably, XGBoost

delivered the best performance, as reflected by both regression and binary classification metrics. These findings support the view that both viral phenotypic traits and socio-economic contexts contribute to viral fitness. More details can be found in the article published in the preprint version.

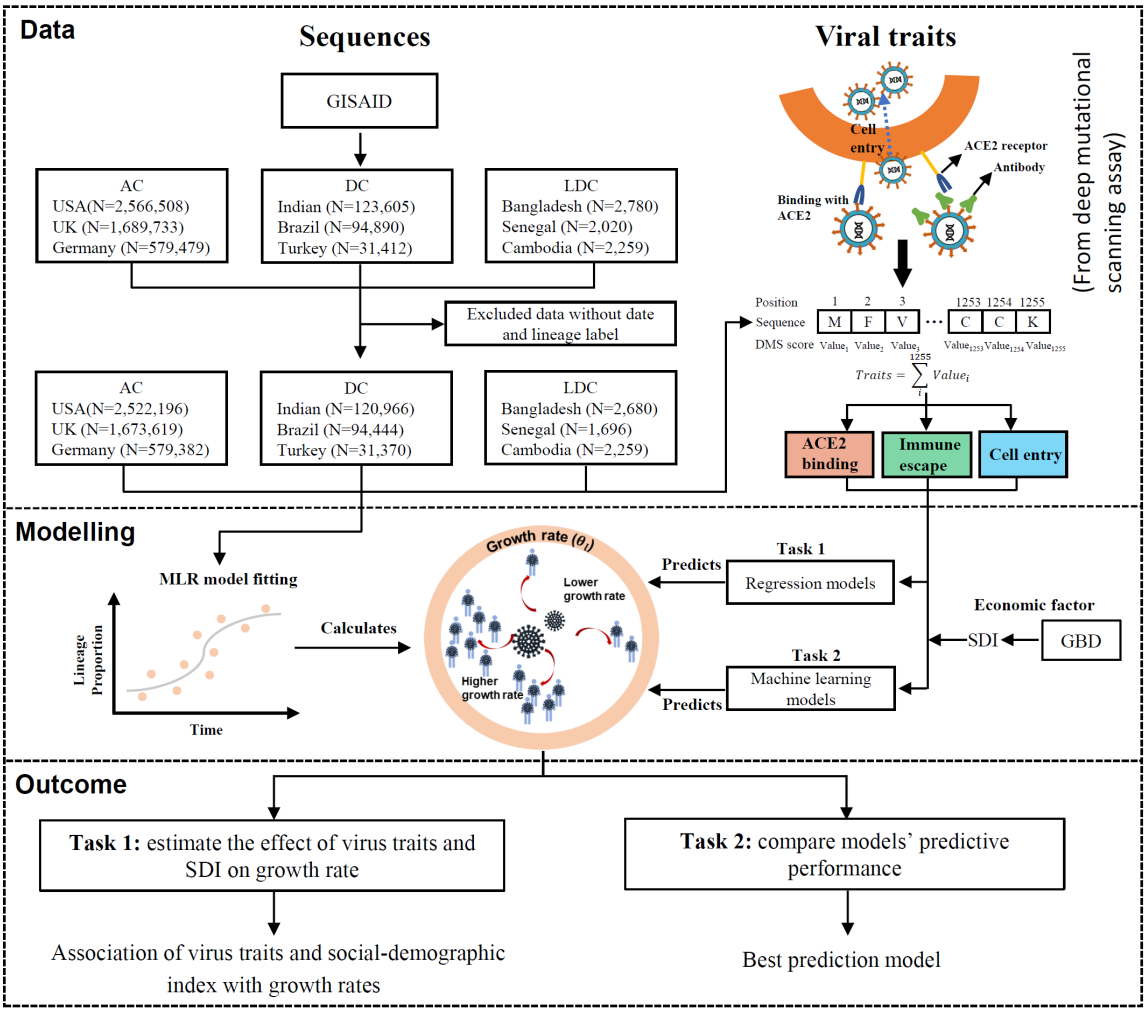

Figure S1. Workflow of virus trait analysis and growth rate estimation across nine countries. Sequence data from GISAID (ranging from November 1, 2020, to November 30, 2022) were gathered for the advanced countries (i.e., USA, UK, and Germany), developing countries (i.e., India, Brazil, and Turkey), and the least developed countries (i.e., Bangladesh, Senegal, and Cambodia). Viral growth rate estimation was conducted using MLR model. The viral traits, including ACE2 binding affinity, immune escape, and cell entry, were quantified based on DMS data. The association between the effects of virus traits and SDI was investigated. The prediction performance of multiple machine learning models was compared.

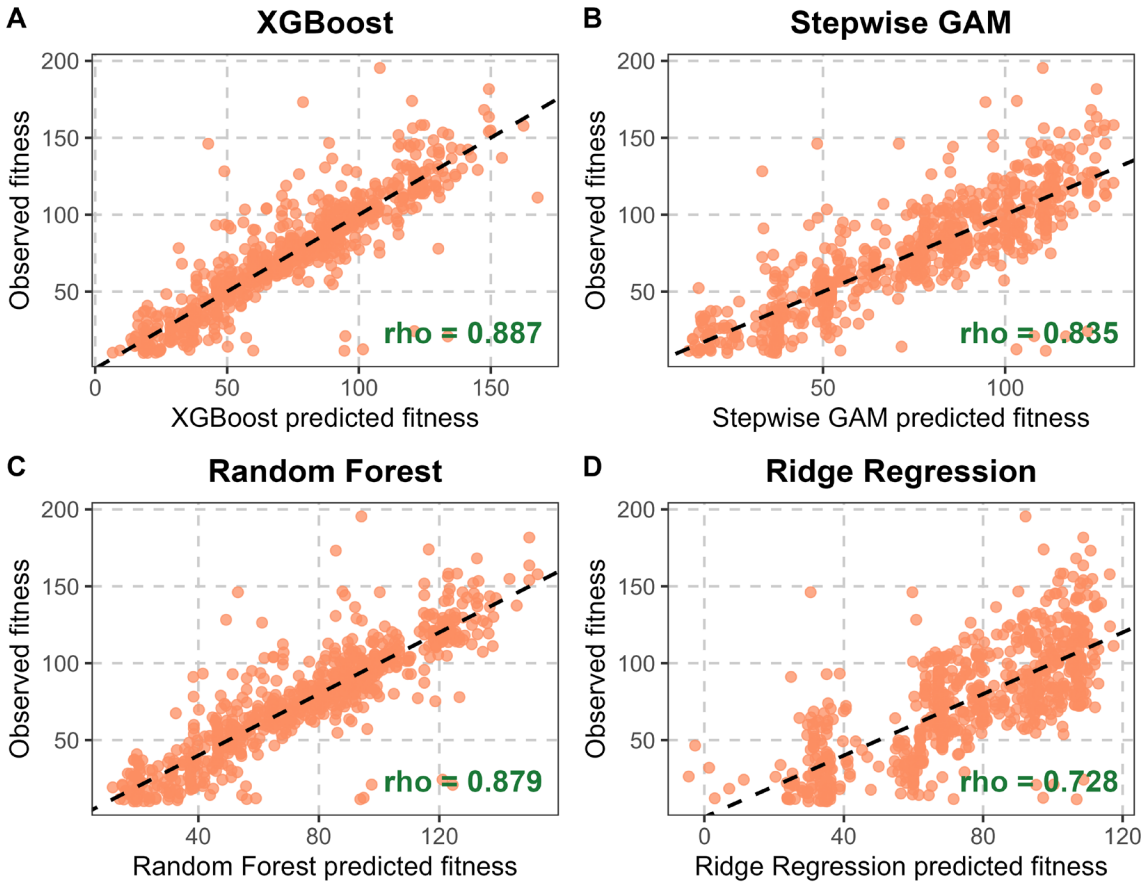

Figure S2 Performance of different machine learning models in predicting SARS-CoV-2 lineage fitness. Each panel shows the scatter plot of predicted versus observed lineage fitness (growth rate), with the black dashed line indicating the line of identity. The Pearson correlation coefficient (i.e., rho) is shown in green for each model. Additive ACE2 binding and antibody binding retained scores, along with cell entry score and SDI, were used as input features.

References

[1] V. Calvaresi, A. G. Wrobel, J. Toporowska, D. Hammerschmid, K. J. Doores, R. T. Bradshaw, R. B. Parsons, D. J. Benton, C. Roustan, E. Reading, and M. H. Malim. Structural dynamics in the evolution of sars-cov-2 spike glycoprotein. *Nature Communications*, 14(1):1421, 2023.

[2] D. Wang, M. Huot, V. Mohanty, and E. I. Shakhnovich. Biophysical principles predict fitness of sars-cov-2 variants. Proceedings of the National Academy of Sciences U.S.A., 121(23):e2314518121, 2024.

[3] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15):e2016239118, 2021.

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

[5] V. Calvaresi, A. G. Wrobel, J. Toporowska, D. Hammerschmid, K. J. Doores, R. T. Bradshaw, R. B. Parsons, D. J. Benton, C. Roustan, E. Reading, and M. H. Malim. Structural dynamics in the evolution of sars-cov-2 spike glycoprotein. Nature Communications, 14(1):1421, 2023.

[6] P. Han, L. Li, S. Liu, Q. Wang, D. Zhang, Z. Xu, P. Han, X. Li, Q. Peng, C. Su, and B. Huang. Receptor binding and complex structures of human ace2 to spike rbd from omicron and delta sars-cov-2. *Cell*, 185(4):630–640, 2022.