Article

A protein language model for exploring viral fitness landscapes

Received: 18 July 2024

Accepted: 22 April 2025

Published online: 13 May 2025

Check for updates

Jumpei Ito (1,2), Adam Strange¹, Wei Liu (1,3,4,19), Gustav Joas (1,5,19), Spyros Lytras (1,6), The Genotype to Phenotype Japan (G2P-Japan) Consortium* & Kei Sato (1,2,6,7,8,9)

Successively emerging SARS-CoV-2 variants lead to repeated epidemic surges through escalated fitness (i.e., relative effective reproduction number between variants). Modeling the genotype–fitness relationship enables us to pinpoint the mutations boosting viral fitness and flag high-risk variants immediately after their detection. Here, we present CoVFit, a protein language model adapted from ESM-2, designed to predict variant fitness based solely on spike protein sequences. CoVFit was trained on genotype–fitness data derived from viral genome surveillance and functional mutation assays related to immune evasion. CoVFit successively ranked the fitness of unknown future variants harboring nearly 15 mutations with informative accuracy. CoVFit identified 959 fitness elevation events throughout SARS-CoV-2 evolution until late 2023. Furthermore, we show that CoVFit is applicable for predicting viral evolution through single amino acid mutations. Our study gives insight into the SARS-CoV-2 fitness landscape and provides a tool for efficiently identifying SARS-CoV-2 variants with higher epidemic risk.

A primary challenge faced in controlling viral infectious diseases stems from the ability of viruses to evolve through mutations¹. Throughout the COVID-19 pandemic, SARS-CoV-2 variants with escalated spreading potential (i.e., fitness) in the host population have successively emerged, leading to repeated epidemic surges^{2,3}. The fitness of a virus is determined by multiple factors such as its ability to replicate efficiently within host cells, its capacity to evade innate immunity, and, most notably, its ability to escape host population-level immunity shaped by vaccination and natural infection. By understanding how viruses enhance their fitness in a pandemic through the lens of SARS-CoV-2 studies, we can learn critical insights for managing not just COVID-19 but future viral infectious diseases as well. SARS-CoV-2 can alter its fitness by acquiring mutations (including substitutions, insertions, and deletions) in viral proteins, with particular emphasis on the spike (S) protein⁴. The S protein is the glycoprotein essential for virus entry into the host cells via interaction with the angiotensin-converting enzyme 2 (ACE2) receptor⁵. Also, the S protein is a primary target for neutralizing antibodies (Abs), which are key components of the humoral immune response triggered by vaccinations or natural infections⁶. Therefore, mutations in the S protein that can affect its binding efficiency with ACE2 and its ability to evade neutralizing Abs tend to have a stronger impact on viral fitness^{2,3,7–19}.

In genetics, fitness is defined as a measure of an organism's reproductive success, passing on its genes to the next generation.

¹Division of Systems Virology, Department of Microbiology and Immunology, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ²International Research Center for Infectious Diseases, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ³Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. ⁴Swiss Institute of Bioinformatics, Geneva, Switzerland. ⁵Division of Immunology and Respiratory Medicine, Department of Medicine, Karolinska Institutet, Stockholm, Sweden. ⁶MRC-University of Glasgow Centre for Virus Research, Glasgow, UK. ⁷Graduate School of Medicine, The University of Tokyo, Tokyo, Japan. ⁸International Vaccine Design Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. ⁹Collaboration Unit for Infection, Joint Research Center for Human Retrovirus Infection, Kumamoto University, Kumamoto, Japan. ¹⁹These authors contributed equally: Wei Liu, Gustav Joas. *A list of authors and their affiliations appears at the end of the paper. e-mail: jampei@g.ecc.u-tokyo.ac.jp; KeiSato@g.ecc.u-tokyo.ac.jp Accordingly, the fitness of SARS-CoV-2 variants can be defined as the relative effective reproduction number (R_e) between variants (hereafter, "fitness" refers to relative R_e in this study)^{4,20–22}. R_e represents the average number of secondary infections caused by an infected individual conditioned by the host population's immunity, which is in turn shaped by the history of vaccination and natural infections. Owing to advancements in virus genome surveillance, it is now feasible to estimate the relative R_e of SARS-CoV-2 variants in almost real-time. Using temporal data on variant detection frequencies with a logistic model, we can estimate the R_e values of various variants relative to a reference variant, under certain assumptions-such as the relative $R_{\rm e}$ between variants remaining constant over time (also see the second-to-last paragraph of the Discussion section)⁷. Based on the estimated relative R_{e} , we can predict which variant(s) has the biggest advantage amongst co-circulating viruses at a given time and will likely become the next epidemic variant. Several research groups, including ours, have successfully utilized R_e or similar indicators to detect upcoming epidemic variants and further elucidated the characteristics and risks these variants possess through virological experiments²³⁻²⁸.

These conventional methods to estimate the fitness (relative R_e) of variants are based on the accumulation rate of variant sequences. This means that the fitness of a newly emerged variant cannot be estimated until a sufficient number (e.g., \geq 30) of sequences for that variant has accumulated. In contrast, if the fitness of variants can be predicted based on their genotypes, it would theoretically become possible to predict the fitness of a newly emerged, unknown variant as soon as a single sequence is obtained. Furthermore, by establishing a fitness prediction model, we can identify mutations that contribute to increased viral fitness under given conditions (e.g., specific immune landscape). Moreover, since viruses like SARS-CoV-2 tend to evolve in a direction that increases their fitness, understanding the virus's fitness landscape enables the prediction of their evolution.

Previous studies, including ours, have developed methods to predict the fitness (relative R_e) of variants based on their mutation patterns using a statistical modeling approach^{4,11,29,30}. However, these models simply represent fitness as a linear combination of individual mutation effects, not considering interactions between mutations, namely epistasis^{4,11,29,30}. Furthermore, these statistical models cannot consider the effect of mutations that have not yet emerged at the time of the training dataset creation. We speculated that these challenges can be addressed using protein language models, an innovative technology that has become increasingly popular in the field of natural language processing³¹. Protein language models are large language models pretrained on extensive datasets of protein sequences, allowing them to capture context-specific patterns within and across amino acid sequences. Protein language models can represent protein sequences as numerical vectors (i.e., embeddings) that reflect their characteristics and functions. By leveraging these embeddings, we can further develop a regression or classification model from a protein language model to address specific tasks, such as fitness prediction. Moreover, by employing a multitask learning framework, we can further inform the fitness prediction process with functional data on individual mutations, potentially enhancing the model's predictive accuracy.

In this study, we developed CoVFit, a model to predict the fitness (relative R_e) of SARS-CoV-2 variants based on the S protein by utilizing the state-of-the-art protein language model ESM-2³². We finetuned a customized ESM-2 model using (i) genotype–fitness information, estimated from virus genome surveillance, and (ii) individual mutation effect information on evasion ability from humoral immunity, determined by high-throughput deep mutational scanning (DMS) experiments³³. Using CoVFit, we have explored the current fitness landscape until late 2023 and the evolutionary potential of SARS-CoV-2.

Results

Introduction of CoVFit

We developed CoVFit, a fitness prediction model based on S protein sequences by finetuning ESM-2 (Fig. 1a). To increase the model's knowledge of the coronavirus S proteins, we first established ESM- $2_{Coronaviridae}$, by performing additional pretraining (i.e., domain adaptation) on the ESM-2 model with S protein sequences obtained from 1506 *Coronaviridae* viruses (Figs. 1b and S1). The ESM- $2_{Coronaviridae}$ model demonstrates enhanced predictive capability in the masked learning task specifically for SARS-CoV-2 S proteins, while retaining its performance across a broader collection of proteins (Fig. S1c). Subsequently, utilizing a multitask learning framework, we finetuned the model on both genotype–fitness (relative R_e) data and DMS data for the ability to escape neutralization by monoclonal Abs (mAbs) (Fig. 1b). Consequently, for a given S protein sequence, CoVFit can predict the country-specific fitness value and the ability to escape from each mAb (Fig. S1a).

To assemble the genotype–fitness dataset, we first classified viral sequences into S protein genotypes, defined as groups of viruses sharing a unique set of mutations in S protein, using the genome surveillance data up to November 2, 2023, obtained from GISAID. Subsequently, we estimated the relative R_e of each genotype in each country by fitting a multinomial logistic model to temporal data on genotype detection frequencies, as previously described^{2,3,7–19}. Consequently, we obtained a total of 21,281 genotype–fitness data points, covering 12,817 genotypes across 17 countries (Figs. 1b and S2a). Consistent with a previous study⁴, there is a clear trend where lateremerging variants exhibit higher R_e values, resulting in the continuous replacement of circulating variants (Fig. S2a). This result suggests that the average relative fitness of circulating variants increases over time likely in response to the rising levels of population immunity in the hosts.

We utilized an in vitro DMS dataset on the neutralization capabilities of mAbs, produced by Cao et al.²⁴. A total of 173,384 mutation–mAb DMS data points, covering 2096 types of mutations in the receptor binding domain (RBD) in the S protein and 1548 mAbs, were included in the dataset (Figs. 1b and S2b). Aligning with previous findings²⁴, the effects of mutations on mAbs varied depending on their epitope classes (Fig. S2b–d). Variants exhibiting higher fitness, such as BQ.1.1 or XBB, exhibited an increased ability to evade these mAbs, supporting the effectiveness of utilizing this information for predicting fitness (Fig. S2d).

Using the five-fold cross-validation scheme, we generated five model instances of CoVFit (Fig. 1b). These model instances were used to evaluate performance for the corresponding test data. Additionally, these instances can provide the mean and variance of predictions for new data (e.g., data acquired after model training). Hereafter these model instances are referred to as CoVFit_{Nov23} (subscript denoting the month and year when the genotype–fitness data were obtained).

Prediction performance of CoVFit

To evaluate the prediction performance of the CoVFit_{Nov23} model instances, we examined the prediction performance of respective model instances using the corresponding test datasets. We selected Spearman's rank correlation score as the main metric because we believe it is necessary to assess the model's ability to predict the ranking of relative fitness between variants in order to prioritize variants with high fitness that are likely to spread. The resulting prediction performance for fitness on data that doesn't require extrapolation is notably high (Spearman's correlation: 0.990) (Fig. 2a–c). Also, prediction of escape ability from neutralization by mAbs reaches moderately high performance (Spearman's correlation for each epitope class: 0.578–0.814) (Figs. 2a and S3a). The model's higher predictive performance was also supported in



Fig. 1 | **Overview of CoVFit. a** Conceptual framework of CoVFit. CoVFit is a protein language model designed to predict the fitness (relative *R*_e) of SARS-CoV-2 variants based on their S protein sequences. **b** Outline of the training process used to develop CoVFit model instances.



Fig. 2 | Prediction performance of CoVFit. a Spearman's correlation scores for predicted relative fitness values and mAb neutralization escape scores. Scores from five cross-validation folds are shown as dots, with the mean represented by a bar and the standard deviation by an error bar. The correlation for mAbs was calculated in each epitope group. b Scatter plot for fitness prediction, aggregating results from five-fold cross-validation. Dot denotes the result of a certain viral genotype in

a specific country. Dot is colored by the Nextclade clade. The relative fitness value was scaled so that the 0.1 percentile and 99.9 percentile points fall between 0 and 1. A dashed line with a slope 1 and intercept 0 is shown. **c** Scatter plot inherited from (**b**) but colored by the emergence date of each genotype. Source data are provided as a Source Data file.

evaluations stratified by sampling country and mAb type for fitness and immune escape ability, respectively (Fig. S3b, c). Even when variants were stratified by country of circulation and emergence time (on a weekly basis), a robust correlation was observed between fitness and predicted fitness (Fig. S3d, e). This result suggests that CoVFit can distinguish the fitness of variants emerging during the same period with an effective level of accuracy. Together, we show that CoVFit has sufficient power to represent the fitness landscape as well as the effect of mutations on evasion from diverse types of mAbs. **Prediction performance of CoVFit for unknown, future variants** This study aims to develop a model that can robustly predict the fitness of yet-to-emerge variants in addition to known variants. However, in the experiment shown in Fig. 2, where the dataset is randomly split into training and test sets, it is highly likely that variants in the test set are very similar to those in the training set. Indeed, we analyzed the minimum amino acid difference between variants in the test dataset and their closest counterparts in the training dataset (referred to as the minimum amino acid distance) and found that most variants have a distance of less than five amino



Fig. 3 | **Prediction performance of CoVFit for unknown, future variants. a** Strategy for evaluating prediction performance on future variants. Model instances, referred to as CoVFit_{Past}, were trained on variant data prior to a specified cutoff date (e.g., January 31, 2022). Prediction performance for future variants was then assessed using data from variants that emerged after this date. **b** Number of sequences from each clade in the past datasets with specific cutoff dates. **c** Fitness predictions for future (gray) and past (light gray) variants in the dataset with a cutoff date of February 28, 2022. Points represent results for each genotype, calculated as average values across countries and five-fold predictions. A dashed line with a slope of 1 and an intercept of 0 is included. **d** Fitness predictions for future variants, with colors indicating Nextclade clade classifications. In addition to the dashed line with a slope of 1 and intercept 0, a gray estimated regression line, based

on mean prediction values, is displayed. **e** Scatter plot based on (**d**) but colored according to the minimum amino acid distance from variants in the past data. **f** Predicted fitness of genotypes within each Nextclade clade. Each clade's distribution (violin) and median value (dot) are shown. Individual panels display results for datasets with different cutoff dates. Clades present in the past data are separated by a dashed vertical line from those absent in the past data. Additionally, the median observed fitness value of each clade is represented by a heatmap on the left side. **g** Comparison of prediction performance metrics across methods, including Spearman's correlation score, *R*-squared value, mean absolute error (MAE), and estimated regression slope. Source data are provided as a Source Data file.

acids (Fig. S3f). This suggests that the task may be easier than the real-world prediction task we aim to solve.

Hence, to evaluate the model's prediction performance for unknown, future variants (i.e., extrapolation capacity), we synthesized datasets for "past" and "future" variants by splitting the existing genotype-fitness dataset based on variants' emergence dates with a specific cutoff date (e.g., February 28, 2022) (Fig. 3a). Subsequently, we generated five instances of CoVFit solely using the past variant dataset with the five-fold cross-validation scheme and then evaluated its performance on the corresponding future variant dataset. We conducted the same experiment for these eight cutoff dates, which were spaced 1 month apart from January 31, 2022, to August 31, 2022 (Fig. 3b). The earliest cutoff date corresponds to just after the emergence of BA.2, while the latest cutoff date is just before the emergence of BQ.1 and XBB lineages (Fig. 3b). BQ.1 is a lineage that arose from BA.5 through a sequential evolutionary process, whereas XBB emerged through a non-sequential (or saltation-like) process, involving recombination in the S gene that resulted in the acquisition of 14 amino acid mutations.

We demonstrated that this series of trained model instances, CoVFit_{Past}, successfully predicted that the fitness of future variants exceeds that of past variants (Figs. 3c and S4a). Furthermore, CoVFit_{Past} showed a notable level of accuracy in predicting the relative fitness ranking of future variants (Figs. 3d, e and S4a). For example, CoVFit_{Feb22}, trained on data up to the emergence of BA.2.12.1, achieved a Spearman correlation score of 0.862 for the future dataset (Fig. 3d, e). Moreover, we calculated the median predicted fitness for each viral clade in the future dataset and found that CoVFit_{Feb22} accurately ranked the clade-median fitness of future variants, including BA.4, BA.5, BA.2.75, BQ.1, XBB, XBB.1.5, EG.5.1, and HK.3 (Fig. 3f). Given that the XBB lineages has \geq 14 amino acid differences from its closest genetic counterparts in the past dataset, CoVFit's extrapolation capacity is notably effective (Fig. 3e). Although prediction accuracy tended to decrease for variants with greater amino acid distances from those in the past (training) dataset, this relationship was neither linear nor strictly monotonic (Fig. S4b). Finally, consistent results were generally observed across datasets with varying cutoff dates, highlighting CoVFit's extrapolation capacity for previously unknown, future variants (Figs. 3f and S4a).

However, we also found that CoVFit does not exhibit high generalization ability in some scenarios. For instance, CoVFit trained on datasets that either completely lack or contain only a few sequences of BA.2.75 (before June 30, 2022) tended to overestimate the fitness ranking of BA.2.75 with the K356T mutation, likely due to this mutation's high immune escape effect by K356T (Figs. 3f and S4a)³⁴. Furthermore, in later cutoff dates (e.g., August 31, 2022), CoVFit tended to underestimate the fitness value of XBB subvariants, although CoVFit was able to correctly rank the fitness values of these variants (Fig. S4a; also see the regression slope in Fig. 3g). Moreover, in some cutoff dates (e.g., August 31, 2022), CoVFit underestimated the fitness ranking of BA.2.86, a variant that emerged through a saltation-like evolutionary process involving 30 mutations in the S protein and exhibited one of the highest fitness levels in late 2023 (Fig. 3f). This underestimation was apparent even when CoVFit was trained on data up to just before the emergence of BA.2.86 (e.g., July 31, 2023) (Fig. S4c). These results show that while CoVFit has high extrapolation capacity, there are limitations to its generalization ability, where prediction accuracy can be inconsistent between training datasets for some more divergent protein contexts.

Performance comparison with other prediction models

Next, we evaluated CoVFit's performance relative to other currently available methods. A previous study by Obermeyer et. al established PyR_0 , a statistical model that can predict the fitness (relative R_e) of unknown variants based on their mutation profiles⁴. Thus, we began by comparing CoVFit with PyR₀. Two versions of the PyR₀ model instances are available: one trained on data up to January 20, 2022 (covering variants up to BA.2), and another trained on data up to September 19, 2022 (just before the emergence of BQ.1 and XBB). While these PyR₀ models achieved an effective level of accuracy in predicting the fitness of past variants, their Spearman correlation scores were negative when applied to future variants (Fig. S5). Particularly, PyR₀ model instances underestimate variants with greater amino acid distances from those in the past dataset, such as XBB lineages (Fig. S5a, b, right). These results suggest that PyR₀ was unable to accurately rank future variant fitness in our experimental setting, suggesting that CoVFit has a higher extrapolation capacity than PyR₀. This outcome likely reflects PyR₀'s design as an interpretable statistical model to identify mutations critical for fitness gains.

To gain additional insight into the extrapolation capacity of CoVFit, we constructed alternative prediction models using non-deep learning methods, including LASSO, Random Forest, and Light Gradient Boosting Machine (LightGBM) (Fig. 3g). These models demonstrated prediction performance comparable to CoVFit in terms of Spearman correlation, *R*-squared, mean absolute error (MAE), and estimated the regression slope when applied to past variants. However, when predicting the fitness of future variants, CoVFit outperformed these models on most metrics and cutoff dates, suggesting that CoVFit has a higher extrapolation capacity than all these models.

To assess the impact of incorporating DMS data into CoVFit on the efficacy of fitness prediction, we generated an additional model instance, CoVFit_{noDMS}, by training without the DMS dataset and subsequently evaluated its predictive performance against the original model (Fig. 3g). In predicting the fitness of past variants, CoVFit_{noDMS} exhibited prediction performance similar to the original CoVFit. However, CoVFit_{noDMS} substantially underperformed in predicting the fitness of future variants across most metrics and cutoff dates. Similarly, we examined the impact of the domain adaptation step on prediction performance and demonstrated the contribution of this step in achieving higher performance (Fig. 3g). Removing the DMS dataset had a much stronger effect than omitting the domain adaptation step, underscoring the critical role of DMS dataset incorporation. Collectively, CoVFit's higher extrapolation capacity is likely attributed to its incorporation of functional information on mutations and the application of domain adaptation frameworks.

Fitness elevation events during SARS-CoV-2 evolution

To deepen our understanding of the fitness landscape of SARS-CoV-2, we developed a CoVFit-based phylogenetic framework to analyze fitness elevation throughout its evolution (Fig. 4). First, we constructed a phylogenetic tree of 11,098 variants, which correspond to viral genome sequences encoding respective S protein genotypes. Subsequently, ancestral S protein sequences at the internal nodes of the tree were reconstructed (Fig. 4a). We then inferred the fitness of all nodes, representing both observed and reconstructed ancestral sequences, utilizing the latest CoVFit_{Nov23} models (Fig. 4b and Fig. S6 for Omicron and all lineages, respectively). We obtained five predicted fitness values per node via a five-fold cross-validation scheme. Finally, we identified branches where fitness elevation was statistically significant (false discovery rate, FDR < 0.1) by comparing predicted fitness values between a given node and its parent node. Of the 9846 branches that acquired mutations in the S protein, 959 (9.7%) branches were identified with significant fitness elevation (Fig. S6), including 542 branches within the Omicron lineages (Fig. 4c). We observed increases in viral fitness both in the branches representing the most recent common ancestor of major lineages and throughout their subsequent diversification (Figs. 4c and S6).

To identify mutations critical for fitness elevation, we next calculated the average fitness gain per mutation by examining branches with the acquisition of a single mutation (Fig. 4d and Fig. S6 for Omicron and all lineages, respectively). Mutations with higher fitness gain [per mutation] score are predominantly found in the RBD of the S protein, particularly in its receptor binding motif (RBM), consisting of the binding interface to ACE2³⁵ (Figs. 4e, f and S6). Additionally, we identified significant non-RBD mutations, like the T19I and Q52H, in the N-terminal domain. We found that mutations with a greater impact on fitness elevation also tend to enhance the virus's ability to evade humoral immunity (Fig. 4g). Furthermore, mutations with significant fitness impact in Omicron lineages tend to have been acquired multiple times in a convergent manner throughout Omicron's evolution (Fig. 4h). This association was evident in RBD mutations (Nagelkerke's pseudo R²: 0.392) but not in non-RBD mutations (Nagelkerke's pseudo R^2 : 0.008).

Context-specific effects of the F456L substitution

We found that acquisitions of some mutations were overrepresented in a specific phylogenetic lineage. For example, while the R346T substitution was convergently acquired across Omicron lineages, the F456L substitution was markedly overrepresented in the XBB lineage, with few occurrences in the BQ.1 lineage (Fig. 5a). To quantify differences in the fitness effect of F456L among lineages, we performed in silico mutational scanning analysis leveraging CoVFit_{Nov23} by computationally inducing F456L in various S protein backbones and inferring the fitness gain caused by this substitution (Fig. 5b). The predicted



Fig. 4 | Detection of fitness elevation events during Omicron diversification. a Scheme to detect phylogenetic branches with fitness elevation utilizing CoVFit models. b Inference of change in fitness through Omicron's evolution. The maximum likelihood (ML) tree of Omicron lineages is shown. Branch color indicates an inferred fitness value for each phylogenetic node, including both observed and reconstructed ancestral genotypes of S proteins in the phylogenetic tree. c Detection of fitness elevation events during Omicron's evolution. Dot color indicates inferred fitness gain in each branch, calculated as the difference in predicted fitness between a node and its parental node. d Mean fitness gain over a specific mutation during Omicron evolution. Since some mutations have been acquired multiple times, the mean value of fitness gain among acquisition events was used as the "fitness gain [per mutation]" score. The top 20 mutations regarding this score are shown with the protein domain information. e Enrichment of fitnessassociated mutations in the RBD, particularly in its RBM. The negative score is clipped to 0. f Mapping the site-wise fitness gain score on the 3D structure of the ancestral D614G S protein (PDB: 7BNN)56. If multiple mutation types are present in a

fitness gain from F456L was markedly higher in the XBB lineage, followed by BQ.1, with other lineages showing substantially lower gains. Together, these results suggest that the fitness-increasing impact of F456L is specific to the XBB lineages (and possibly to BQ.1 as well).

To gain mechanistic insights into the XBB-specific F456L effect on viral fitness, we analyzed published DMS data. This includes the data on mAb neutralization by Cao et al.²⁴ and those on the RBD ACE2binding affinity and protein stability by Taylor and Starr³⁶. Substitutions at F456 has one of the largest impacts on neutralization escape, according to the escape estimator data³⁷, in various lineages including the ancestral D614G strain's S protein (Fig. 5c). On the other hand, the effects of this mutation on ACE2 binding and protein expression were different among S protein backbones. While F456L enhances ACE2 binding and protein expression in the XBB S protein, this substitution has a negative effect on ACE2 binding and/or protein expression in all tested S backgrounds other than XBB (Fig. 5d). The non-deleterious specific site, the maximum value is shown as the "fitness gain [per site]" score. Amino acid side chains for the top 15 sites regarding this score are shown as sphere. The plot was generated using Chimera X⁵⁷. **g** Association of fitness gain rank with the mean mAb escape score. This escape score was calculated as the mean of the escape score across mAbs over a mutation. The ND group includes mutations not observed in our phylogenetic analysis. The categories 1–50, 51–100, 101–, and ND include 39, 24, 75, and 1964 entries, respectively. The box represents the interquartile range (IQR; 25th to 75th percentile), with the horizontal line indicating the median (50th percentile). The whiskers extend to the smallest and largest values within 1.5 times the IQR from the lower and upper quartiles, respectively. **h** Association of the fitness gain [per mutation] score with the inferred acquisition count. The estimated regression curve (line) with standard error (ribbon) by Poisson regression using all mutations is shown. In addition, Nagelkerke's pseudo *R*² values for Poisson regression analyses using all mutations, RBD mutations, and non-RBD mutations are shown. Source data are provided as a Source Data file.

effect of F456L in ACE2 binding and expression, unique to XBB, has been confirmed in both RBD and full S DMS assays in previous studies^{24,38}. These results suggest that F456L confers preferable effects on the XBB's S protein but confers a double-edged sword effect on other lineages' S protein. Together, the XBB-specific positive effect of F456L on fitness can be explained by the XBB-specific removal of the deleterious effects of this mutation. This example validates the effectiveness of CoVFit to infer mutational effects in a context-specific manner.

CoVFit-based in silico DMS to predict subsequent mutations

To evaluate the potential of CoVFit in predicting viral evolution, we developed a simulation method, namely CoVFit-based in silico DMS. In this method, we computationally introduced every possible single amino acid substitution into the S protein sequence of a specific variant. Subsequently, the fitness gain from each substitution was inferred using CoVFit.



Fig. 5 | **Context-specific effect of the F456L substitution.** a Examples of convergent acquisitions of specific substitutions. A node indicates the acquisition events, and node color denotes fitness gain at the acquisition events. Branch color denotes the presence (gray) or absence (light gray) of specific substitutions in the reconstructed ancestral S protein sequences. b Fitness gain upon F456L in each backbone S protein sequence, inferred by in silico mutational scanning using CoVFit. Variants with available DMS data (shown in (d)) were included in this

analysis. **c** Site-wise immune escape score for the ancestral D614G strain, BA.2, and XBB variants, estimated by mAb escape estimator³⁷ based on Cao's DMS data²⁴. The top 5 sites regarding the escape score are annotated. **d** Effect of F456L on the S protein's expression (stability) and ACE2-binding affinity, extracted from publicly available DMS data from Taylor and Starr³⁶. The dot color indicates inferred fitness gain shown in (**b**). Higher values indicate enhanced higher expression and ACE2-binding affinity values. Source data are provided as a Source Data file.

In this study, we selected BA.2.86.1 as the target for in silico DMS analysis to evaluate CoVFit's performance in predicting the evolution of the BA.2.86.1 lineage. Following the emergence of BA.2.86.1, its descendant lineage, JN.1, arose through the L455S mutation and rapidly spread worldwide. JN.1 demonstrated higher fitness than both BA.2.86.1 and the XBB subvariants circulating at the time, leading to a near-complete replacement of the dominant lineage from XBB to the BA.2.86 lineage by early 2024²³. Subsequently, JN.1 underwent further diversification, giving rise to various subvariants, such as KP.2 and KP.3.

The genotype-fitness dataset we used included one sequence of JN.1 and 11 sequences from other BA.2.86 lineages. To create a model that completely excluded knowledge of JN.1, we removed its sequence from the dataset and trained a new CoVFit model instance, referred to as CoVFit_{woJN1}, for conducting in silico DMS targeting BA.2.86.1. Finally, we calculated per-site fitness gains based on the in silico DMS analysis.

The results showed that the top 2% (20 sites) with the highest predicted fitness gains were significantly enriched in the top 2% of sites with the highest mutation frequencies within the BA.2.86.1 lineage (odds ratio: 29.4; P<1.0E-6) (Fig. 6a). This finding suggests that mutations with the highest fitness gains predicted by our model were indeed the ones acquired and subsequently in the BA.2.86 lineage. Notably, the mutation at site 455 (ranked 3rd; e.g., L455S) was the first to rapidly fix within the viral population, coinciding with the expansion of JN.1 (Fig. 6b, c). Similarly, the mutation at site 456 (ranked 1st; e.g., F456L) became fixed second in the population during the spread of KP.2 and KP.3, descendants of JN.1 that harbored both L455S and F456L. Furthermore, the mutation at site 346 (ranked 6th; e.g., R346T) initially increased in frequency alongside the expansion of KP.2, which carried L455S, F456L, and R346T. However, its frequency later declined as KP.3, which carried L455S and F456L but lacked R346T, outcompeted KP.2. Moreover, as of December 2024, JN.1 subvariants harboring R346T in addition to L455S and F456L-such as LP.8.1-have emerged and spread rapidly worldwide³⁹. These findings suggest that CoVFitwoINI successfully predicted, with an effective level of accuracy, the mutations that were likely to be acquired and spread within the BA.2.86 population. This highlights the utility of CoVFit for predicting viral evolution through single amino acid mutations.

CoVFit-CLI tool

The trained CoVFit_{Nov23} model instances used in this study are available as a command-line tool, CoVFit-CLI, from our GitHub repository (https://github.com/TheSatoLab/CoVFit). Additionally, the instances trained on the latest genome surveillance data, CoVFit_{Nov24}, are also available. Researchers can use the standalone program to conveniently provide fitness and DMS escape predictions for their own SARS-CoV-2 S protein sequences. The CoVFit-CLI tool will receive periodic updates to the model instances, trained on the latest genomic surveillance data.

Discussion

In this study, we established CoVFit, a protein language model to predict the fitness (i.e., relative R_e) of SARS-CoV-2 variants, which represents a variant's relative ability to spread in the host population. Previous studies have proposed statistical models predicting viral fitness^{4,11,29,30} as well as machine learning models predicting viral phenotypes strongly associated with fitness, such as immune evasion ability^{40,41} or protein *grammaticality*^{42,43}. However, to our knowledge, CoVFit is the first machine learning model designed to directly predict viral fitness beyond conventional statistical models.

We demonstrated that while CoVFit exhibits higher extrapolation capacity compared to other methods, it also has certain limitations. CoVFit was able to predict the fitness of future, unknown variants, such as XBB and its descendant lineages, which are ~15 amino acids distant from the training data, with reasonable accuracy. This strong extrapolative ability likely stems from CoVFit's integration of multitask learning with DMS data, as indicated by ablation experiments (Fig. 3g). However, CoVFit was unable to consistently predict the fitness ranking of BA.2.86, which has >30 amino acid differences compared to variants in the training data (Figs. 3f and S4c). In other words, CoVFit appears capable of extrapolating to moderate saltation-like events (e.g., the emergence of XBB) but struggles with larger saltation-like events (e.g., the emergence of BA.2.86, which is comparable to the emergence of Omicron in the number of spike mutations). Furthermore, we found



Fig. 6 | **CoVFit-based in silico DMS on the BA.2.86.1 lineage. a** Association between the fitness gain [per site] score and the mutation frequency at each site in the BA.2.86.1 lineage. Points represent amino acid sites, while dashed lines indicate the 98th percentile (top 2%) for both the fitness gain score and mutation frequency. Statistical measures quantifying the degree of overlap between data points within the top 2% for these two metrics are shown. The *p* value was calculated using a two-sided Fisher's exact test. **b** Temporal trend in mutation frequency at individual

amino acid sites within the BA.2.86.1 population. The genome surveillance data from October 1, 2023, to July 31, 2024, was used. Frequencies were calculated using 7-day bins. **c** Temporal trends in viral lineage frequencies within the BA.2.86.1 population. Each viral lineage category includes its descendant lineages unless those descendant lineages are explicitly defined as separate categories. Mutations in the S protein relative to BA.2.86.1 are indicated, with emphasis on those with higher fitness gain [per site] scores. Source data are provided as a Source Data file.

that CoVFit tended to underestimate the fitness values of XBB subvariants in later cutoff dates (e.g., August 31, 2022), although it was still able to correctly rank their relative fitness values (Fig. S4a). Collectively, this study provides comprehensive insights into both the utility and limitations of CoVFit.

The CoVFit model design is expected to be most effective in the mid-to-late stages of a pandemic, when sufficient genome data has been accumulated, but surveillance is declining. As of 2024, with the SARS-CoV-2 pandemic persisting over a prolonged period of time, it has become increasingly difficult to sustain intensive viral genome surveillance (https://gisaid.org/hcov-19-variants-dashboard/). Conventional methods estimate fitness based on the accumulation rate of variant sequences. Therefore, to apply these methods effectively, a large number of viral genomes must be continuously sequenced and shared without delay. In contrast, CoVFit leverages past genomic information to directly predict fitness from genotype, enabling immediate predictions without waiting for new variant sequences to accumulate. Moreover, CoVFit's predictions do not rely on variant classification, eliminating the need for the processes of variant identification, classification, and naming, which require manual curation by experts (https://github.com/cov-lineages/pango-designation). A CoV-Fit model instance can be trained within 24 h using a single Nvidia A100 GPU, making it feasible to regularly update the model with the latest genome surveillance data. Thus, deploying the latest CoVFit model to monitor viral genome databases presents a promising strategy for improving the efficiency of high-risk variant surveillance, a task that remains essential even in a prolonged pandemic scenario. Moreover, based on the reasons outlined above, CoVFit-like methods would also be useful for monitoring pathogens already circulating in humans, such as Influenza, RSV, and common cold coronaviruses, where there's sparse sequencing data covering a wider period of their evolution.

In this study, we developed a framework that combines CoVFit with a phylogenetic approach to identify mutations that enhance viral fitness (Figs. 4 and 5). Despite the common challenge of interpretability in machine learning models, our framework is designed to assess not just the average effect of specific mutations but also their

context-specific or epistatic effects. For instance, it can distinguish the varying impacts of a particular mutation (e.g., F456L) across different contexts (e.g., within the XBB lineage versus other lineages) (Fig. 5). This framework is highly versatile and can be applied in combination with various phenotype prediction models beyond CoVFit. Compared to in silico DMS, this approach aims to retrospectively interpret observed evolutionary steps and identify patterns in their occurrence.

Previous studies have reported that some mutations are convergently acquired through the SARS-CoV-2 evolution, probably because these convergent mutations confer a positive effect on viral fitness^{24,44}. Indeed, some studies have inferred the importance of mutations for viral fitness according to how often these mutations are acquired convergently, under the assumption that these two features are correlated^{24,45}. However, despite our prior investigations focusing on a limited set of mutations¹¹, the degree of association between a mutation's acquisition frequency and its impact on fitness has remained unclear. In this study, we comprehensively investigated this issue and found an association between these two features, with a particularly strong correlation observed in mutations within the RBD (Fig. 4h). This result suggests, for accurate predictions of viral evolution, it is effective to consider both the acquisition frequency of mutations and their fitness effects. By steering future development towards a machine learning model based on a phylodynamics approach integrating both factors, we could potentially achieve precise predictions and simulations of viral evolution.

Despite the utility of CoVFit, it still has several limitations in its present form. First, as the objective variable, we utilized the fitness (relative R_e) of a variant, which was estimated from viral genome surveillance data and inherently carries a degree of uncertainty and bias. This bias arises partly because the surveillance does not employ random sampling. Often, samples from specific infection clusters are disproportionately represented in the surveillance data, potentially leading to a biased estimation of the fitness of certain variants. Second, the logistic model used to estimate variant fitness operates under a strong assumption that the relative fitness among variants remains constant over time^{4,20–22}. Although this model is widely used, this

assumption might not always hold true. In reality, the relative fitness of variants can vary in response to changes in the host population's immune status, influenced by factors such as natural infections and vaccinations^{2,3}. For example, in 2024, when the majority of the population has some level of immunity to SARS-CoV-2, the impact of immune escape on fitness is likely more pronounced compared to the early stages of the COVID-19 pandemic in 2020. Similarly, although it is likely that the effects of mutations on fitness may also change over time, the method to identify fitness-elevating mutations with CoVFit does not account for this temporal variation. Third, in our fitness prediction task, we utilized sequences from real-world variants as training inputs. Mutations with a substantial negative impact on fitness must be underrepresented in our training dataset since such mutations are usually eliminated through natural selection. Consequently, CoVFit is likely to underestimate the negative effects of certain mutations on fitness. Finally, because CoVFit requires a large amount of viral genome sequences for training, its utility would be limited in the early stages of future viral outbreaks, before sufficient genomic data has been accumulated. In the early phases, methods like EVEscape⁴¹, which use expert-designed scoring systems to estimate immune escape potential without requiring a training process, may be more practical. Other technical limitations and their potential solutions are discussed in the "Room for improvement of CoVFit" section in the "Methods" section. Together, the fitness predictions made by our model should be carefully interpreted in the context of complimentary information.

Despite the limitations mentioned, CoVFit holds the potential to decipher the fitness landscape of viruses. Our approach is poised to contribute to the development of innovative methods for the early prediction of future epidemic variants and for advancing viral evolutionary predictions. These advancements are critical for efficient epidemic control, vaccine development, and drug discovery. Furthermore, the methodologies employed in CoVFit can be applied to predicting the fitness of other viruses, including viral pathogens causing future pandemics. In anticipation of the next pandemic, it is imperative to continually develop foundational bioinformatics methods that assist in epidemic control, leveraging the extensive genomic data efforts of SARS-CoV-2 as an archetypical forerunner.

Methods

Preparation of genotype-fitness dataset

We retrieved all SARS-CoV-2 genome sequences and their associated metadata available as of November 2, 2023, from GISAID. We then assigned the most recent PANGO lineage classification available at that time to each sequence in our dataset using Nextclade v.2.14.0⁴⁶ with Nextclade dataset version "2023-10-26T12:00:00Z." Then we excluded low-quality sequences based on the following criteria: (i) absence of collection date information; (ii) samples derived from animals other than humans; or (iii) more than 1% undetermined nucleotide characters.

To develop a classification system for SARS-CoV-2 with a higher resolution than the PANGO lineage and based solely on the sequence of the S protein, we defined the genotype of the S protein and utilized it as the virus classification system in this study. The S protein genotypes refer to groups of viral sequences that share a unique set of mutations in the S protein. To achieve this, we first identified mutations in the S protein observed in more than 100 sequences. We then analyzed the mutation patterns across each sequence, enabling us to categorize sequences into genotypes based on these patterns. Only genotypes represented by 20 or more sequences in any country were considered for our analysis. As a result, a total of 13,643 S protein genotypes were included in our dataset. Each genotype was linked to the Nextclade PANGO lineage, Nextclade clade, and a representative genome sequence. This representative sequence was randomly chosen from the collection of sequences of the genotype. The emergence date of a genotype was determined as the 1st percentile date of collection for the viral isolates within the genotype. The viral genome sequences contained within this dataset are summarized under the EPI_SET_ID: EPI_SET_240311ma, which can be accessed through the GISAID website (https://gisaid.org/).

To estimate the relative R_e value of each genotype in each country, we began by tallying the daily count of each genotype within each country's dataset. We applied a multinomial logistic model to the count data of each country as previously described⁸. In this model, time (date) was used as the explanatory variable and the detection count for each variant as the dependent variable. Parameters in the model were estimated using the maximum likelihood method with the "multinom" function in the "nnet" package v.7.3.18 in R v.4.2.1.

In estimating R_e within the multinomial logistic model, it is necessary to select one variant as the reference variant with $R_e = 1$. Ideally, the reference variant should have the largest possible number of sequences. Additionally, a common reference variant must be chosen for all countries included in the analysis. To do this, we first counted how many sequences of each variant were detected in each country. Then, we compared the detection counts across countries for each variant and calculated the minimum count for each variant across countries. Finally, the variant with the highest minimum count was selected as the common reference variant for all countries. This approach allows us to choose a reference variant with the largest number of sequences registered across all countries, or at least in the country with the fewest sequences. Accordingly, the major genotype of the BA.5 lineage (equivalent to BA.5.2.1 in the PANGO lineage) was selected as the reference.

We then extracted the estimated growth rate (slope parameter) of each genotype relative to a reference genotype. The relative R_e of the viral lineage l, r_l , was calculated according to the slope parameter β_l of the lineage as $r_l = \exp(\gamma\beta_l)$, where γ is the average viral generation time (2.1 days) (http://sonorouschocolate.com/covid19/index.php? title=Estimating_Generation_Time_Of_Omicron)⁷. The estimated R_e value for each genotype is summarized in Supplemental Data 1.

Estimating relative $R_{\rm e}$ with the multinomial logistic regression model has certain limitations. The first is that it can only estimate relative R_e values between variants; in other words, it cannot provide absolute (non-relative) values of R_{e} . As a result, we cannot determine solely based on these values whether an epidemic caused by a particular variant will expand or subside. In this model, a variant's disappearance from the viral population (its relative frequency becoming negligible) is driven only by the expansion of other variants with a higher $R_{\rm e}$. The second limitation of this model is the assumption that the relative R_e between variants remains constant over time. In reality, the relative fitness of variants can fluctuate in response to changes in the host population's immune status, influenced by factors such as natural infections and vaccinations. In 2024, when most of the population has some immunity to SARS-CoV-2 due to vaccination or prior infection, the impact of immune escape on fitness is likely more pronounced compared to the early stages of the COVID-19 pandemic in 2020. The third limitation of this model is the assumption that the average viral generation time remains constant across variants. This assumption does not always hold, as previous studies have estimated differences in generation time between the Delta and Omicron variants⁴⁷. Therefore, it is essential to interpret the estimated relative $R_{\rm e}$ values with these limitations in mind.

In each viral clade, a small number of viral genotypes had emergence dates that were exceptionally earlier than the majority of other genotypes. For instance, the earliest emergence date in clade 21 K (BA.1) is associated with hap_89965, which is recorded as January 1, 2021. However, considering that BA.1 was first identified in November 2021, this date is likely erroneous. Additionally, in the analysis of prediction performance for future variants described later, it is preferable for the past-future cutoff dates (set at 1-month intervals) to align as

https://doi.org/10.1038/s41467-025-59422-w

closely as possible with the actual emergence dates of the viral clades. Therefore, we excluded a small number of these earliest genotypes with anomalously early emergence dates from the genotype–fitness dataset. Specifically, genotypes with emergence dates earlier than November 30, 2021 for clade 21K (BA.1); December 31, 2021 for clade 21L (BA.2); March 31, 2022 for clades 22A (BA.4) and 22B (BA.5); January 31, 2022 for clade 22C (BA.12.1); May 31, 2022 for clade 22D (BA.2.75); August 31, 2022 for clades 22E (BQ.1) and 22F (XBB); November 30, 2022 for clade 23A (XBB.1.5); February 28, 2023 for clade 23B (XBB.1.16); September 30, 2022 for clade 23C (CH.1.1); and May 31, 2023 for clade 23F (EG.5.1) were removed from the dataset.

Prior to the training step, we excluded S protein sequences containing more than 5 ambiguous characters and more than 30 amino acid deletions from the dataset. Furthermore, we excluded viral genotypes classified as Recombination clades (i.e., recombinant lineages, excluding major recombinant lineages such as XBB) in the Nextclade classification. Also, we focused on countries where more than 300 genotypes were detected, which led to collecting data for 17 countries: Australia, Belgium, Brazil, Canada, Denmark, France, Germany, India, Italy, Japan, Netherlands, South Korea, Spain, Sweden, Switzerland, the UK, and the USA. As a result of this additional filtering, a total of 21,281 genotype–fitness (relative R_e) data points, encompassing 12,817 genotypes across 17 countries, were included in our genotype–fitness dataset.

The estimated fitness value was transformed using the natural logarithm function, and then the data was scaled so that the 0.1 percentile and 99.9 percentile points fall between 0 and 1 before model training.

Preparation of DMS data for mAbs evasion

In this study, we utilized DMS data on evasion from mAbs provided by Cao et al.²⁴. The processed DMS data, specifically the mutation-wise immune escape score prepared for the antibody-escape estimator developed by Greaney et al.37, was downloaded from the Bloom lab GitHub repository on April 11, 2023 (https://github.com/ibloomlab/ SARS2 RBD Ab escape maps/blob/main/processed data/escape data mutation.csv). We applied specific exclusion criteria to the DMS data: (i) mAbs categorized as "SARS convalescents" and "WT-engineered"; and (ii) mAbs with an IC₅₀ value \geq 10, indicative of very weak binding affinity, for the target virus. The escape score in this repository was calculated using a DMS experiment using the ancestral D614G strain's RBD. Following the methods of Greaney et al., we defined a weighted escape score for each target virus (e.g., D614G and BA.2) from this escape score, following the method of Greaney et al. Specifically, the escape score was multiplied by the IC₅₀ value for the S protein of the target virus, followed by negative log transformation with a pseudo count of 1. The weighted escape score was scaled so that the 0 and 95 percentiles fell within the range 0-1, and values above 95 percentile were clipped to 1. For the comprehensive training of CoVFit, the weighted escape values for D614G were employed. On the other hand, considering that variants predominantly circulating after early 2022 are related to the BA.2 lineages, the weighted escape values for BA.2 were used in the training for CoVFit_{Past}.

Dataset preparation for domain adaptation

The S protein sequences for *Coronaviridae*, except for SARS-CoV-2, were downloaded from the NCBI Identical Protein Groups database (https://www.ncbi.nlm.nih.gov/ipg) on July 3, 2023, using the following search query: query: ("Alphacoronavirus"[Organism] OR "Betacoronavirus"[Organism] OR "Gammacoronavirus"[Organism] OR coronavirus[All Fields]) AND (spike[All Fields] OR S[All Fields] OR surface[All Fields]) NOT ("Severe acute respiratory syndrome coronavirus 2"[Organism] OR ("Severe acute respiratory syndr

("Severe acute respiratory syndrome coronavirus 2"[Organism] OR SARS-CoV-2[All Fields])))))))) NOT ("unidentified"[Organism] OR ("unidentified"[Organism] OR ("unidentified"[Organism] OR ("unidentified"[Organism] OR ("unidentified"[Organism] OR "Unknown"[All Fields])))) NOT "unidentified human coronavirus"[Organism] NOT ("synthetic construct" [Organism]) AND ("1000" [SLEN]: "1500" [SLEN]). The metadata associated with these sequences were also downloaded. The S protein for the SARS-CoV-2 Wuhan-Hu-1 strain was downloaded using NCBI Datasets Command-line tools v.15.6.1 (https://www.ncbi.nlm. nih.gov/datasets/docs/v2/download-and-install/) and subsequently incorporated into our dataset. Sequences with more than 5% unidentified amino acids were filtered out. Next, we removed redundant sequences using CD-HIT v.4.8.148 with a clustering threshold of 99% sequence identity. However, even after the CD-HIT filtering, a large number of sequences (i.e., 796 sequences) corresponding to the porcine epidemic diarrhea virus (PEDV) remained in the dataset. To reduce the redundancy of this virus group, we randomly selected 10 representative PEDV sequences to retain in our dataset. Consequently, 1392 Coronaviridae sequences were included in our dataset.

("Severe acute respiratory syndrome coronavirus 2"[Organism] OR

In addition to the *Coronaviridae* S protein dataset, we prepared a dataset for the SARS-CoV-2 S protein specifically for domain adaptation. Of the S protein genotypes we defined in this study (see the section "Preparation of genotype–fitness dataset"), we eliminated genotypes with an emergence date later than August 31, 2022, in order to prevent the model from accessing data beyond this cutoff date during the domain adaptation process. Subsequently, we removed redundant sequences using CD-HIT with a clustering threshold of 99%, in accordance with the method described above. Consequently, 114 S protein sequences of SARS-CoV-2 were retained in the dataset. Finally, we combined the *Coronaviridae* and SARS-CoV-2 S protein datasets, resulting in 1506 sequences, for use in domain adaptation step is summarized in Supplemental Data 2.

Introduction of CoVFit

We developed CoVFit, a fitness prediction model based on S protein sequences, by finetuning the ESM-2 protein language model (Fig. 1a). To enhance the model's performance, we employed three key techniques: domain adaptation, multitask learning, and Low-Rank Adaptation (LoRA)⁴⁹. Domain adaptation is an additional pretraining phase using a custom data collection. In this study, we performed domain adaptation using S protein sequences from various human and animal coronaviruses (see the "Dataset preparation for domain adaptation" section). This technique enabled the model to better learn the general properties of these proteins (see the "Domain adaptation" section). Multitask learning, a framework that trains a model on different types of data simultaneously, was utilized to allow the model to capture critical information shared across tasks, thereby enhancing its generalization capabilities. In CoVFit, the model was finetuned using a total of 1,565 regression tasks, including to predict fitness values for 17 countries and to predict relative binding affinity for 1548 mAbs (see the "Model architecture of CoVFit" section). Finally, LoRA is a technique to fine-tune large models efficiently that reduces the requirements for GPU memory resources without compromising the model's prediction performance. LoRA additionally contributes to mitigating the model's tendency to overfit (see the "Model finetuning and performance evaluation" section for further information).

The input for CoVFit consists of amino acid sequences of SARS-CoV-2 S proteins, aligned with the S protein of the Wuhan-Hu-1 strain. These aligned sequences can be generated by Nextclade. CoVFit can predict the fitness (relative R_e) value across 17 countries and the ability to evade 1548 types of mAbs for a given S protein sequence (Fig. S1a).

Training of the CoVFit model completes within 24 h on a computational node with a single Nvidia A100 GPU (40GB) for each instance. Consequently, the model can be updated routinely using the latest genome surveillance data without intensive computational resource requirements.

Utilizing a five-fold cross-validation scheme, we generated five instances of the CoVFit model, which enabled us to estimate both the average prediction value and its uncertainty across these models (Fig. 1b). This approach was chosen because the predicted values, especially regarding the fitness of future variants, can vary among different instances of the trained models (Fig. 3).

CoVFit implementation

ESM-2 models with various parameter sizes are available³² (https:// github.com/facebookresearch/esm). Of these models, we used the version with 650M parameters, prioritizing a balance between prediction performance and computational cost. According to the official benchmark using the unsupervised contact prediction task, this 650M parameter model achieves a performance 1.7 times superior compared to the 35M parameter model, which possesses 20 times fewer parameters. However, the performance gain when comparing the 650M model to the 15 billion (B) parameter model, which has 20 times more parameters, is relatively modest at only 1.08 times. Furthermore, a systematic analysis presented in a recent preprint indicates that enlarging the parameter size of a protein language model does not necessarily enhance prediction performance for tasks outside of protein structure prediction⁵⁰. Given this insight, we opted not to employ models larger than the 650M model, such as the 3B or 15B models, in our study.

The ESM-2 model has a maximum input sequence length (1024 amino acids) due to the computational demands of self-attention, which requires memory in proportion to the square of the sequence length ($O(L^2)$). Unfortunately, the S protein of the Wuhan-Hu-1 strain is composed of 1273 amino acids, exceeding the model's limit. Consequently, amino acid sequences beyond the 1024th position (amino acids 1025–1273; the C-terminus of the S2 subunit) are truncated and not utilized in the ESM-2 model. This constitutes a technical limitation of CoVFit. Nonetheless, this limitation is anticipated to minimally impact performance, considering that while mutations predominantly occur within the S1 subunit (amino acids 1–681), the S2 subunit (amino acids 682–1273) remains highly conserved and with fewer mutations.

CoVFit was implemented using Python v.3.11.4, NVIDIA CUDA v.12.1.0, PyTorch v.2.1.0, Transformers v.4.31.0, and PEFT v.0.5.0. Further information about the system requirements for CoVFit can be found in the GitHub repository (https://github.com/TheSatoLab/CoVFit). The computational codes were executed on a supercomputer node equipped with a single NVIDIA A100 GPU with 40 GB RAM unless otherwise noted.

Domain adaptation

To establish the ESM- $2_{\text{Coronaviridae}}$ model, domain adaptation was carried out using the masked language learning scheme as described in Delvin et al.⁵¹. For domain adaptation, the S protein dataset prepared in the "Dataset preparation for domain adaptation" section was used. For our model's domain adaptation training, each input sequence had 15% of its positions masked randomly, with each instance of a position's masking having an 80% chance to be a <mask> token, a 10% chance to be incorrect, and a 10% chance to be the original. Subsequently, amino acid or token types for these 15% of positions were predicted in batched training steps, and model weights were updated using a cross-entropy loss function.

Using the scheme above, we trained the 650M parameter ESM-2 model with the provided MaskedLM layer, downloaded via functions implemented in the Hugging Face Transformers library. The model was trained for 30 epochs. The batch size was set at 5. A base learning rate of 2e-5 was used with one epoch of warmup, and a cosine-based learning rate scheduler was implemented to successively lower the learning rate during training.

To compare the inference ability of the ESM-2_{Coronaviridae} model to the original ESM-2 model, we performed inference with both models on masked SARS-CoV-2 S protein sequences. Since our dataset for domain adaptation training includes the S proteins of genotypes that emerged up to August 31, 2022, we used genotypes with emergence dates later than September 1, 2022, for inference. The same masking parameters as in the training were used. The results on the test dataset were converted to perplexity scores as the exponential of the crossentropy loss value calculated during inference. Given as *perplexity* = $e^{-\sum_{x} P(x) \log Q(x)}$ where P(x) is the true probability distribution and Q(x) is the probability distribution from the model's predictions, the perplexity score represents how certain the model is in making its predictions, with lower values demonstrating higher certainty. For our inference results, the original ESM-2 model produced a perplexity score of 11.38, whereas the ESM-2_{Coronaviridae} model achieved a low perplexity score of 1.17, demonstrating higher prediction certainty after domain adaptation training (Fig. S1c).

To assess the possibility of the domain adaptation negatively impacting the model's original ability to provide inference on a wide variety of proteins, we again compared the original EMS-2 and ESM- $2_{Coronaviridae}$ models, this time with protein sequences sampled from the UniRef50 released in March 2018. A subset consisting of 29,950 sequences was randomly sampled from the full 3,016,211 sequences and used for the evaluation. The perplexity values of the models were checked as above for the two models, with the original ESM-2 model's perplexity score at 6.76 and the ESM- $2_{Coronaviridae}$ model's perplexity score at 6.86, demonstrating that the model retains its certainty on general proteins after domain adaptation (Fig. S1c).

We conducted the domain adaptation training in Python v3.10.9 with CUDA v12.1.1 and torch v2.1.0.dev20230601 using the Hugging Face Transformers v4.34.1 library.

The computation was executed on a single NVIDIA RTX 6000 Ada GPU with 48 GB RAM. More detailed information on implementation is available in the GitHub repository (https://github.com/TheSatoLab/CoVFit).

Model architecture of CoVFit

For the multitask learning component, we engineered custom taskspecific regression heads for the ESM-2 model (Fig. S1a). On the embedding layer of ESM-2, a linear layer with dimensions equal to the number of tasks was set as task-specific heads. Additionally, an intermediate linear layer with 252 dimensions connecting the embedding layer and the task-head layer was set.

In CoVFit, a single input sequence is linked to multiple target variables due to the multitask learning framework. For example, regarding DMS data, a typical S protein mutant (input sequence) is linked to relative binding affinity values for >1000 mAbs (target variables). To boost computational efficiency, CoVFit utilizes an architecture that processes a single input sequence alongside its multiple corresponding target variables in parallel, rather than processing pairs of the same input sequence and one target variable sequentially (Fig. S1b). As a result, the loss values for multiple target variables associated with a single input sequence are calculated simultaneously.

However, the number of tasks linked to each input sequence can differ greatly, especially when comparing the variant S protein sequences used for fitness prediction (up to 17 tasks) against the mutant sequences used for DMS predictions (up to 1548 tasks).

Consequently, the magnitude of the loss value for each dataset can vary significantly based on the number of associated tasks, which can lead to training instability. To stabilize the training process, CoVFit utilized non-overlapping random sampling to create data chunks where a single input sequence is associated with target variables for a maximum of 10 tasks. These generated sequence-variable chunks were then used as the training inputs.

For the loss function, CoVFit utilizes a custom least squares approach weighted according to individual tasks. In principle, the weights were determined to be proportional to the reciprocals of the task frequencies. One exception was implemented where, for fitness prediction tasks for genotypes that emerged after January 1, 2022, we adjusted the weights by doubling them.

Model finetuning and performance evaluation

In CoVFit, we finetuned the custom model based on ESM-2 using the LoRA technique implemented in the Hugging Face PEFT v.0.5.0. Low-rank adapters were injected into the weight matrices of the key, query, and value components, as well as those for the dense layers. Full finetuning was applied to these adapters and the custom regression heads added onto ESM-2, while the other, original layers were kept frozen in their pretrained state. A rank parameter of r = 8 and a scale parameter of alpha = 16 were used. Consequently, out of the total 659,741,475 parameters, the model has 7,768,974 trainable parameters, which constitutes -1.18% of the total. The LoRA dropout rate was set at 0.05.

For finetuning, the AdamW optimizer was used with a weight decay parameter of 0.02. The maximum learning rate was set at 2.0E-4 with a linear learning rate scheduler, and the training was conducted over 20 epochs with a warmup ratio of 0.05. The batch size was set at 4 with gradient accumulation steps of 2.

The genotype-fitness and DMS datasets were randomly divided into training, evaluation, and test datasets in a 6:2:2 ratio. For the genotype-fitness dataset, we considered the combinations of country and Nextclade clade, ensuring that data representing each combination were evenly distributed across the training-evaluation and test datasets. Similarly, for the DMS dataset, the types of mAbs were considered during the data splitting process. We conducted the data splitting with a five-fold cross-validation approach.

In our experiments aimed at assessing the model's ability to predict the performance of future variants, we began by dividing the genotype-fitness dataset into two: one for past variants and another for future variants, based on their emergence dates. We generated eight datasets using eight cutoff dates, spaced 1 month apart, from January 31, 2022, to August 31, 2022 (Fig. 3b). Training was done using these datasets according to the scheme described above.

In our experiments designed to assess the importance of including DMS data for immune evasion, we trained alternative instances without incorporating the DMS dataset for comparison. Likewise, in our experiments evaluating the significance of the domain adaptation step, we employed the original ESM-2 model rather than the version adapted to the coronaviral S protein dataset.

CoVFit-CLI

The CoVFit-CLI tool packages $CoVFit_{Nov23}$ via pyinstaller 6.4.0 using Python v.3.10.9, torch v.2.1.2, transformers 4.37.1, and bio v.1.5.9 with CUDA v.12.3 on x86_64 Linux, kernel 5.15.0.

Development of fitness prediction models based on non-deep learning models

We constructed fitness prediction models based on LASSO, Random Forest, and LightGBM to compare the prediction performance of CoVFit with those of these models. LASSO employs a linear regression framework enhanced with L1 regularization, offering a method to include penalty terms that reduce overfitting by shrinking some coefficients to zero. In contrast, Random Forest and LightGBM are advanced, decision tree-based models known for their greater expressive capability. These models aim to predict a variant's fitness value based on its amino acid mutation profile in the S protein and the country of origin. Both the mutation profile and the country data were one-hot encoded to serve as input features for the models.

We trained these models and evaluated their performance using the past–future variant datasets with eight cutoff dates, spaced 1 month apart, from January 31, 2022, to August 31, 2022, and the fivefold cross-validation scheme as described in the "Model finetuning and performance evaluation" section. In the training dataset, we selected 200 features to be used as inputs of the models according to the feature importance estimated by Random Forest. We trained the models with hyperparameter-tuning using a Bayesian optimization method. In this process, R^2 was used as the optimization metric, and the number of iterations was set at 20. The parameter spaces searched in this step are described in detail in the GitHub repository (https:// github.com/TheSatoLab/CoVFit).

The machine learning models above were reconstructed using Python v.3.9.13, pandas v.1.4.4, numpy v.1.21.5, lightgbm v.3.3.5, scikit-learn v.1.0.2, and scikit-optimize v.0.9.0.

Phylogenetic analysis

We created the dataset for phylogenetic analysis as a subset of the dataset of the representative viral genome sequences encoding respective S protein genotypes (EPI_SET_ID: EPI_SET_240311ma; https://www.gisaid.org; see "Preparation of genotype–fitness dataset" section). We removed sequences matching the following criteria: (i) sequences with >3% ambiguous characters across positions 265 to 29,673 (in alignment with the Wuhan-Hu-1 reference (GenBank accession number: NC_045512.2 [https://www.ncbi.nlm.nih.gov/nuccore/1798174254])) and (ii) sequences classified as "recombinant" according to Nextclade clade assignments. Additionally, we included the Wuhan-Hu-1 reference genome sequence to our dataset. The dataset for viral genome sequences used in the phylogenetic analysis, except for the Wuhan-Hu-1 reference genome, is summarized under the EPI_SET_ID: EPI_SET_240311rk, which can be accessed through the GISAID website (https://gisaid.org/).

The nucleotide viral genome sequences were aligned to the reference sequence of Wuhan-Hu-1 using Minimap2 v.2.1752. This alignment was then converted into a multiple sequence alignment following the GISAID phylogenetic analysis pipeline (https://github. com/roblanf/sarscov2phylo). Sites corresponding to positions 1-265 and 29,674-29,903 in the reference genome were masked, that is, converted to "NNN," to exclude them from subsequent analyses. The maximum likelihood phylogenetic tree was constructed using IQ-TREE v.2.1.4_beta, adopting the GTR+I+G nucleotide substitution model⁵³. To assess the reliability of the phylogenetic tree nodes, an ultrafast bootstrap analysis was performed with 1000 replicates. A timeresolved phylogenetic tree was inferred from the constructed tree using TreeTime v.0.11.1, with the rerooting strategy set to "oldest"⁵⁴, resulting in rerooting by the Wuhan-Hu-1 strain. The S protein sequences for ancestral nodes were also reconstructed using Time-Tree with the default options.

Detection of phylogenetic branches with fitness elevation using CoVFit

To infer the impact of mutations on fitness through the observed evolution of SARS-CoV-2, we analyzed the increase in predicted fitness across all branches of the SARS-CoV-2 phylogenetic tree, as outlined in the "Phylogenetic analysis" section. We employed five CoVFit_{Nov23} models, developed via a five-fold cross-validation, to predict the fitness values for both existing and reconstructed ancestral S protein sequences within the tree. Since CoVFit predicts fitness across multiple countries, we averaged these predictions to obtain a single

representative fitness value for each sequence, resulting in five representative fitness values per sequence. We compared these values between each node and its ancestral node, calculating the mean fitness gain for the branches connecting them. Statistical significance of fitness changes was determined using a paired Welch's *t*-test, with multiple testing correction applied via the Benjamini–Hochberg method. Branches with an FDR less than 0.1 were considered statistically significant. We also identified mutations in the S protein acquired along each branch by comparing the S protein sequences at both ends of the branch. The detected fitness elevation events are summarized in Supplemental Data 3.

Characterization of the F456L substitution using publicly available DMS data

Position-wise scores for escape from humoral immunity were calculated using escape estimator³⁷ based on DMS data for the ability to evade mAbs presented in Cao et al.²⁴ (https://github.com/jbloomlab/ SARS2_RBD_Ab_escape_maps) (shown in Fig. 5c). For the ACE2 binding and protein expression DMS data, we retrieved the per-site variant score results presented by Taylor and Starr³⁶ (https://github.com/ tstarrlab/SARS-CoV-2-RBD_DMS_Omicron-XBB-BQ/blob/main/results/ final_variant_scores/final_variant_scores.csv). We filtered for mutant L on position 456 and retrieved the "delta_bind" and "delta_expr" values presenting the mean of values across replicates minus the mean for the reference residue for each variant target (shown in Fig. 5d).

CoVFit-based in silico (deep) mutational scanning analysis

Instances of the CoVFit_{nonJN1} model, models trained on the dataset without the sequences of JN.1, were utilized to infer the fitness of S protein mutants. First, the mean fitness value for each mutant was calculated across different countries. Subsequently, these mean values were averaged across all five CoVFit_{nonJN1} model instances, yielding a singular average fitness value for each S protein mutant. This streamlined fitness value was then compared to the fitness of the original backbone S protein sequence.

Epidemic analysis on JN.1 subvariants

We retrieved all SARS-CoV-2 genome sequences and their associated metadata available up to October 21, 2024, from GISAID. To ensure data quality, sequences were excluded from analysis based on the following criteria: (i) absence of collection date; (ii) samples taken from animals other than humans; (iii) more than 2% undetermined nucleotides; or (iv) samples collected during quarantine. We analyzed the BA.2.86 lineage collected between October 1, 2023, and July 31, 2024 (EPI SET ID: EPI_SET_241126wq).

We calculated the proportion of viral sequences harboring mutations at specific sites within the BA.2.86 lineage. Each viral lineage category includes its descendant lineages unless those descendant lineages are explicitly defined as separate categories. For the temporal trends in mutation and variant detection frequencies, calculations were performed at 7-day intervals for each geographic region within the BA.2.86 population. Results from Africa and South America were excluded due to the low total sequence count.

Methodological discussion: room for improvement of CoVFit

We recognize the presence of multiple areas where CoVFit could potentially be improved with future development. First, since our current model is solely trained on S protein sequences, it may be possible to improve its performance by including information on additional viral proteins. Previous studies have identified mutations associated with increased fitness also in non-S proteins, particularly in the nucleocapsid (N) protein, supporting the possible effectiveness of this approach⁴. In the current setting, the effects of mutations in non-S proteins are absorbed into the effects of mutations in the S protein that have linkage disequilibrium relationships with these mutations. However, it is certain that the S protein has a particularly strong impact on fitness compared to other viral proteins⁴. Therefore, it is unclear to what extent prediction accuracy would be improved by adding information from other viral proteins. There is even a possibility that generalizability could be decreased by including other viral proteins due to the decrease in signal-to-noise ratio. Similarly, although the amino acid sequences of the S protein beyond the 1024th position (amino acids 1025–1273; the C-terminus of the S2 subunit) are truncated and not utilized in the ESM-2 model, it is unclear whether this limitation has a negative effect on prediction performance, as the S2 subunit (amino acids 682–1273) remains highly conserved and has fewer mutations.

Second, it may also be possible to improve the performance by including various DMS data, such as those on other viral phenotypes. In this study, we only used DMS data on the immune evasion ability against mAbs. However, given the significant impact of the ACE2binding affinity of the S protein on fitness, employing DMS data for this trait could improve predictive performance. In our preliminary experiments, however, we found that the convergence speed for DMS data on binding affinity to ACE2 was much slower compared to other tasks. Considering the difficulty of simultaneous learning, we decided not to use this DMS data in this study. Similarly, while we used DMS data obtained by experiments using S proteins of other variants (particularly variants emerged recently) could potentially further improve predictive performance.

The third consideration is the scaling up of the model. It is generally known that language models improve in performance as they increase in size. Indeed, benchmark tests using ESM-2 have shown that changing the model size from the 650M, used in this study, to 3B or 15B can lead to slight improvements in performance regarding the prediction of protein structures³². For our method, using larger models like 3B or 15B models may also enhance performance. Recently developed techniques, such as QLoRA, a quantization-based LoRA⁵⁵, make it possible to fine-tune even 15B models in a limited GPU resource. However, we faced issues with incompatibility between CoVFit and QLoRA, leading us to abandon the development of a QLoRA-based model. It is also important to note that scaling up the model can significantly increase training and inference times.

The fourth consideration is data augmentation for fitness data. In viral genome surveillance, there can be a delay of several weeks to months between the date of sample collection and the date of data submission to databases. Furthermore, in viral genome surveillance, the number of viral genomes newly sequenced have gradually decreased in recent years (https://gisaid.org/). Consequently, the most recent genomic data tends to be under-sampled. However, this recent data is considered to contain more critical information for predicting future variants compared to older data. Therefore, increasing the proportion of recent data in the training dataset or employing data augmentation techniques, which artificially expand the dataset, might enhance the model's ability to generalize to future variants.

Lastly, since we have not conducted an exhaustive investigation of the model's hyperparameters, the model's performance could be improved by adjusting them. Adjustable hyperparameters include data normalization methods, network architecture, task weight balance, the optimizer algorithm, learning rate, and maximum epoch numbers.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Surveillance datasets of SARS-CoV-2 genomes are available from the GISAID database (https://www.gisaid.org; EPI_SET_240307pq; EPI_SET_240311ma; EPI_SET_240311rk; EPI_SET_241126wq). The supplemental table for each GISAID dataset is available in the GitHub

Article

repository (https://github.com/TheSatoLab/CoVFit). Source data are provided with this paper.

Code availability

The computational codes used in this study are available in the GitHub repository (https://github.com/TheSatoLab/CoVFit).

References

- 1. Pybus, O. G. & Rambaut, A. Evolutionary analysis of the dynamics of viral infectious disease. *Nat. Rev. Genet.* **10**, 540–550 (2009).
- 2. Carabelli, A. M. et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat. Rev. Microbiol.* **21**, 162–177 (2023).
- Markov, P. V. et al. The evolution of SARS-CoV-2. Nat. Rev. Microbiol. 21, 361–379 (2023).
- Obermeyer, F. et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).
- 5. Zhou, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
- 6. Chen, Y. et al. Broadly neutralizing antibodies to SARS-CoV-2 and other human coronaviruses. *Nat. Rev. Immunol.* **23**, 189–199 (2023).
- 7. Suzuki, R. et al. Attenuated fusogenicity and pathogenicity of SARS-CoV-2 Omicron variant. *Nature* **603**, 700–705 (2022).
- 8. Yamasoba, D. et al. Virological characteristics of the SARS-CoV-2 Omicron BA.2 spike. *Cell* **185**, 2103–2115.e2119 (2022).
- Kimura, I. et al. Virological characteristics of the SARS-CoV-2 Omicron BA.2 subvariants, including BA.4 and BA.5. *Cell* 185, 3992–4007.e3916 (2022).
- 10. Saito, A. et al. Virological characteristics of the SARS-CoV-2 Omicron BA.2.75 variant. *Cell Host Microbe* **30**, 1540–1555.e1515 (2022).
- Ito, J. et al. Convergent evolution of SARS-CoV-2 Omicron subvariants leading to the emergence of BQ.1.1 variant. *Nat. Commun.* 14, 2671 (2023).
- Tamura, T. et al. Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants. *Nat. Commun.* 14, 2800 (2023).
- Uriu, K. et al. Enhanced transmissibility, infectivity, and immune resistance of the SARS-CoV-2 omicron XBB.1.5 variant. *Lancet Infect. Dis.* 23, 280–281 (2023).
- 14. Yamasoba, D. et al. Virological characteristics of the SARS-CoV-2 omicron XBB.1.16 variant. *Lancet Infect. Dis.* **23**, 655–656 (2023).
- Kaku, Y. et al. Antiviral efficacy of the SARS-CoV-2 XBB breakthrough infection sera against omicron subvariants including EG.5. *Lancet Infect. Dis.* 23, e395–e396 (2023).
- Uriu, K. et al. Transmissibility, infectivity, and immune evasion of the SARS-CoV-2 BA.2.86 variant. *Lancet Infect. Dis.* 23, e460–e461 (2023).
- 17. Kosugi, Y. et al. Characteristics of the SARS-CoV-2 omicron HK.3 variant harbouring the FLip substitution. *Lancet Microbe* **5**, e313 (2024).
- 18. Tamura, T. et al. Virological characteristics of the SARS-CoV-2 BA.2.86 variant. *Cell Host Microbe* **32**, 170–180.e112 (2024).
- 19. Kaku, Y. et al. Virological characteristics of the SARS-CoV-2 JN.1 variant. *Lancet Infect. Dis.* **24**, e82 (2024).
- 20. Davies, N. G. et al. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. *Science* **372**, eabg3055 (2021).
- 21. Vöhringer, H. S. et al. Genomic reconstruction of the SARS-CoV-2 epidemic in England. *Nature* **600**, 506–511 (2021).
- Ito, K., Piantham, C. & Nishiura, H. Estimating relative generation times and reproduction numbers of Omicron BA.1 and BA.2 with respect to Delta variant in Denmark. *Math. Biosci. Eng.* 19, 9005–9017 (2022).
- Yajima, H. et al. Molecular and structural insights into SARS-CoV-2 evolution: from BA.2 to XBB subvariants. *mBio* 15, e0322023 (2024).

- 24. Cao, Y. et al. Imprinted SARS-CoV-2 humoral immunity induces convergent Omicron RBD evolution. *Nature* **614**, 521–529 (2023).
- 25. Planas, D. et al. Distinct evolution of SARS-CoV-2 Omicron XBB and BA.2.86/JN.1 lineages combining increased fitness and antibody evasion. *Nat. Commun.* **15**, 2254 (2024).
- 26. Khan, K. et al. Evolution and neutralization escape of the SARS-CoV-2 BA.2.86 subvariant. *Nat. Commun.* **14**, 8078 (2023).
- 27. Wang, Q. et al. Alarming antibody evasion properties of rising SARS-CoV-2 BQ and XBB subvariants. *Cell* **186**, 279–286.e278 (2023).
- Meng, B. et al. Altered TMPRSS2 usage by SARS-CoV-2 Omicron impacts infectivity and fusogenicity. *Nature* 603, 706–714 (2022).
- 29. Tsujino, S. et al. Virological characteristics of the SARS-CoV-2 Omicron EG.5.1 variant. *Microbiol Immunol.* **68**, 305–330 (2024).
- Masuda, Y. et al. Characterization of the evolutionary and virological aspects of mutations in the receptor binding motif of the SARS-CoV-2 spike protein. *Front. Virol.* 3, 1328229 (2023).
- 31. Bepler, T. & Berger, B. Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654–669.e653 (2021).
- 32. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
- Starr, T. N. et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *Cell* 182, 1295–1310.e1220 (2020).
- 34. Yajima, H. et al. Structural basis for receptor-binding domain mobility of the spike in SARS-CoV-2 BA.2.86 and JN.1. *Nat. Commun.* **15**, 8574 (2024).
- Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. Nat. Rev. Mol. Cell Biol. 23, 3–20 (2022).
- Taylor, A. L. & Starr, T. N. Deep mutational scans of XBB.1.5 and BQ.1.1 reveal ongoing epistatic drift during SARS-CoV-2 evolution. *PLoS Pathog.* **19**, e1011901 (2023).
- Greaney, A. J., Starr, T. N. & Bloom, J. D. An antibody-escape estimator for mutations to the SARS-CoV-2 receptor-binding domain. Virus Evol. 8, veac021 (2022).
- 38. Dadonaite, B. et al. Spike deep mutational scanning helps predict success of SARS-CoV-2 clades. *Nature* **631**, 617–626 (2024).
- 39. Chen, L. et al. Virological characteristics of the SARS-CoV-2 LP.8.1 variant. *Lancet Infect. Dis.* **25**, e193 (2025).
- 40. Han, W. et al. Predicting the antigenic evolution of SARS-COV-2 with deep learning. *Nat. Commun.* **14**, 3478 (2023).
- 41. Thadani, N. N. et al. Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818–825 (2023).
- 42. Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).
- 43. Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Syst.* **13**, 274–285.e276 (2022).
- Focosi, D., Quiroga, R., McConnell, S., Johnson, M. C. & Casadevall A. Convergent evolution in SARS-CoV-2 spike creates a variant soup from which new COVID-19 waves emerge. *Int. J. Mol. Sci.* 24, 2264 (2023).
- 45. Bloom, J. D. & Neher, R. A. Fitness effects of mutations to SARS-CoV-2 proteins. *Virus Evol.* **9**, vead055 (2023).
- 46. Hadfield, J. et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**, 4121–4123 (2018).
- 47. Madewell, Z. J. et al. Rapid review and meta-analysis of serial intervals for SARS-CoV-2 Delta and Omicron variants. *BMC Infect. Dis.* **23**, 429 (2023).
- Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150–3152 (2012).
- Hu, E. J. et al. LoRA: low-rank adaptation of large language models. Preprint at https://ui.adsabs.harvard.edu/abs/ 2021arXiv210609685H (2021).

Article

- Li, F.-Z., Amini, A. P., Yue, Y., Yang, K. K. & Lu A. X. Feature reuse and scaling: understanding transfer learning with protein language models. *bioRxiv*, 2024.2002.2005.578959 (2024).
- Devlin, J., Chang, M.-W., Lee, K. & Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://ui.adsabs.harvard.edu/abs/2018arXiv181004805D (2018).
- 52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
- Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37, 1530–1534 (2020).
- 54. Sagulenko, P., Puller, V. & Neher, R. A. TreeTime: maximumlikelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).
- Dettmers, T., Pagnoni, A., Holtzman, A. & Zettlemoyer, L. QLoRA: efficient finetuning of quantized LLMs. Preprint at https://ui.adsabs. harvard.edu/abs/2023arXiv230514314D (2023).
- Benton, D. J. et al. The effect of the D614G substitution on the structure of the spike glycoprotein of SARS-CoV-2. Proc. Natl. Acad. Sci. USA 118, e2022586118 (2021).
- 57. Meng, E. C. et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).

Acknowledgements

We gratefully acknowledge all data contributors, i.e., the authors and their originating laboratories responsible for obtaining the specimens, and their submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based. We would like to thank all members belonging to The Genotype to Phenotype Japan (G2P-Japan) Consortium. The super-computing resources were provided by the Human Genome Center and the Information Technology Center, the University of Tokyo. This study was supported in part by JST PRESTO (JPMJPR22R1, to J.I.); JSPS KAKENHI Grant-in-Aid for Scientific Research B (25K00116, to J.I.); JSPS KAKENHI Grant-in-Aid for Early-Career Scientists (23K14526, to J.I.); AMED SCARDA Japan Initiative for World-leading Vaccine Research and Development Center "UTOPIA" (JP223fa627001, JP243fa627001h0003, and 253fa627001, to J.I.; JP223fa627001, JP243fa627001h0003, and 253fa627001 to K.S.); AMED SCARDA Program on R&D of new generation vaccine including new modality application (253fa727002 to K.S.); SHIONOGI Infectious Disease Research Promotion Foundation (to J.I.); AMED ASPIRE Program (25jf0126002, to G2P-Japan Consortium and K.S.); AMED Research Program on Emerging and Re-emerging Infectious Diseases (23fk0108583, 24fk0108690, 24wm0225041, to K.S.); JSPS KAKENHI Fund for the Promotion of Joint International Research (International Leading Research) (25wm022504 and 25fk0108690, to K.S.; 25wm022504 to J.I.); JSPS KAKENHI Grant-in-Aid for Scientific Research A (JP24H00607, to K.S.); Mitsubishi UFJ Financial Group, Inc. Vaccine Development Grant (to J.I. and to K.S.); and The Cooperative Research Program (Joint Usage/ Research Center program) of Institute for Life and Medical Sciences, Kyoto University (to K.S.); JSPS International Research Fellow (to G.J.).

Author contributions

Jumpei Ito designed the study, the main conceptual ideas, and the proof outline. Adam Strange developed ESM- $2_{\rm Coronaviridae}$ and the CovFit-CLI

standalone command-line tool. Wei Liu constructed a series of non-deep learning prediction methods. Gustav Joas collected and created the *Coronaviridae* dataset. Spyros Lytras performed a detailed analysis on the F456L substitution. Jumpei Ito performed the other parts of CoVFit development and computational analyses. Jumpei Ito made the figures and wrote the initial draft of the manuscript. Adam Strange and Spyros Lytras provided editing. Kei Sato and The Genotype to Phenotype Japan (G2P-Japan) Consortium contributed to the project administration. All authors reviewed and proofread the manuscript.

Competing interests

J.I. has consulting fees and honoraria for lectures from Takeda Pharmaceutical Co. Ltd Spyros Lytras has consulting fees from EcoHealth Alliance. K.S. has consulting fees from Moderna Japan Co., Ltd and Takeda Pharmaceutical Co. Ltd, and honoraria for lectures from Gilead Sciences, Inc., Moderna Japan Co., Ltd, and Shionogi & Co., Ltd. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59422-w.

Correspondence and requests for materials should be addressed to Jumpei Ito or Kei Sato.

Peer review information *Nature Communications* thanks Amalio Telenti, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http:// creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025

The Genotype to Phenotype Japan (G2P-Japan) Consortium

Keita Matsuno¹⁰, Naganori Nao¹⁰, Hirofumi Sawa¹⁰, Keita Mizuma¹⁰, Isshu Kojima¹⁰, Jingshu Li¹⁰, Tomoya Tsubo¹⁰, Shinya Tanaka¹⁰, Masumi Tsuda¹⁰, Lei Wang¹⁰, Yoshikata Oda¹⁰, Zannatul Ferdous¹⁰, Kenji Shishido¹⁰, Takasuke Fukuhara¹⁰, Tomokazu Tamura¹⁰, Rigel Suzuki¹⁰, Saori Suzuki¹⁰, Shuhei Tsujino¹⁰, Hayato Ito¹⁰, Yu Kaku¹, Naoko Misawa¹, Arnon Plianchaisuk¹, Ziyi Guo¹, Alfredo A. Hinay Jr.¹, Kaoru Usui¹, Wilaiporn Saikruang¹, Keiya Uriu¹, Yusuke Kosugi¹, Shigeru Fujita¹, Jarel Elgin M.Tolentino¹, Luo Chen¹, Lin Pan¹, Wenye Li¹, Mai Suganami¹, Mika Chiba¹, Ryo Yoshimura¹, Kyoko Yasuda¹, Keiko Iida¹, Naomi Ohsumi¹, Shiho Tanaka¹, Kaho Okumura¹, Kazuhisa Yoshimura¹¹, Kenji Sadamas¹¹, Mami Nagashima¹¹, Hiroyuki Asakura¹¹, Isao Yoshida¹¹, So Nakagawa¹², Akifumi Takaori-Kondo¹³, Kotaro Shirakawa¹³, Kayoko Nagata¹³, Ryosuke Nomura¹³, Yoshihito Horisawa¹³, Yusuke Tashiro¹³, Yugo Kawai¹³, Kazuo Takayama¹³, Rina Hashimoto¹³, Sayaka Deguchi¹³, Yukio Watanabe¹³, Yoshitaka Nakata¹³, Hiroki Futatsusako¹⁴, Ayaka Sakamoto¹³, Naoko Yasuhara¹³, Takao Hashiguchi¹³, Tateki Suzuki¹³, Kanako Kimura¹³, Jiei Sasaki¹³, Yukari Nakajima¹³, Hisano Yajima¹³, Takashi Irie¹⁴, Ryoko Kawabata¹⁴, Kaori Sasaki-Tabata¹⁵, Terumasa Ikeda¹⁶, Hesham Nasse¹⁶, Ryo Shimizu¹⁶, MST Monira Begum¹⁶, Michael Jonathan¹⁶, Yuka Mugita¹⁶, Sharee Leong¹⁶, Otowa Takahashi¹⁶, Kimiko Ichihara¹⁶, Takamasa Ueno¹⁶, Chihiro Motozono¹⁶, Mako Toyoda¹⁶, Akatsuki Saito¹⁷, Maya Shofa¹⁷, Yuki Shibatani¹⁷, Tomoko Nishiuchi¹⁷, Jiri Zahradni¹⁸, Prokopios Andrikopoulos¹⁸, Miguel Padilla-Blanco¹⁸ & Aditi Konar¹⁸

¹⁰Hokkaido University, Sapporo, Japan. ¹¹Tokyo Metropolitan Institute of Public Health, Tokyo, Japan. ¹²Tokai University, Tokyo, Japan. ¹³Kyoto University, Kyoto, Japan. ¹⁴Hiroshima University, Hiroshima, Japan. ¹⁵Kyushu University, Fukuoka, Japan. ¹⁶Kumamoto University, Kumamoto, Japan. ¹⁷University of Miyazaki, Miyazaki, Japan. ¹⁸Charles University, Prague, Czech Republic.