Deep-learning-enabled generative prevalent mutation prediction through host-to-herd *in silico* virus evolution

Xudong Liu[1,2†], Zhiwei Nie[1,2†], Haorui Si[3,4,5†], Xurui Shen[5,3*], Yutian Liu[6,2], Xiansong Huang[2], Tianyi Dong[7,8,5], Fan Xu[2], Jie Chen[1,2*], Peng Zhou[5,3*], Yonghong Tian[1,6,2*]

[1]School of Electronic and Computer Engineering, Peking University, China.
[2]Peng Cheng Laboratory, Shenzhen, China.
[3]Guangzhou Medical University, Guangzhou, China.
[4]State Key Laboratory of Respiratory Disease, Guangzhou, China.
[5]Guangzhou Laboratory, Guangzhou, China.
[6]School of Computer Science, Peking University, China.
[7]CAS Key Laboratory of Special Pathogens, Wuhan Institute of Virology, Chinese Academy of Sciences, Wuhan, China.
[8]University of Chinese Academy of Sciences, Beijing, China.

*Corresponding author(s). E-mail(s): shen_xurui@gzlab.ac.cn; jiechen2019@pku.edu.cn; zhou_peng@gzlab.ac.cn; yhtian@pku.edu.cn;
†These authors contributed equally to this work.

Abstract

Predicting the mutation prevalence trends of emerging viruses in the real world is an efficient means to update vaccines or drugs in advance. It is crucial to develop a computational method spanning from the host level to the herd level for prevalent mutation prediction as virus evolves within and between hosts involving the impact of multiple selective pressures. Here, a deep-learning generative prediction framework for real-world prevalent mutations, GenPreMut, is developed with a novel host-to-herd selective pressure simulation strategy. Through the paradigm of host-to-herd *in silico* virus evolution, GenPreMut reproduces previous real-world prevalent mutations for multiple lineages with significant accuracy improvements over state-of-the-art methods. More importantly, GenPreMut correctly predicts future prevalent mutations that dominate the pandemic in the

real world more than half a year in advance with *in vitro* experimental validation. Overall, GenPreMut demonstrates a proactive approach to the prevention of emerging viral infections, accelerating the process of discovering future prevalent mutations with the power of generative prediction.

1 Introduction

The representative emerging viral infection COVID-19, caused by Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1, 2], has had a catastrophic impact on public health worldwide, including extremely high morbidity and millions of deaths [1]. Regardless of geographic location, SARS-CoV-2 evolves within and between hosts [3], involving the action of multiple selective pressures [4–7]. The evolutionary trend of SARS-CoV-2 is profoundly influenced by those selective pressures [8], which induces its adaptive evolution to produce various variants, most notably multiple sub-lineages of Omicron [9].

Predicting the mutation prevalence trends of SARS-CoV-2 or other emerging viruses that may appear is a powerful means to update vaccines or drugs in advance. Previous studies [8, 10–18] analyzed the mutation-induced fitness changes of infectious viruses through deep mutation scanning (DMS) profiles [19]. However, with the continuous emergence of new variants, the significant experimental cost prevents such experimental method from becoming a real-time ex ante prediction approach, making it more suitable for post-hoc analysis. Therefore, it is crucial to develop efficient computational methods to predict potential prevalent mutations in the future.

Previous computational methods, especially deep-learning models, for virus prediction have focused on predicting the properties of viral proteins, reproducing evolutionary trajectories, and predicting evolutionary trends at the host level. For the property prediction of viral proteins, the spike protein, the receptor binding domain (RBD), the angiotensin converting enzyme 2 (ACE2), the antibody are jointly or individually input into a neural network to predict fitness changes caused by mutations [20–28]. For the reproduction of evolutionary trajectories, evo-velocity [29] and PyR_0 [30] were proposed to predict the directionality of virus evolution based on detected viral sequences. For the evolutionary trend prediction, some previous studies [22, 31–33] achieved the prediction of evolutionary trends through host-level fitness prediction.

However, the above computational methods are difficult to predict mutations that are likely to become prevalent in the future. The property prediction models of viral proteins [20–28] are customized according to specific properties, thus lacking a comprehensive consideration of selective pressures. The methods of reproducing evolutionary trajectories [29, 30] can only perform post hoc panoramic analyses of the observed variants, and cannot predict the prevalence of unknown lineages. As for the existing evolutionary trend prediction models [22, 31–33], the lack of simultaneous consideration of intra-host and inter-host selective pressures makes it difficult for their

[1] https://data.who.int/dashboards/covid19

predictions to be strongly correlated with the prevalence in the real world. Therefore, it is crucial to develop a computational method for predicting prevalent mutations that spans from the host level to the herd level.

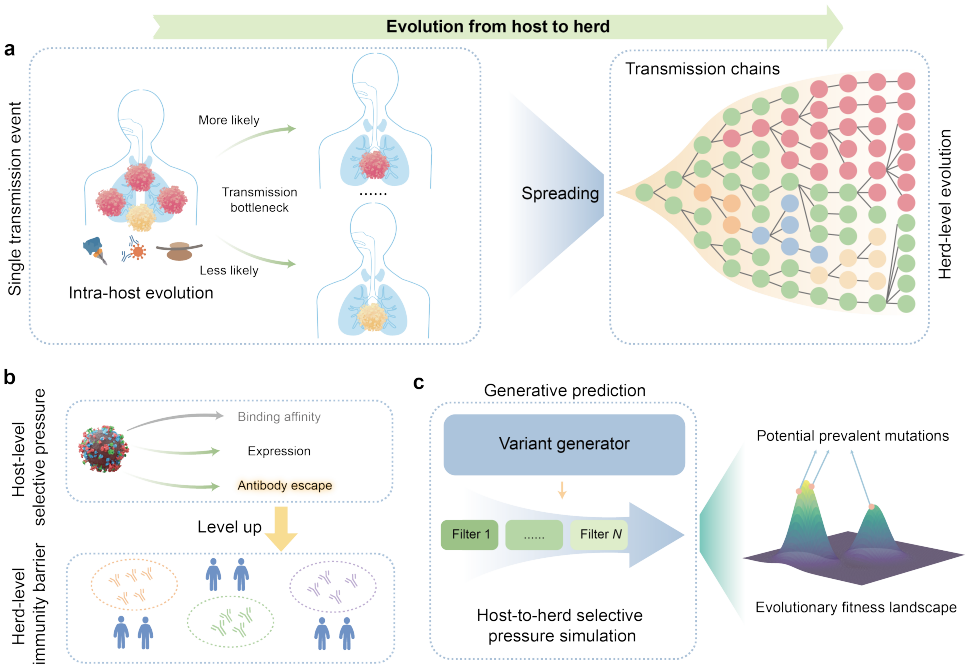

Fig. 1 Our motivation and methodology. **a**, Illustration of the virus evolution from the host level to the herd level. After undergoing intra-host evolution, the variants spreads through the transmission bottleneck in a single transmission event (left panel), thus forming a large number of transmission chains, in which lineages with selective advantages become dominant (right panel). **b**, Our host-to-herd selective pressure simulation strategy. Intra-host adaptive evolution involves main host-level selective pressures including ACE2 binding affinity, expression, and antibody escape, among which host-level antibody escape is upgraded to herd-level immunity barrier to introduce herd-level selective pressures. **c**, The methodology of our deep-learning generative prediction framework for potential prevalent mutations. Massive variants generated by the variant generator are subjected to host-to-herd selective pressure simulation, thereby recommending potential real-world prevalent mutations in the future in the vast evolutionary fitness landscape.

As shown in Fig.1a, as the link between intra-host evolution and inter-host evolution, the transmission bottleneck is often very narrow, in which variants with significant selective advantages are more likely to infect new hosts in a single transmission event [3, 34]. With continuous spreading, variants with significant selective advantages account for most of the transmission chains, thus showing the dominance of the pandemic at the herd level [3]. In order to make the computational methods take into account the selective pressures of intra-host evolution and inter-host evolution at the same time, it is a feasible strategy to introduce herd-level mutational drivers based on the host-level mutational drivers. As shown in Fig.1b, for the intra-host

adaptive evolution of SARS-CoV-2, ACE2 binding affinity, expression, and antibody escape are the main mutational drivers, i.e., host-level selective pressures [11, 35, 36]. First, according to a large number of previous studies [37–46], ACE2 binding affinity can be partially sacrificed during the intrapandemic evolution of SARS-CoV-2. Therefore, we do not consider the ACE2 binding affinity of variants at the host level in this work because it cannot be an effective criterion for judging the transmission advantage of variants. Second, considering the importance of breaking through antibody immunity at the herd level for the occurrence of transmission events, we upgrade host-level antibody escape with estimated herd-level immunity barrier [47, 48].

With the above host-to-herd selective pressure simulation strategy, a deep-learning **gen**erative prediction framework for real-world **pre**valent **mut**ations, GenPreMut, is developed. GenPreMut realizes generative prediction through host-to-herd *in silico* virus evolution, including variant generation and selective pressure screening. The protein language model fine-tuned by real-time evolutionary information is employed as the variant generator for mutated-site-guided variant generation, and the predictive models of intra-host and inter-host selective pressures are adopted as diverse filters for variant screening (Fig.1c). Finally, potential prevalent mutations are recommended by GenPreMut in the vast evolutionary fitness landscape, not only reproducing previous real-world prevalent mutations with *in silico* virus evolution, but also correctly predicting future prevalent mutations that dominate the pandemic in the real world more than half a year in advance with *in vitro* experimental validation. Overall, GenPreMut facilitates the predictive discovery of future prevalent mutations and provides insights for proactive prevention of emerging viral infections.

2 Results

2.1 A generative prediction framework across host to herd

GenPreMut is a generative prediction framework across host to herd to predict potential prevalent mutations in the future. Four modules make up this deep-learning framework, namely protein language model (PLM) fine-tuning, mutated-site-guided variant generation, host-to-herd selective pressure screening, and mutation ranking. Following this host-to-herd *in silico* virus evolution process, the top-ranked point mutations are recommended as the most likely prevalent mutations in the real world.

First, for the PLM fine-tuning module, the real-time evolutionary information is adopted to fine-tune the pretrained PLM to internalize evolutionary patterns of viral proteins (Fig.2a). Specifically, considering the ability of PLM to internalize evolutionary patterns [49–52], we choose the pretrained PLM (ESM-2 with 650M parameters here) as the basic variant generator. However, the off-the-shelf PLM is pretrained on massive natural protein sequences and lacks understanding of the mutational patterns of SARS-CoV-2 viral proteins. Therefore, we integrate the variant sequences of specific lineages to fine-tune the pretrained PLM to adjust the learned evolutionary patterns towards the mutational fitness of viral proteins (Fig.2a). In detail, we employ the real-time evolutionary trajectory information of the starting lineage (the reference sequence of *in silico* evolution) as the future mutational pattern, where all variant sequences (RBDs here) that appears after the starting lineage from GISAID

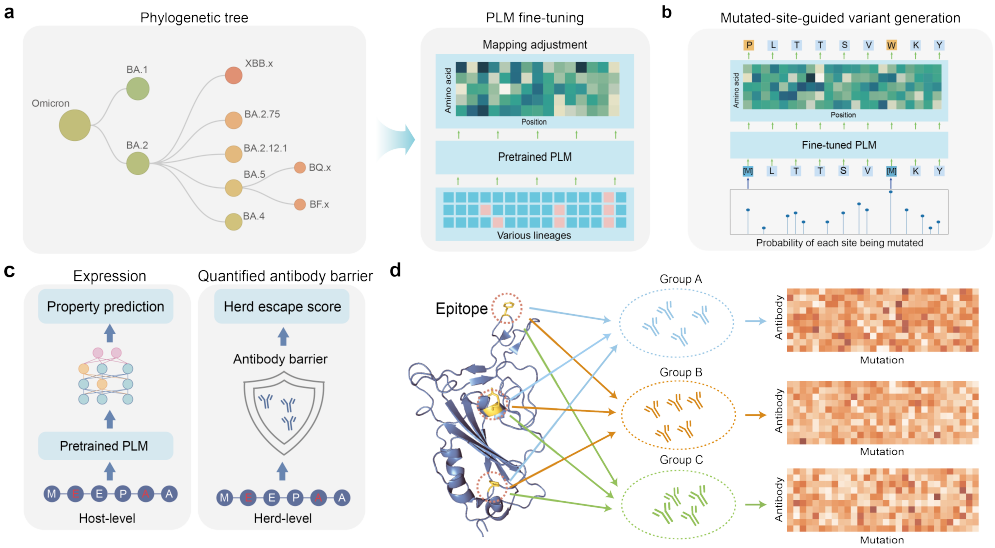

Fig. 2 Module details. **a**, Protein language model (PLM) fine-tuning module, in which variant sequences of specific lineages carrying real-time evolutionary information are adopted to fine-tune the pretrained PLM to adjust the learned evolutionary patterns towards the mutational fitness of virus. **b**, Mutated-site-guided variant generation module, where the probability of each site being mutated in real-time evolutionary trajectory is used as the probability of each site being masked (where to mutate) for the prediction of residue type to be mutated to (how to mutate). **c**, Host-to-herd selective pressure screening module, in which the *in silico* generated variant sequences are screened through the selective pressures of virus at host-level (using expression prediction model) and herd-level (using quantified antibody barrier model). **d**, Illustration of the proposed quantified antibody barrier model. Based on the deep mutational scanning data, monoclonal antibodies isolated from convalescent individuals are divided into multiple groups according to antigenic epitope, and the herd escape score of a variant is obtained by the average escape score of each of its mutations for each group.

database [2] [53] are integrated into a sequence set. Subsequently, the real-time virus evolutionary information of the starting lineage carried by the above sequence set is captured through unsupervised fine-tuning to adjust the initially learned mapping of position-by-position residue type distributions.

Second, after PLM fine-tuning, mutated-site-guided variant generation module is conducted to simulate the real-world mutation process of SARS-CoV-2 RBD (Fig.2b). In general, the mutation process is determined by two factors, i.e. where to mutate (determine the site to be mutated) and how to mutate (determine the mutation type at that site). For the determination of mutated sites, the site-by-site mutated probabilities of real-time evolutionary trajectory of the starting lineage are adopted to determine the sites to be mutated during variant generation. Specifically, we calculate the distribution of mutated probabilities of all sites in the integrated variant sequence set for PLM fine-tuning that carries the real-time evolutionary trajectory information. Subsequently, taking the starting lineage as the reference sequence of *in silico*

[2] https://gisaid.org/

evolution, the probability of each site being mutated is used as its masked probability, determining where to mutate. Finally, the fine-tuned PLM predicts the residue type to be mutated to, determining how to mutate. In this way, massive variants are generated on demand (Methods 4.3).

Third, host-to-herd selective pressure screening module is designed to screen the *in silico* generated variants through the selective pressures of viral proteins in two levels, i.e. host-level that determines survival and herd-level that determines transmission (Fig.2c). Based on our host-to-herd selective pressure simulation strategy, we consider the expression at the host level and the immunity barrier at the herd level. For the expression at the host level, a unified deep-learning framework for virus mutation drivers prediction [33] is adopted to predict the expression level of the *in silico* generated variants (Methods 4.4.1, Fig.2c left panel). For the immunity barrier at the herd level, we propose a quantified antibody barrier model to quantify the capability of variants to break through the herd-level immune barrier with herd escape scores (Fig.2c right panel). Specifically, as shown in Fig.2d, the deep mutational scanning (DMS) data of monoclonal antibodies isolated from convalescent individuals [8, 12] are adopted to calculate the herd escape scores of the *in silico* generated variants. According to the previous studies [8, 12], these isolated antibodies can be divided into multiple groups according to antigenic epitope. Considering the differences in the neutralization patterns of antibodies in these groups, for each mutation, we first calculate the average escape score in each antibody group. Subsequently, given that a variant may contain several mutations, we add up the average escape score corresponding to each mutation to obtain the group escape score of a variant for each antibody group. Finally, the herd escape score of a variant is calculated as the weighted sum of its corresponding group escape score for each antibody group. In this work, the proportion of antibodies in each group is used as the weight reflecting the importance of each antibody group (Methods 4.4.2).

Fourth, the mutation ranking module is conducted to recommend potential prevalent mutations in the future. Specifically, we first retain the variants with enhanced RBD expression (classification probability larger than 0.5) at the host level, and then further retain the top 50% variants in the herd escape score ranking list. Subsequently, we calculate the frequency of each mutation in the screened variant sequences. Finally, mutations are ranked from high frequency to low frequency, where the top-ranked mutations are considered to have the potential to become prevalent in the future.

2.2 Prediction reliability of GenPreMut

To ensure the reliability of GenPreMut, we conduct multiple ablation experiments from three perspectives, including the completeness of generation scale, prediction stability, and screening effectiveness. In the retrospective experimental scenario of this section, we define a lineage as starting lineage (the reference sequence of *in silico* evolution), and a lineage as cutoff lineage (the endpoint of variant sequence collection for PLM fine-tuning). Variant sequences between the starting and cutoff lineages are integrated into a sequence set for the PLM fine-tuning and calculation of mutated probabilities of all sites. The real-world prevalent mutations of the lineages emerging after the cutoff lineage along the phylogenetic tree are target mutations, i.e., the ground truth.

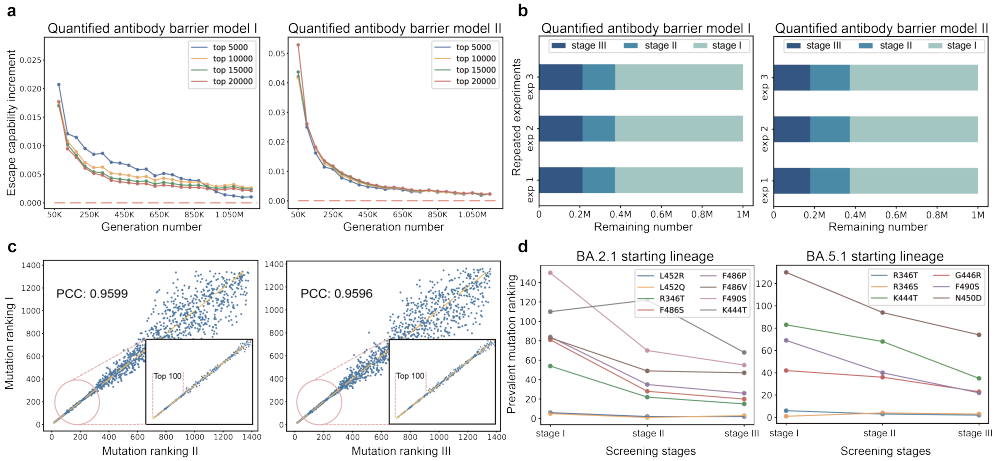

Fig. 3 Ablation experiments. **a**, Variant generation scale ablation experiments under two types of quantified antibody barrier models. The x-axis represents the number of generated variants with an interval of 50,000 and the y-axis represents the escape capability increment referring to the difference in the average herd escape score of the top K ($K = 5,000, 10,000, 15,000, 20,000$) variants sorted by scores in two consecutive generation experiments. **b**, The remaining number of variants at different screening stages of three repeated experiments under two types of quantified antibody barrier models. "Stage I" refers to the stage of initial variant generation, "Stage II" refers to the stage where the variants are screened by the expression prediction model, and "Stage III" refers to the stage where the variants are further screened by the quantified antibody barrier model. **c**, The pearson correlation coefficient (PCC) in mutation rankings across three repeated experiments. The higher the ranking of the mutation, the better the correlation across the repeated experiments. **d**, The ranking trends of previous real-world prevalent mutations (target mutations) across different screening stages. 8 of the 14 target mutations in BA.2.1 (i.e., 57%, left panel) and 6 of the 7 target mutations in BA.5.1 (i.e., 86%, right panel) are ranked higher and higher overall with the implementation of selective pressure screening models.

Importantly, the predicted ranking of a mutation takes the average ranking from two types of quantified antibody barrier models using different DMS data of monoclonal antibodies [8, 12] to mitigate the impact of experimental data bias (Methods 4.5).

First, to avoid missing potential prevalent mutations with strong immune escape ability across the population, we perform variant generation scale ablation experiment to determine the reasonable number of generated variants covering the fitness landscape of *in silico* SARS-CoV-2 evolution, i.e., the complete generation scale. In this ablation experiment, BA.2.1 are adopted as the starting lineage, with BA.3 as the cutoff lineage. Specifically, after fine-tuning the PLM using the integrated variant sequence set from BA.2.1 to BA.3 lineage, variant generation are conducted at different scales. As shown in Fig.3a, multiple sets of generation experiments with an interval of 50,000 are performed to explore whether the escape capability increment approaches 0 under two types of quantified antibody barrier models. The escape capability increment here is the difference in the average herd escape score of the top K variants sorted by scores in two consecutive generation experiments. We continue to expand the generation scale from an initial scale of 50,000 and observe the downward trend of escape capability increment under different K values, including 5,000, 10,000,

15,000, and 20,000. Overall, for both types of quantified antibody barrier models, the generation scale of one million makes the escape capability increment begin to stabilize and approach 0 when setting different K, which appears to cover the evolutionary fitness landscape for herd-level immune escape ability. Therefore, we set the variant generation scale to one million in all experiments of this work.

Second, considering the PLM fine-tuning and the variant generation modules involve randomness to the whole prediction pipeline, we conduct experiments from two perspectives to test the prediction stability of GenPreMut, including quantitative consistency across different screening stages (Fig.3b) and correlation of the mutation rankings in repeated experiments (Fig.3c). In this ablation experiment, BA.2.1 is adopted as the starting lineage, with BA.3 as the cutoff lineage. On the one hand, as shown in Fig.3b, we conduct three repeated experiments under the two types of quantified antibody barrier models to observe the remaining number of variants at different screening stages. Specifically, three stages are considered, namely, the stage of initial variant generation (Stage I), the stage where the variants are screened by the expression prediction model (Stage II), and the stage where the variants are further screened by two types of quantified antibody barrier models (Stage III). It can be noticed that regardless of the type of quantified antibody barrier model, the number of variants in each screening stage is very close (Supplementary information XX) in the three repeated experiments, demonstrating strong stability from the perspective of quantitative consistency. On the other hand, as shown in Fig.3c, the correlation in mutation rankings are observed across three repeated experiments. We can find that the mutation rankings of the three repeated experiments show strong correlation, with the Pearson Correlation Coefficient (PCC) exceeding 0.95. It is worth noting that the higher the ranking of the mutation, the better the correlation in the repeated experiments, and the top 100 mutations are close to perfect correlation, which ensures that our prediction results are stable because we ultimately recommend the top-ranked mutations as candidates for prevalent mutations in the future.

Third, to demonstrate the effectiveness of selective pressure screening, we visualize the ranking trends of previous real-world prevalent mutations (target mutations) across different screening stages in Fig.3d. For this ablation experiment, in addition to BA.2.1 as the starting lineage, BA.5.1 is also adopted as the starting lineage, which is paired with BQ.1 as the cutoff lineage. When BA.2.1 is used as the starting lineage, there are 14 target mutations (339H, 346T, 368I, 444T, 445P, 446S, 452Q, 452R, 460K, 478R, 486P, 486S, 486V, 490S), while when BA.5.1 is used as the starting lineage, there are 7 target mutations (346S, 346T, 444T, 446R, 450D, 460K, 490S). We can find that 8 of the 14 target mutations in BA.2.1 (i.e., 57%, Fig.3d left panel) and 6 of the 7 target mutations in BA.5.1 (i.e., 86%, Fig.3d right panel) continue to rank higher with the implementation of selective pressure screening models, including the expression prediction model in Stage II and two types of quantified antibody barrier models in Stage III. Overall, most of the previous real-world prevalent mutations are ranked higher and higher as the screening models are gradually executed, proving the contribution of the host-to-herd selective pressure screening module to the recommendation of potential prevalent mutations.

2.3 Reproduction of previous real-world prevalent mutations

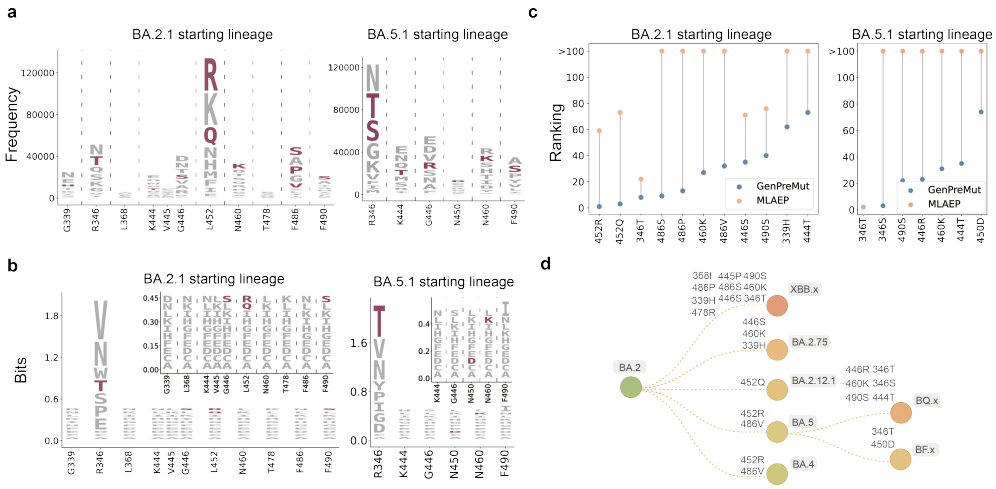

Fig. 4 Reproduction of previous real-world prevalent mutations. **ab**, Sequence logos of our GenPreMut (a) and the state-of-the-art method MLAEP (b) on the sites carrying previous real-world prevalent mutations with BA.2.1 (left panel) or BA.5.1 (right panel) as the starting lineage. A certain previous prevalent mutation ranked within the top 100 by GenPreMut or MLAEP is considered to be correctly predicted and colored in purple. For MLAEP, the sequence logos at sites other than R346 are enlarged for clear observation. **c**, Ranking improvement of correctly predicted previous real-world prevalent mutations with BA.2.1 (left panel) or BA.5.1 (right panel) as the starting lineage, in which the mutations ranked after 100 are uniformly represented by ">100". **d**, Annotated phylogenetic tree, where previous real-world prevalent mutations correctly predicted by our GenPreMut are annotated on the corresponding lineages. The "x" suffix represents a collection of related lineages.

To demonstrate the predictive power of GenPreMut, we perform *in silico* validation experiments on the recurrence of previous prevalent mutations taking BA.2.1 and BA.5.1 as the starting lineage, respectively. The settings of the *in silico* validation experiments are consistent with those of ablation experiments in Section 2.2. In addition, a certain previous prevalent mutation ranked within the top 100 by GenPreMut is considered to be correctly predicted.

To the best of our knowledge, MLAEP [31] is the state-of-the-art method for predicting antigenic evolution through *in silico* directed evolution, so it is used to compare the prediction performance with our GenPreMut (Supplementary information S4). In general, the number of previous prevalent mutations correctly predicted by GenPreMut (i.e., the prevalent mutations ranked within top 100) is two to seven times that of MLAEP (11 vs. 5 for BA.2.1 and 7 vs. 1 for BA.5.1), and all the correctly predicted prevalent mutations of MLAEP are included in the predicted ranking list of GenPreMut. In other words, GenPreMut's prediction accuracy is 42.9% higher than MLAEP's for BA.2.1 (78.6% vs. 35.7%) and 85.7% higher for BA.5.1 (100% vs. 14.3%).

As shown in Fig.4ab, the prediction performance of our GenPreMut and MLAEP is presented in the form of sequence logos [54, 55], where sites carrying previous

real-world prevalent mutations are selected to visualize the distribution of predicted mutations. The heights of the letters corresponding to the correctly predicted previous real-world prevalent mutations in GenPreMut's sequence logos are more uniform than those in MLAEP, indicating that GenPreMut ranks the the ground-truth target mutations higher overall. To show the improvement of GenPreMut's predicted ranking for previous prevalent mutations in more detail, we compare the rankings of correctly predicted target mutations in GenPreMut and MLAEP in Fig.4c, where mutations ranked after 100 are represented by ">100" for better visualization. We can find that, compared with MLAEP, our GenPreMut gives significantly higher rankings to previous real-world prevalent mutations, demonstrating a significant improvement in prediction performance.

As shown in Fig.4d, we annotate the previous real-world prevalent mutations correctly predicted by GenPreMut on the corresponding lineages in the phylogenetic tree. We find that whether BA.2.1 or BA.5.1 is used as the starting lineage, there are correctly predicted real-world prevalent mutations on the subsequent lineages. When BA.2.1 is taken as the starting lineage, although XBB lineages appear relatively late among its subsequent lineages, multiple prevalent mutations they contain are correctly predicted (the top panel of Fig.4d). This indicates that the potential prevalent mutations predicted by GenPreMut may not appear in the near future, but are likely to spread at a certain time point, which is closely related to the element of chance introduced by the transmission bottlenecks and geographical locations [3, 34, 56].

2.4 Recommendation of future prevalent mutations more than half a year in advance

After confirming GenPreMut's ability to reproduce previous prevalent mutations, we explore how to use it to recommend potential prevalent mutations in the future in this section. Specifically, the predictions of GenPreMut are validated *in vitro* to recommend mutations with potential to become prevalent in the future, with XBB.1.5 as the starting lineage. Unlike the retrospective experiments in the previous sections, the predictive experiments in this section do not involve a cutoff lineage. The recommendation pipeline consists of 5 steps as described below. In the first step, the variant sequences appearing after XBB.1.5 are integrated to fine-tune the PLM (ESM-2 with 650M parameters here). In the second step, the mutated-site-guided variant generation module is conducted to generate one million variants *in silico*. In the third step, host-to-herd selective pressure screening module is performed to identify high-risk variants with selective advantages. In the fourth step, we average the mutation rankings from the two types of quantified antibody barrier models to obtain the final mutation ranking list, and the top 12 point mutations are adopted for *in vitro* experimental validation at December 2023. In the fifth step, we recommend high-risk mutations based on the validation results and track their real-world prevalence at June 2024.

Entry efficiency is indicative of the viral invasiveness into host cells [57, 58], and viral evasion of neutralizing antibodies could exacerbate disease symptoms and facilitate transmissions [59–61]. Therefore, as shown in Fig.5ab, for the top 12 point mutations predicted by GenPreMut in the fourth step, we perform entry efficiency

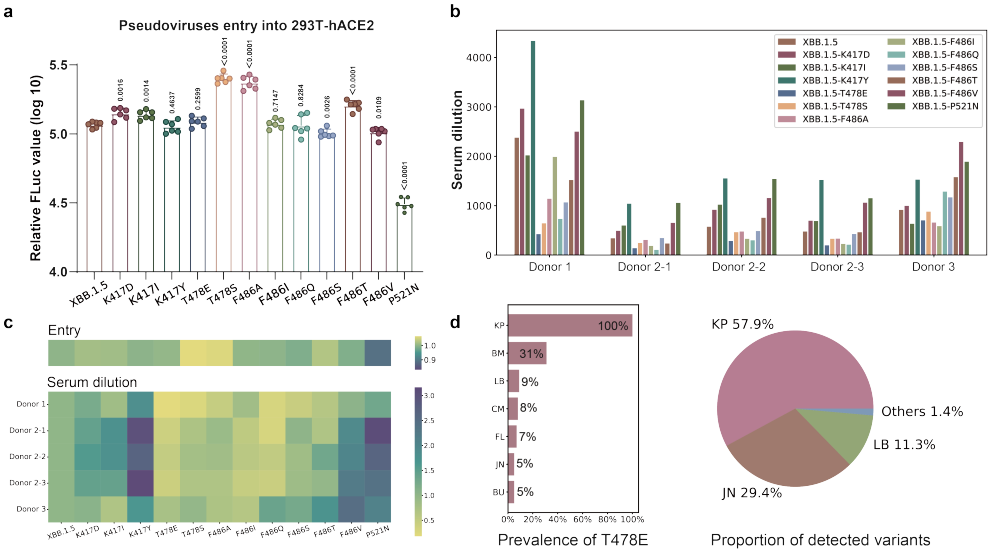

Fig. 5 *In vitro* validation and real-world prevalence analysis of recommended mutations. **a**, VSV-FLuc-based pseudotyed virus entry. 293T-hACE2 cells are infected with pseudoviruses containing original spike protein of XBB.1.5 and its variants carrying predicted point mutations. 24hr later, entry efficiency of pseudoviruses is determined by measuring FLuc activity in cell lysates. Error bars reveals the standard deviation of the means from six biological repeats and the data are analyzed by Student's t-test. **b**, Neutralizing assay with convalescent sera. Spike-pseudotyed viruses are incubated with serial-dilute sera for 1 hr and added to 293T-hACE2 cells. 24hr post infection, neutralizing IC50 of sera samples to different spike-pseudotyped viruses is calculated by measuring the FLuc activity. **c**, Summary of entry efficiency and serum dilution. Entry efficiency (upper panel) and serum dilution to reach IC50 (lower panel) of predicted mutated spike pseudoviruses are compared with original XBB.1.5 spike pseudovirus. For entry assay, yellow color means higher viral infectivity, and for neutralizing assay, yellow color means stronger immune evasion. **d**, Real-world prevalence analysis of recommended representative T478E mutation, including the mutation prevalence of T478E from outbreak.info and GISAID at June 2024 (left panel) and the proportion of detected variants from CDC COVID Data Tracker at the collection week of June 24, 2024 (right panel).

and neutralization experiments to determine their influences on viral entry efficiency and immune evasion.

First, in order to investigate the impact of the top 12 single-amino-acid mutations on viral entry stage in susceptible cells, vesicular stomatitis virus (VSV)-derived pseudovirus assays are conducted. Specifically, pseudoviruses containing original or mutated spike proteins of XBB.1.5 are packaged and diluted to equal copy numbers. Subsequently, 293T-hACE2 cells are infected by pseudoviruses, and the firefly luciferase (FLuc) activity of infected cells is quantified (Methods 4.6.1). As shown in Fig.5a, compared with the original XBB.1.5 spike protein, seven variants with single-amino-acid mutations, including K417D, K417I, T478E, T478S, F486A, F486I, and F486T, exhibit higher FLuc activity. Notably, T478S, F486A, and F486T show significantly increased entry efficiency (P value < 0.0001), suggesting enhanced invasiveness into susceptible cells. Conversely, mutations like F486S, F486V, and P521N in the RBD

of XBB.1.5 reduce infection in 293T-hACE2 cells, particularly the P521N mutation. Mutations such as K417Y, T478E, F486I, and F486Q do not affect the entry process significantly, with minimal differences in entry efficiency compared to the original XBB.1.5 spike protein.

Second, the evasive capabilities of these variants carrying predicted mutations for neutralizing antibodies are further evaluated. Specifically, Sera from three convalescent individuals are serially diluted and incubated with an equal volume of pseudoviruses at 37 ℃. Following incubation, 293T-hACE2 cells are infected with the serum-pseudovirus mixture. The half-maximal inhibitory concentration (IC50) of these sera against the original XBB.1.5 spike protein and its variants is determined by measured FLuc activity (Methods 4.6.2). As shown in Fig.5b, among the predicted mutations, T478E, T478S, F486A, and F486I demonstrate substantial immune evasion against all five sera tested. F486Q and F486S attenuate the neutralizing ability of four sera, whereas F486T is not effectively neutralized by three sera.

According to the overall results of the *in vitro* validation in Fig.5c, T478E, T478S, F486A, and F486I appear to pose greater risks among the 12 predicted mutations, as they display increased entry efficiency in susceptible cell lines and enhanced immune evasion against all the convalescent sera compared with the original XBB.1.5 spike protein. Therefore, the above four mutations are recommended for real-world prevalence analysis in the fifth step. Specifically, with GISAID [3], outbreak.info [4], and CDC COVID Data Tracker [5] as references, the above four mutations are tracked for prevalence in the real world at June 2024. According to the prevalence data of real world, three of the four mutations (including T478E, F486A, F486I) have appeared at extremely high frequency (more than 80% of all sequences) in multiple lineages during the period from the time of *in vitro* experimental validation (December 2023) to the time of prevalence analysis (June 2024), confirming their widespread prevalence in the real world. As shown in Fig.5d left panel, taking the representative mutation T478E as an example, its mutation prevalence from outbreak.info and GISAID at June 2024 shows that 100% of the variant sequences with the prefix KP contain this mutation, and the prefixes of the remaining lineages containing this mutation that account for no less than 5% are BM (31%), LB (9%), CM (8%), FL (7%), JN (5%), and BU (5%), respectively. Meanwhile, as shown in Fig.5d right panel, the proportion of detected variants from CDC COVID Data Tracker (the collection week of June 24, 2024) shows that the lineage prefixed with KP is dominant (57.9%), and the remaining lineages with a proportion of more than 10% are prefixed with JN (29.4%) and LB (11.3%). Combining the above results, we can find that the prevalence of predicted T478E has a clear correlation with the proportion of variants detected in the real world, which indicates that the point mutations recommended by our prediction pipeline not only occur widely in the real world, but also have the potential to dominate the pandemic.

[3] https://gisaid.org/
[4] https://outbreak.info/
[5] https://covid.cdc.gov/covid-data-tracker/#datatracker-home

3 Discussion

As a generative prediction framework, GenPreMut achieves anchoring of potential prevalent mutations in a huge evolutionary space through *in silico* virus evolution based on the host-to-herd selective pressure simulation strategy. The ablation experimental results show that GenPreMut has achieved the completeness of the generation scale, maintained the prediction stability, and ensured the screening effectiveness. Overall, GenPreMut has been proven to have superior prediction ability from two perspectives, including the reproduction of previous prevalent mutations and the recommendation of potential prevalent mutations in the future. First, GenPreMut accurately reproduces previous prevalent mutations and achieves 42.9% and 85.7% improvement in prediction performance on two representative lineages, respectively. Second, the predicted mutations verified by *in vitro* experiments are recommended as future prevalent mutations more than half a year in advance, and the real-world prevalence analysis shows that the mutations recommended by GenPreMut not only occur widely in the real world, but also dominate the pandemic.

GenPreMut still has limitations, lacking consideration of transmissibility, pathogenicity, virulence, and geographic location. The above factors lack refined machine learning modeling strategies, limited by insufficient data or dynamic changes. For transmissibility, it is gratifying that there are some mathematical forms for SARS-CoV-2 infectiousness [62, 63], but the difficulty of obtaining real-time transmission data cannot be ignored. For pathogenicity and virulence, the isolation of existing clinical data between different hospitals makes it difficult to integrate for machine learning modeling. For geographic location, the different public health strategies adopted by different countries or regions make the impact of geographic location complex and variable, significantly increasing the difficulty of using mathematical forms to quantitatively describe it. Therefore, in the future, introducing some reasonable mathematical forms for transmissibility, developing machine learning modeling methods based on a small amount of clinical data for pathophysiology, and modeling the impact of geographic location as finely as possible are three potential directions to improve the predictive performance of GenPreMut.

4 Methods

4.1 Data

4.1.1 Dataset for PLM fine-tuning and variant generation

The sequence set for PLM fine-tuning and variant generation are integrated from GISAID database [6] (downloaded at December 2023). Specifically, for the ablation experiments in Section 2.2 and *in silico* validation experiments in Section 2.3, the RBD sequences of variants between the starting and cutoff lineages are integrated into a sequence set for the PLM fine-tuning and calculation of mutated probabilities of all sites. In addition, for the predictive experiments in Section 2.4, the RBD sequences of

[6] https://gisaid.org/

variants that appears after the starting lineage are integrated into a sequence set for the PLM fine-tuning and calculation of mutated probabilities of all sites.

4.1.2 Dataset for host-level screening model

For host-level expression prediction, the deep mutational scanning (DMS) datasets of previous work [11] are adopted to train a predictive model. There are seven groups of DMS datasets including wild type, Alpha, Beta, Delta, Eta, Omicron BA.1 and Omicron BA.2. We only use the samples of Omicron BA.1 and Omicron BA.2 due to the notable difference in lineages before and after Omicron [64, 65]. We normalize the RBD expression labels with the value of Omicron BA.2 to get the final labels for training. Positive samples are defined as those with labels greater than 1, while others are negative samples. Finally, there are 7,182 negative samples and 639 positive samples in the training dataset.

4.1.3 Dataset for herd-level screening model

For herd-level escape score prediction, two sets of DMS datasets from previous studies [8, 12] are adopted to build the quantified antibody barrier model that quantifies the capability of variants to break through the herd-level immune barrier. There are high-throughput mutation escape profiles for thousands of antibodies in the above DMS datasets, including mutations, escape scores, antibody names, and the group that each antibody belongs to. In the work of [8], the antibodies were isolated from the convalescents of SARS, SARS-CoV-2 Wild-type, BA.1, BA.2 and BA.5 variants. In the work of [12], the antibodies were isolated from the convalescents of BA.1, BA.2, BA.5 and BF.7 variants, the reinfection convalescents of BA.1, BA.2 variants and long-term BA.1 convalescents. Since Omicron sub-lineages are currently prevalent, we only use the DMS data of antibodies isolated from convalescents of Omicron variants.

4.2 Protein language model fine-tuning

For a given starting lineage x_p with length L, we collect the variant sequences emerging after x_p (also before the cutoff lineage x_t the ablation experiments in Section 2.2 and *in silico* validation experiments in Section 2.3) to build a sequence set F. The integrated sequence set F is used to fine-tune the pretrained PLM with the masked residue prediction task. The widely recognized ESM-2 (650M) is adopted in this work, which can be replaced by any powerful pretrained PLM, such as ProtTrans [51] and ProteinBERT [66]. To balance the novelty and diversity of the generated variant sequences, we set the fine-tuning epoch as 1 epoch so that the PLM is trained on all the collected variant sequences while preserving the diversity of generation.

4.3 Variant generation

The core of variant generation lies in two points: where to mutate (determine the site to be mutated) and how to mutate (determine the mutation type at that site). First, we calculate the mutated probability of each site P using sequences in F as

$$P_i = \frac{1}{|F|}\sum_f (1 - \mathbb{I}(f[i] == x_s[i])), f \in F, i \in [1, L], \tag{1}$$

where $|F|$ denotes the number of sequences in F, and $\mathbb{I}(x) = 1$ if x is $true$, otherwise $\mathbb{I}(x) = 0$.

Second, with the calculated P and the fine-tuned PLM, we conduct variant sequence generation through a mutation process based on the starting lineage x_p. Each site i of x_p performs a binomial trial with the success probability of P_i, i.e. $p_i \sim B(P_i)$, and the site i is updated as

$$x_s^{'}[i] = p_i \cdot [MASK] + (1 - p_i) \cdot x_s[i], p_i \in \{0, 1\}, \tag{2}$$

where $x_p[i]$ denotes the ith residue of the sequence and $[MASK]$ denotes the mask token. The equation means that each residue i has the probability of P_i to be replaced by mask token. We reject the $x_s^{'}$ with mask token number greater than MAX_{mask}. In our experiments, MAX_{mask} is set as 5.

Third, the fine-tuned PLM FT is used to complete the masked sequences following the adjusted residue distribution as

$$\hat{x}_s[i] = \mathrm{FT}(x_s^{'})[i], i \in \{i | x_s^{'}[i] = [MASK]\}. \tag{3}$$

The above generation process is repeated until the generated variant sequences reach the generation scale S.

4.4 Screening models

4.4.1 Host-level expression prediction model

The host-level expression prediction model adopts the deep-learning framework proposed by previous work [33], which includes three steps, namely sequence embedding extraction, local-global dependence coupling and multi-task focal learning. The implementation details are as follows. First, for sequence embedding extraction, the off-the-shelf pretrained PLM is adopted to extract amino-acid-level embeddings of mutated RBD sequences. Second, the CNN and ContextPool-attention [67] are adopted to capture the local and global dependencies of the mutations respectively. The one-dimension CNN adopted here is to build the impact of a mutation on their sequentially neighboring residues, i.e., the local dependencies of the mutation. It consists of 3 convolution layers with Layer Normalization and uses leaky rectified linear unit (Leaky ReLU) as activation function. Correspondingly, the ContextPool-attention with dynamic granularity is adopted here to build the motif-level global dependencies of a mutation. The Gaussian mask vector is normalized with its maximum value during calculation for stable training. The above local feature and motif-level global feature of mutations are subsequently concatenated and max-pooled on the residue dimension to obtain the protein-level embeddings. Third, the model is trained through multi-task learning of binary classification task (expression increases or not) and regression task (expression strength estimation), where a multi-task focal loss is adopted to alleviate the imbalance between positive and negative samples mentioned in Methods 4.1.2.

4.4.2 Herd-level quantified antibody barrier model

The DMS datasets from previous studies [8, 12] mentioned in Methods 4.1.3 are adopted to build the quantified antibody barrier model, in which the isolated antibodies can be divided into multiple groups according to antigenic epitope for the calculation of herd escape scores.

First, considering the differences in the neutralization patterns of antibodies in different groups, we calculate the average escape score in each antibody group for each mutation. Specifically, for the antibody group G, we define $|G|$ as the number of antibodies of G. For the mutation (s, m), where s denotes mutated site and m denotes the residue mutated to, we calculate the average of all the escape scores i of this mutation in group G, i.e.

$$f_G(s, m) = \frac{1}{|G|} \sum_{site_i = s,\ mut_i = m} score_i, \tag{4}$$

where *site* denotes the mutated site, *mut* denotes the residue type and *score* denotes the escape score. Here, we can get a function for each group that maps mutation to the average escape score of the group, i.e.

$$f_G : (s, m) \rightarrow score_G. \tag{5}$$

Second, for a generated variant (i.e. mutated RBD), it may contain many mutations compared with wild-type RBD sequence, which can be denoted as $\{(s_1, m_1), (s_2, m_2), ..., (s_n, m_n)\}$. Based on the mapping function above, we sum the average escape scores of all the mutations as the group escape score of the variant for each antibody group. For the antibody group G, the group escape score of a variant is calculated as

$$F_G = \sum_{i=1}^{n} f_G(s_i, m_i). \tag{6}$$

Third, the group escape scores of different antibody groups are weighted summed based on the importance of each antibody group to get the final herd escape score $Barrier$, i.e.

$$Barrier = \sum_{G} w_G F_G, \tag{7}$$

where w_G denotes the weight of antibody group G. Here, we use the proportion of antibodies in each group as weight.

4.5 Mutation ranking criterion

For the prediction results of host-level expression prediction model, we select the variants with enhanced expression level compared to BA.2, i.e., the variants with output probabilities of classification prediction larger than 0.5 is remained. Subsequently, for the prediction results of herd-level quantified antibody barrier model, we sort all the variants with herd escape score, and select top 50% variants with relatively high escaping capability against population. Finally, the frequency of each mutation in the

screened variant sequences are calculated for ranking, where the top-ranked mutations are considered to have the potential to become prevalent in the future. It should be noted that two sets of DMS datasets from previous studies [8, 12] are adopted to build two types of quantified antibody barrier models, and the two rankings for the same mutation from the above two types of quantified antibody barrier models are averaged to mitigate the impact of experimental data bias to obtain the final ranking result.

4.6 *In vitro* experimental validation

4.6.1 Pseudotyped viruses entry 293T-hACE2 cells

First, we construct plasmids of SARS-CoV-2 XBB.1.5 and its variants with the top 12 predicted mutations. The spike sequences are mammalian codon-optimized and introduced to pCDNA3.1(+) vector by Genscript company. For package of pseudotyped viruses, spike plasmids are tranfected to 293T cells with lipofectamine 3000 (Invitrogen). 24 hr later, VSV-ΔG-Fluc is added to cells and incubated for 8 hr. The supernatants are subsequently discarded and cells are washed with PBS for three times. DMEM (Gibco) containing 2% of FBS (Omega) are added to culture cells for another 24 hr. The supernatants containing pseudotyped viruses are harvested and centrifuged at 1000 g for 10 min to remove the cell debris. Pseudotyped viruses are then aliquoted and stored at -80℃. Titer of pseudotyped viruses are determined by infecting 293T-hACE2 cells. To compare the entry efficiency to 293T-hACE2 cells, spike-pseudotype viruses of wild-type and variants of XBB.1.5 are diluted to the same copy numbers before use. 293T-hACE2 cells are seeded to 96-well plates and cultured with DMEM containing 10% FBS for 24 hr. Pseudotyped viruses are added to cells and incubated at 37℃. 24 hr later, the supernatants are discarded and cells are treated with Bright-Lite Luciferase Assay System (Vazyme). Entry efficiency is evaluated by measuring FLuc activity.

4.6.2 Neutralizing assay with convalesent serum

The pseudotyped viruses containing original and mutated spike proteins of XBB 1.5 are diluted to 1.3 E+04 TCID50/mL. 7.5μL of convalescent sera are added to 142.5μL of DMEM medium and six three-fold serial gradient dilutions are performed with 50μL of mixture. 50μL of pseudotyped viruses are subsequently added to 96-well plates containing diluted sera and incubated at 37℃ for 1 hr. After incubation, 7E+04 of 293T-hACE2 cells are added to each well containing serum-pseudovirus mixture and cultured for 24 hr. Cells are washed once with PBS and lysed with same reagent described above. IC50 of convalescent sera for pseudoviruses is calculated.

Data availability

The sequence set for PLM fine-tuning and variant generation is freely available at https://gisaid.org/. The deep mutational scanning datasets of screening models are publicly available from previous studies [8, 11, 12]. The data used in our experiments can be reached at https://github.com/Kevinatil/GenPreMut.

Code availability

Relevant code and models are available at https://github.com/Kevinatil/GenPreMut.

References

[1] Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature* **579**, 270–273 (2020).

[2] Hu, B., Guo, H., Zhou, P. & Shi, Z.-L. Characteristics of sars-cov-2 and covid-19. *Nature reviews microbiology* **19**, 141–154 (2021).

[3] Markov, P. V. *et al.* The evolution of sars-cov-2. *Nature Reviews Microbiology* **21**, 361–379 (2023).

[4] Greaney, A. J. *et al.* Mapping mutations to the sars-cov-2 rbd that escape binding by different classes of antibodies. *Nature communications* **12**, 4196 (2021).

[5] Starr, T. N. *et al.* Shifting mutational constraints in the sars-cov-2 receptor-binding domain during viral evolution. *Science* **377**, 420–424 (2022).

[6] Harrison, A. G., Lin, T. & Wang, P. Mechanisms of sars-cov-2 transmission and pathogenesis. *Trends in immunology* **41**, 1100–1115 (2020).

[7] Anderson, R. M., Vegvari, C., Truscott, J. & Collyer, B. S. Challenges in creating herd immunity to sars-cov-2 infection by mass vaccination. *The Lancet* **396**, 1614–1616 (2020).

[8] Cao, Y. *et al.* Imprinted sars-cov-2 humoral immunity induces convergent omicron rbd evolution. *Nature* **614**, 521–529 (2023).

[9] Roemer, C. *et al.* Sars-cov-2 evolution in the omicron era. *Nature microbiology* **8**, 1952–1959 (2023).

[10] 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 (2020).

[11] Starr, T. N. *et al.* Deep mutational scans for ace2 binding, rbd expression, and antibody escape in the sars-cov-2 omicron ba. 1 and ba. 2 receptor-binding domains. *PLoS pathogens* **18**, e1010951 (2022).

[12] Yisimayi, A. *et al.* Repeated omicron exposures override ancestral sars-cov-2 immune imprinting. *Nature* **625**, 148–156 (2024).

[13] Lei, R. *et al.* Mutational fitness landscape of human influenza h3n2 neuraminidase. *Cell reports* **42** (2023).

[14] Sourisseau, M. *et al.* Deep mutational scanning comprehensively maps how zika envelope protein mutations affect viral growth and antibody escape. *Journal of virology* **93**, 10–1128 (2019).

[15] Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J. & Bloom, J. D. Mapping mutational effects along the evolutionary landscape of hiv envelope. *Elife* **7**, e34420 (2018).

[16] Wu, N. C. *et al.* High-throughput profiling of influenza a virus hemagglutinin gene at single-nucleotide resolution. *Scientific reports* **4**, 4942 (2014).

[17] Doud, M. B. & Bloom, J. D. Accurate measurement of the effects of all amino-acid mutations on influenza hemagglutinin. *Viruses* **8**, 155 (2016).

[18] Soh, Y. S., Moncla, L. H., Eguia, R., Bedford, T. & Bloom, J. D. Comprehensive mapping of adaptation of the avian influenza polymerase protein pb2 to humans. *Elife* **8**, e45079 (2019).

[19] Araya, C. L. & Fowler, D. M. Deep mutational scanning: assessing protein function on a massive scale. *Trends in biotechnology* **29**, 435–442 (2011).

[20] Chen, J., Wang, R., Wang, M. & Wei, G.-W. Mutations strengthened sars-cov-2 infectivity. *Journal of molecular biology* **432**, 5212–5226 (2020).

[21] Taft, J. M. *et al.* Deep mutational learning predicts ace2 binding and antibody escape to combinatorial mutations in the sars-cov-2 receptor-binding domain. *Cell* **185**, 4008–4022 (2022).

[22] Wang, G. *et al.* Deep-learning-enabled protein–protein interaction analysis for prediction of sars-cov-2 infectivity and variant evolution. *Nature Medicine* **29**, 2007–2018 (2023).

[23] Chen, C. *et al.* Computational prediction of the effect of amino acid changes on the binding affinity between sars-cov-2 spike rbd and human ace2. *Proceedings of the National Academy of Sciences* **118**, e2106480118 (2021).

[24] Han, J. *et al.* D3ai-spike: A deep learning platform for predicting binding affinity between sars-cov-2 spike receptor binding domain with multiple amino acid mutations and human angiotensin-converting enzyme 2. *Computers in Biology and Medicine* **151**, 106212 (2022).

[25] Wang, E. Prediction of antibody binding to sars-cov-2 rbds. *Bioinformatics Advances* **3**, vbac103 (2023).

[26] Swanson, K., Chang, H. & Zou, J. *Predicting immune escape with pretrained protein language model embeddings*, 110–130 (PMLR, 2022).

[27] Thadani, N. N. *et al.* Learning from prepandemic data to forecast viral escape. *Nature* **622**, 818–825 (2023).

[28] Hie, B., Zhong, E. D., Berger, B. & Bryson, B. Learning the language of viral evolution and escape. *Science* **371**, 284–288 (2021).

[29] Hie, B. L., Yang, K. K. & Kim, P. S. Evolutionary velocity with protein language models predicts evolutionary dynamics of diverse proteins. *Cell Systems* **13**, 274–285 (2022).

[30] Obermeyer, F. *et al.* Analysis of 6.4 million sars-cov-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).

[31] Han, W. *et al.* Predicting the antigenic evolution of sars-cov-2 with deep learning. *Nature Communications* **14**, 3478 (2023).

[32] Chen, J. *et al.* Running ahead of evolution—ai-based simulation for predicting future high-risk sars-cov-2 variants. *The International Journal of High Performance Computing Applications* **37**, 650–665 (2023).

[33] Nie, Z. *et al.* Evolution-guided large language model is a predictor of virus mutation trends. *bioRxiv* 2023–11 (2023).

[34] Zwart, M. P. & Elena, S. F. Matters of size: genetic bottlenecks in virus infection and their potential impact on evolution. *Annual Review of Virology* **2**, 161–179 (2015).

[35] Magazine, N. *et al.* Mutations and evolution of the sars-cov-2 spike protein. *Viruses* **14**, 640 (2022).

[36] Harari, S. *et al.* Drivers of adaptive evolution during chronic sars-cov-2 infections. *Nature medicine* **28**, 1501–1508 (2022).

[37] Moulana, A. *et al.* The landscape of antibody binding affinity in sars-cov-2 omicron ba. 1 evolution. *Elife* **12**, e83442 (2023).

[38] Dejnirattisai, W. *et al.* Sars-cov-2 omicron-b. 1.1. 529 leads to widespread escape from neutralizing antibody responses. *Cell* **185**, 467–484 (2022).

[39] Starr, T. N. *et al.* Sars-cov-2 rbd antibodies that maximize breadth and resistance to escape. *Nature* **597**, 97–102 (2021).

[40] Hong, Q. *et al.* Molecular basis of receptor binding and antibody neutralization of omicron. *Nature* **604**, 546–552 (2022).

[41] Meng, B. *et al.* Altered tmprss2 usage by sars-cov-2 omicron impacts infectivity and fusogenicity. *Nature* **603**, 706–714 (2022).

[42] Cameroni, E. *et al.* Broadly neutralizing antibodies overcome sars-cov-2 omicron antigenic shift. *Nature* **602**, 664–670 (2022).

[43] Triveri, A. *et al.* Sars-cov-2 spike protein mutations and escape from antibodies: a computational model of epitope loss in variants of concern. *Journal of Chemical Information and Modeling* **61**, 4687–4700 (2021).

[44] Gruell, H. *et al.* Neutralisation sensitivity of the sars-cov-2 omicron ba. 2.75 sublineage. *The Lancet Infectious Diseases* **22**, 1422–1423 (2022).

[45] Wu, L. *et al.* Sars-cov-2 omicron rbd shows weaker binding affinity than the currently dominant delta variant to human ace2. *Signal transduction and targeted therapy* **7**, 8 (2022).

[46] Imai, M. *et al.* Efficacy of antiviral agents against omicron subvariants bq. 1.1 and xbb. *New England Journal of Medicine* **388**, 89–91 (2023).

[47] Omer, S. B., Yildirim, I. & Forman, H. P. Herd immunity and implications for sars-cov-2 control. *Jama* **324**, 2095–2096 (2020).

[48] Britton, T., Ball, F. & Trapman, P. A mathematical model reveals the influence of population heterogeneity on herd immunity to sars-cov-2. *science* **369**, 846–849 (2020).

[49] Chowdhury, R. *et al.* Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology* **40**, 1617–1623 (2022).

[50] Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

[51] Elnaggar, A. *et al.* Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence* **44**, 7112–7127 (2021).

[52] Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics* **55**, 1512–1522 (2023).

[53] Shu, Y. & McCauley, J. Gisaid: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **22**, 30494 (2017).

[54] Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. Weblogo: a sequence logo generator. *Genome research* **14**, 1188–1190 (2004).

[55] Schneider, T. D. & Stephens, R. M. Sequence logos: a new way to display consensus sequences. *Nucleic acids research* **18**, 6097–6100 (1990).

[56] Grubaugh, N. D., Hanage, W. P. & Rasmussen, A. L. Making sense of mutation: what d614g means for the covid-19 pandemic remains unclear. *Cell* **182**, 794–795 (2020).

[57] Kalia, M. & Jameel, S. Virus entry paradigms. *Amino acids* **41**, 1147–1157 (2011).

[58] Schneider, U., Von Messling, V., Devaux, P. & Cattaneo, R. Efficiency of measles virus entry and dissemination through different receptors. *Journal of virology* **76**, 7460–7467 (2002).

[59] Vossen, M. T., Westerhout, E. M., Söderberg-Nauclér, C. & Wiertz, E. J. Viral immune evasion: a masterpiece of evolution. *Immunogenetics* **54**, 527–542 (2002).

[60] Thorne, L. G. *et al.* Evolution of enhanced innate immune evasion by sars-cov-2. *Nature* **602**, 487–495 (2022).

[61] Minkoff, J. M. & tenOever, B. Innate immune evasion strategies of sars-cov-2. *Nature Reviews Microbiology* **21**, 178–194 (2023).

[62] Grassly, N. C. & Fraser, C. Mathematical models of infectious disease transmission. *Nature Reviews Microbiology* **6**, 477–487 (2008).

[63] Ferretti, L. *et al.* Quantifying sars-cov-2 transmission suggests epidemic control with digital contact tracing. *science* **368**, eabb6936 (2020).

[64] Karim, S. S. A. & Karim, Q. A. Omicron sars-cov-2 variant: a new chapter in the covid-19 pandemic. *The lancet* **398**, 2126–2128 (2021).

[65] Fan, Y. *et al.* Sars-cov-2 omicron variant: recent progress and future perspectives. *Signal transduction and targeted therapy* **7**, 1–11 (2022).

[66] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics* **38**, 2102–2110 (2022).

[67] Huang, C., Talbott, W., Jaitly, N. & Susskind, J. M. *Efficient representation learning via adaptive context pooling*, 9346–9355 (PMLR, 2022).

Acknowledgements

We appreciate the support of AI for Science (AI4S)-Preferred Program, Peking University Shenzhen Graduate School, China. This work was financially supported by the National Key R&D Program of China (No. 2022ZD0118201), Natural Science Foundation of China (No. 61972217, 32071459, 62176249, 62006133, 62271465, 61825101, 62088102), R&D Program of Guangzhou Laboratory (Grant No.SRPG22-001, QNPG23-07, ZL-EKPG21030305).

Competing interests

The authors declare no competing interests.

Supplementary information

S1 Variant generation scale ablation experiments

The data of variant generation scale ablation experiments under two types of quantified antibody barrier models are stored in two separate files (antibody_barrier_model_I.csv, antibody_barrier_model_II.csv).

S2 The remaining number of variants at different screening stages

Repeat experiment	Stage I	Stage II	Stage III
Experiment 1	1,000,000	373,201	213,784
Experiment 2	1,000,000	373,313	214,045
Experiment 3	1,000,000	373,418	213,597

Table S1 The remaining number of variants at different stages under quantified antibody barrier model I

Repeat experiment	Stage I	Stage II	Stage III
Experiment 1	1,000,000	373,201	179,474
Experiment 2	1,000,000	373,313	179,519
Experiment 3	1,000,000	373,418	179,477

Table S2 The remaining number of variants at different stages under quantified antibody barrier model II

S3 The ranking trends of previous real-world prevalent mutations across different screening stages

Mutation	stage I	stage II	stage III
L452R	6	2	2
L452Q	5	1	3
R346T	54	22	15
F486S	81	28	20
F486P	84	35	26
F486V	83	49	47
F490S	150	70	55
K444T	110	122	68

Table S3 The ranking trends of previous real-world prevalent mutations across different screening stages with BA.2.1 as the starting lineage

Mutation	stage I	stage II	stage III
R346T	7	4	3
R346S	2	5	4
K444T	84	69	36
G446R	43	37	24
F490S	70	41	23
N450D	131	95	75

Table S4 The ranking trends of previous real-world prevalent mutations across different screening stages with BA.5.1 as the starting lineage

S4 Ranking improvement of correctly predicted previous real-world prevalent mutations

Mutation	GenPreMut	MLAEP
452R	1	59
452Q	3	73
346T	8	22
486S	9	2,397
486P	13	2,768
460K	27	1,366
486V	32	2,376
446S	35	71
490S	40	76
339H	62	218
444T	73	851

Table S5 Ranking improvement of correctly predicted previous real-world prevalent mutations with BA.2.1 as the starting lineage

Mutation	GenPreMut	MLAEP
346T	2	2
346S	3	3,606
490S	22	3,615
446R	23	3,364
460K	31	276
444T	35	2,578
450D	74	341

Table S6 Ranking improvement of correctly predicted previous real-world prevalent mutations with BA.5.1 as the starting lineage

720 # S5 VSV-FLuc-based pseudotyed virus entry

Variant	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	Replicate 6
XBB.1.5	5.043944	5.031437	5.082566	5.088551	5.070030	5.079000
XBB.1.5-K417D	5.140725	5.116847	5.169771	5.181701	5.085687	5.186795
XBB.1.5-K417I	5.156855	5.113629	5.156161	5.179865	5.119470	5.094090
XBB.1.5-K417Y	5.108128	5.023075	5.086132	5.070056	5.015984	5.000469
XBB.1.5-T478E	5.131070	5.092124	5.111834	5.056794	5.036525	5.089202
XBB.1.5-T478S	5.376013	5.458484	5.403956	5.403354	5.401323	5.363738
XBB.1.5-F486A	5.405845	5.429111	5.327661	5.393680	5.352414	5.310732
XBB.1.5-F486I	5.100277	5.026758	5.066699	5.079000	5.116538	5.043896
XBB.1.5-F486Q	5.019951	5.152187	5.042044	5.039486	5.141183	4.958134
XBB.1.5-F486S	4.979862	5.059609	4.983739	5.018572	4.987675	4.992973
XBB.1.5-F486T	5.225066	5.248432	5.144200	5.182318	5.214195	5.212944
XBB.1.5-F486V	5.035434	5.018426	4.938945	5.001561	5.036094	5.030462
XBB.1.5-P521N	4.542539	4.540342	4.491025	4.427681	4.467105	4.476716

Table S7 The relative FLuc value (log 10) of the entry of pseudoviruses containing original spike protein of XBB.1.5 and its variants carrying predicted point mutations

721 # S6 Neutralizing assay with convalescent sera

Variant	Donor 1	Donor 2-1	Donor 2-2	Donor 2-3	Donor 3
XBB.1.5	2379	340.9	576.2	480.1	916.5
XBB.1.5-K417D	2965	493.6	918.4	696.9	998
XBB.1.5-K417I	2020	600.2	1022	689.7	636.2
XBB.1.5-K417Y	4335	1041	1554	1521	1529
XBB.1.5-T478E	423.3	140.2	287.8	198.6	704.2
XBB.1.5-T478S	642.8	246.1	463.8	332.3	883.1
XBB.1.5-F486A	1143	307.8	480.7	335.2	662.3
XBB.1.5-F486I	1988	187.2	332.1	227.2	585.8
XBB.1.5-F486Q	732.7	107.5	299.9	213	1287
XBB.1.5-F486S	1069	345.4	487.2	426.7	1170
XBB.1.5-F486T	1519	236.1	756.9	463.5	1577
XBB.1.5-F486V	2501	653.5	1155	1063	2291
XBB.1.5-P521N	3135	1058	1541	1153	1892

Table S8 Serum dilution of pseudoviruses containing original spike protein of XBB.1.5 and its variants carrying predicted point mutations

S7 Hyperparameters of host-level expression prediction model

Hyperparameter	Value
Learning rate	0.00002
Number of epochs	200
γ	1
β	0.93
ESM embedding dim	1280 (650M)
Dropout rate	0.6
Convolution layers	3
Kernel size	3

Table S9 Hyperparameters of expression prediction model