Predicting the antigenic evolution of SARS-COV-2 with deep learning

Wenkai Han[1,2,#], Ningning Chen[1,2,#], Yu Wang[3], Shiwei Sun[4,5], Xin Gao[1,2,*]

[1]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Kingdom of Saudi Arabia

[2]Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia

[3]Syneron Technology, Guangzhou, 510000, China

[4]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

[5]University of Chinese Academy of Sciences, Beijing 100049, China

[#]The first two authors contributed equally to this paper.

*To whom all correspondence should be addressed. E-mail: xin.gao@kaust.edu.sa, Tel: +966-12-8080323.

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) antigenic profile evolves in response to the vaccine and natural infection-derived immune pressure, resulting in immune escape and threatening public health. Exploring the possible antigenic evolutionary potentials improves public health preparedness, but it is limited by the lack of experimental assays as the sequence space is exponentially large. Here we introduce the Machine Learning-guided Antigenic Evolution Prediction (MLAEP), which combines structure modeling, multi-task learning, and genetic algorithm to model the viral fitness landscape and explore the antigenic evolution via *in silico* directed evolution. As demonstrated by existing SARS-COV-2 variants, MLEAP can infer the order of variants along antigenic evolutionary trajectories, which is also strongly correlated with their sampling time. The novel mutations predicted by MLEAP are also found in immunocompromised covid patients and newly emerging variants, like BA. 4/5. In sum, our approach enables profiling existing variants and forecasting prospective antigenic variants, thus may help guide the development of vaccines and increase preparedness against future variants.

Introduction

As the number of infection cases increased and the virus spread globally, novel mutations in the virus genome emerged[1-4]. At the time of April 2022, there are more than one million variants in the virus genome identified and uploaded to the Global Initiative on Sharing Avian Influenza Database (GISAID). The mutations often implicate the changes to the SARS-COV-2 properties[3]. Although most mutations decrease the virulence and transmissibility of the virus[5], some individual or combinatorial mutations substantially improve the transmissibility with enhanced cell entry efficacy[6], or ablate the neutralizing antibodies response elicited by infection or vaccine[1, 7], resulting in high-risk variants (HRVs). For example, the Alpha (B.1.1.7) variant of concern (VOC) spread worldwide through a higher human ACE2 binding affinity and transmissibility than the original Wuhan strain[8]. The Beta and the Gamma lineage abolished the neutralizing antibodies elicited by approved COVID-19 vaccines[9]. The Delta variant became a dominant strain worldwide with the increased transmissibility and morality[10, 11]. Recently, the heavily mutated Omicron variant caused new waves due to the extremely high rate of spread and the ability to evade the double-vaccinated person[12].

A substantial fraction of neutralizing antibodies, including monoclonal antibodies and those induced by the vaccines, target the spike receptor-binding domain (RBD)[13-15]. Antibodies targeting the RBD have been divided into four categories according to their binding epitopes[16]. Class 1 and class 2 antibodies bind the surface of the receptor-binding motif (RBM) and thus compete with ACE2 for RBD binding. Mutations in the RBM region, in turn, decreased neutralization by these antibodies. Class 3 antibodies bind the opposite side of the receptor-binding motif, contain less overlap with the ACE2-binding footprint, provide the potential for synergistic effects when combined with Class 1 and 2 antibodies for intercepting ACE2 binding[17]. Class 4 antibodies target a highly conserved region among sarbecoviruses and thus are generally more resistant to the variants[18]. However, the emerging viral lineages such as Omicron and BA.2 can still lead to a substantial loss of neutralization[19].

Understanding the role of the mutations and how they are linked to transmissibility and immune escape are thus of great importance. There have been an expanding set of analyses characterizing these problems[5, 18, 20-23]. Starr et al.[5] and Greaney et al.[18] performed deep mutational scanning (DMS) on the entire Spike RBD sequences of SARS-COV-2 on the yeast surface to determine the impact of single-position substitutions on the

binding ability to ACE2 and monoclonal antibodies. These assayed experiments provide a unique resource for understanding the properties of variants. However, the wet-lab experiments are resource and time-consuming, and cannot be scaled to the large protein sequence space. Maher et al.[23] characterized the potential risks of the single-position substitutions with a computational model and forecasted the driver mutations that may appear in emerging VOCs. Despite their effectiveness in modeling the risks at the single-mutant level, the newly emerging VOCs (e.g., Delta, Omicron) often possess multiple mutations in the RBD region, which directly influences the ACE2 binding and antibody escape. For example, the Omicron variant contains 15 mutations in the RBD region and obtains considerable antigenic escape ability[24]. Moreover, the effects of mutations are context-dependent, such that the epistatic interactions among the mutations limit the application scenario of the single-mutant-based methods[25].

The sequence space of protein variants grows exponentially when multiple mutations are considered, while measuring the functionality of the variant sequences far exceeding the capacity of wet-lab experiments. Machine learning methods have been proposed for solving the problem[26-28]. Alexander et al.[29] trained a large-scale transformer model with the self-supervised protein language modeling objective, while the model can infer the effects of mutations without supervision. Chloe et al.[30] combined linear regression with the Potts model, resulting in a data-efficient variant fitness inference model. These models have been proven to be effective in the protein engineering field for inferring the fitness landscape of proteins.

Inspired by these tools, Hie et al.[20] showed that language models trained on a set of evolutionarily related sequences are capable of predicting the potential risks of COVID variants with multiple mutations, and Karim et al.[22] further combined the language model score with structural modeling to monitor the risks of existing variants. These computational tools can work as HRV monitors and help us predict the risks of the emerging variants. However, as these methods focus on prediction and rely on existing data, they do not provide detailed views for ‘perspective’ variants and antigenic evolutionary potential. Joseph et al.[21] performed deep learning on the RBM sequences and built a predictive profile for the COVID variants in ACE2 binding and antibody escape for class 1, 2, and 3 antibodies. The proposed framework works quite well in finding prospective mutations, but they still have limitations: the mutations are found by brute-force search, so they only focused on a small subset of the RBD region, missed a large part of the Class 3 antibody epitopes and did not take the

class 4 antibodies into consideration.

In this work, we presented the MLAEP, built upon the existing data and approaches to forecast the combinatorial mutations in the entire RBD region that contains high antigenic evolutionary potential and may occur in the future. We hypothesized that under high immune pressure, the virus would tend to escape the antibody neuralization over a short-term time scale, and therefore the forecasting problem transforms into a search problem: starting from an initial sequence, it searches for a variant sequence within some edit distance range that has an improved antibody escape potential without losing much ACE2 binding ability. With the DMS datasets that directly measure the binding affinity of RBD variants towards ACE2 and eight antibodies from four classes, we built a multi-task deep learning model that could simultaneously predict the binding/escaping specificity of the variants towards the ACE2 and eight antibodies. After building the model, we used an *in vitro* pseudovirus neutralization test (pVNT) dataset[22] to validate the model performance in predicting the escape potential. Furthermore, we used existing variants with their sampling date from the GISAID database to validate our hypothesis: we found a surprisingly high correlation between our model scores and the variants' sampling time (Spearman r=0.65, p<1e-308). Next, with our model as the scoring function, we used the genetic algorithm[31, 32] to generate synthetic RBD variants with high ACE2 binding and antibody escape potential. Interestingly, the *in silico* directed evolution shares similar mutations with the adaptive evolution in immunocompromised COVID-19 patients[33-35] and newly emerging variants like BA.4/5. Further computational inference and docking experiments of top generated sequences reveal synthetic sequences with comparable or even better escaping potential than the Omicron variant, which may become the high-risk variants in the future.

Results

Overview of MLAEP

We first developed and trained a multi-task deep neural network model capable of predicting the variant RBD binding specificity towards the ACE2 and antibodies from four classes, as shown in Figure 1. The model receives two inputs: the variant RBD sequences and the ACE2/antibody 3D structures, and outputs the binding specificities of the two inputs. The model is then trained with a multi-task objective function to predict the

binding specificities of the variant sequences towards all targets simultaneously.

We fine-tuned the ESM-1b (evolutionary scale modeling) language model[29] for the sequence feature extraction. The model is pre-trained on ~27 million nature protein sequences in the UniRef50 database[36]. Fine-tuning the model has been proven to be effective for a broad range of downstream tasks, including biophysical properties prediction, structure prediction, and mutation effects prediction. With the ESM-1b model, the amino acid sequences are converted into a dense vector representation. For the ACE2/antibodies structures, we first transformed the 3D structures into graphs based on their contact maps and biophysical properties, then used the structured transformer[37] for the structural feature extraction. With the two models as feature extraction modules, we added nine parallel linear classification layers to learn the sequence to function mapping conditioned on the binding target structures (Fig. 1). As we have multiple binding targets for the variants, we used a hard-parameter sharing scheme to perform multi-task learning, where all modules share the same parameters across all nine tasks. Then, we trained the entire framework in an end-to-end manner. Finally, the model learns how to predict binding specificity for ACE2 and eight antibodies. Given an input RBD variant sequence, our model outputs nine scores corresponding to the ACE2 and eight antibodies. We defined the average of eight antibody scores as the predicted antibody escaping potential.

Our key hypothesis is based on the antigenic evolution: the future viral variants tend to have a higher antibody escaping potential without losing much ACE2 binding ability under high immune pressure. Thus, the antibody/ACE2 binding specificity learned by our model can be used to provide a meaningful direction in searching for novel variants that may cause future concern. Inspired by the progress in the machine learning-guided protein engineering field[26, 27], we used the trained multi-task model as the scoring function, took the average prediction scores from all nine tasks as the fitness score, and used a modified genetic algorithm for searching for novel variants with improved fitness (Fig. 1). The genetic algorithm is inspired by the process of natural selection, which iteratively evolves a group of candidates towards better fitness. The population of each iteration is called a generation. In each generation, the fitness of the candidate sequences is evaluated with the trained model. Then we filtered the populations by selecting the ones with higher fitness with higher probabilities for breeding the next generation. Random mutations and crossover are also introduced to better explore the search space. Genetic algorithm is known for performing well in solving

combinatorial optimization problems, thus fitting our needs in searching for novel variants. More details can be found in Methods.

The effectiveness of the multi-task learning model

MLAEP follows the machine learning-guided directed evolution paradigm, while the quality of generated sequences largely depends on the sequence to function model. First, we validated the generalization ability of the models to newly seen variants with cross-validation. We collected and cleaned nine deep mutational datasets containing 19132 variant sequences and their corresponding binding specificities towards ACE2 and eight antibodies from four functional classes. (Methods) We then compared a range of models specifically designed for protein engineering and accessed their classification performance in classifying the binders and non-binders (Methods, Extended Data Fig. 1) from the variant sequences, including the augmented Potts[30] model, the global UniRep[38] model, the eUniRep[26] model, and our model. We used 5-fold cross-validation to evaluate the performance of all models. The dataset is imbalanced regarding the number of positive and negative samples for all nine tasks. Thus, we reported the macro precision, macro recall, and macro F1 score to add more weights to the minor classes. Combined with the structure features, our model outperforms the other state-of-the-art (SOTA) methods in predicting the effects of mutations in all nine tasks (Fig. 2a, Supplementary Fig. 1). As a result, we focused on our model in the downstream analysis. We also performed ablation study for our model to show the importance of each module. We found that both the fine-tuning step and the structure representations improves the overall model performance (Extended Data Fig. 2).

To further validate the model predictions for the immune escape, we used the *in vitro* pseudovirus neutralization test (pVNT) assays datasets[22], which accessed the cross-neutralizing effect of BNT162b2-immune sera against vesicular stomatitis virus (VSV)-SARS-CoV-2-S pseudoviral bearing the Spike protein of selected variants of concern and reported the observed pVNT50 reduction rate of antibody response for each VOC. Higher reduction rates indicate more immune escape than the Wuhan wild type reference pseudovirus. Most VOCs pseudo viral show higher reduction rates compared with the wild type, especially for the Omicron pseudovirus, which is the most immune escaping variant with >97% reduction rate

of the pVNT50 compared with the wild type. Across all pseudovirus tested, there is a surprisingly high positive correlation (Fig. 2b, Pearson r= 0.76, p=0.001) between the predicted antibody escape potential and the pVNT50 reduction rate.

The Evo-velocity[39] enables the inference of evolutionary dynamics for proteins with a deep learning model. We next accessed our model's ability in inferring the evolutionary trajectory of the existing RBD sequences using the Evo-velocity. We used the existing SARS-COV-2 RBD sequences from the GISAID database across a timescale of around 27 months, from Dec. 2019 to Mar. 2022. The existing GISAID variant sequences were first transformed into embeddings with our multi-task model. On top of the embeddings, we assigned directions among them based on the changes in the average score predicted by our model, which forms the evolutionary "vector field". We visualized the embeddings in the two-dimensional space with the Uniform Manifold Approximation and Projection (UMAP)[40] (Methods). The variants of concern, including Alpha, Beta, Delta, and Omicron, were mapped into different clusters, and the velocities among these variants matched well with the known evolutionary trajectory (Fig. 3a). Despite the model being trained only with the RBD sequences, the pseudo time inferred with our model had a Spearman correlation of 0.55 (p<1e-308) with the known variant sampling time (Extended Data Fig. 3a, Supplementary Fig. 2). While using the ESM-1b (the Evo-velocity default setting) model, the score dropped to -0.38(p=1.05e-243) quickly (Extended Data Fig. 3b, c). We noted that a large set of mutations occur outside the RBD region; this may explain the weak correlation between the ESM-1b model pseudo time and the sampling time. Longer sequence length (e.g., using the entire Spike protein region) would lead to better performance for the ESM-1b model[39]. We attempted to explain our model's unique ability to infer pseudo time with only the RBD region. We explored the effectiveness of labels in our supervised learning, as it provides alternative directions rather than the language model preference[39, 41]. Interestingly, we found that the model prediction scores alone have an even higher Spearman correlation score of 0.65 (p<1e-308) with the sampling time (Fig. 3c) compared with that of the inferred pseudo time, while for the predicted antibody escape potential, the Spearman correlation is 0.67 (p<1e-308). These findings verify our assumptions: under the immune selection pressure, the virus evolves in the direction of immune escape, and our model can capture the antibody escape potential of the viral variants.

We next accessed the antigenic evolution on a short time scale by comparing the model predictions against the

sampling time (Fig. 3c, Extended Data Fig. 4). We evaluated three types of scores, the ACE2 binding score, the antibody escape potential, and the weighted average of the two scores. The predictiveness of the antibody escapes score increases from nearly noninformative early in the pandemic to a stronger correlation during the Omicron wave. It also gains predictiveness with the emergence and spread of Alpha variants in Early 2021 but subsequently loses the predictiveness along with the emergence of other variants. We noted that the antigenic evolution For the ACE2 binding probability score, it tends to become more informative during the first year, while soon it becomes non-informative when the new VOC like Delta and Omicron emerged. These results suggests that the antigenic evolution happens along with the infection waves.

We then examined the model sequence representations against the binding specificities. We found that after the training, there are strong correlations between embeddings' primary and secondary axis of variation and the binding specificities for all nine targets (Fig. 3d, Supplementary Fig. 3). The correlations are observed for both ACE2 binding and antibody escape, suggesting that our multi-task learning strategies enable the model to learn the functional properties simultaneously. Given that the variant sequence embeddings are shared across tasks, this suggests that our model split the sequences based on an antigenic meaningful sense of binding preference.

In summary, our model effectively infers the immune escape potential and the ACE2 binding specificity, while the predicted scores correlate positively with the real-world sampling time, especially for the newly emerging Omicron wave. Taken together, we hypothesize that our model can work as a good scoring function for searching for high-risk mutations and the corresponding variants.

In silico **directed evolution as a predictive tool**

With our model as the scoring function, we used the genetic algorithm to search for novel RBD variant sequences with high antigenic evolutionary potential. The search process consists of selecting an initial sequence from the GISAID database, generating and selecting "better-than-initial" sequences with the genetic algorithm to produce 38870 putatively high-risk variants within a 15-mutations "trust radius" of the initial sequence (Methods). We performed the search process for the sequences in the GISAID database from January 1, 2022 to March 8, 2022, yielding a total of 971 distinct sequences. We then visualized the generated sequences together with the existing sequences using the distance-preserving multidimensional scaling plot[42]

(Fig. 4a). While the sequences from the deep mutational scanning experiments only occupy a small region around the wild type sequences, the prevalent variants (e.g., Omicron) locate in different regions, far from the wild type. The sequences searched with our model shown are diverse, largely expanding the sequence space.

Compared with the seed sequences, the synthetic sequences generated by our model include key mutations for ACE2 binding and antibody escape. To visualize the difference and further explore the patterns of the generated mutations, we constructed the position frequency matrix (PFM) for the two sequence sets and calculated the Kullback-Leibler divergence (KL divergence) for each position based on the two PFMs (Methods). Fig. 4b and Supplementary Fig. 4 provides structure-based visualizations and projects the Kullback-Leibler divergence per site onto a crystal structure of the RBD (PDB id: 6m0j). As an alternative representation, Fig. 4c provides a probability-weighted Kullback-Leibler logo plot[43] for the top 50 most divergence sites, where the total height of the letters depicts the KL divergence of the site, while the size of the letters is proportional to the relative log-odds score and observed probability (see Methods). The logo plot for all positions can be found in Extended Data Fig. 5. Enriched amino acids locate at the positive side of the y-axis and depleted amino acids locate at the negative side.

The logo plot shows that the mutations searched by our model largely overlap with the antibody escape maps. For example, Y453, F456, and A475 are key sites for class 1 antibody escape[18], while they are also present in many synthetic variant sequences. Mutations escaped class 2 antibodies at sites E484, F490, and P491[18]. The logo plot shows that these sites ranked high as “active sites”. Class 3 antibodies, which bind the opposite side of the receptor-binding motif, tend to be escaped by sites like N437, N448, and Q498[18], which are also vulnerable sites suggested by the model. Class 4 antibodies bind to a conserved motif among the sarbecorviues, far away from the RBM. Our model still captures the conservation and assigns mutations to the motif. However, some sites with a large KL divergence do not locate in the epitope regions. This has several explanations. Firstly, it is clear that some top sites (e.g., L368, C480) are not the direct binding sites but the proximal contact sites in the structures, which may influence the binding as well. Secondly, as there are epistasis relationships among the mutations, some combinatorial mutations may influence the RBD function nonlinearly and then modify the antibody escape, which is not directly revealed by the epitope map. Moreover, these non-epitope sites with high KL divergence need to be taken into consideration as they may perform an

important role in future variants. Another concern is that some sites in the epitope region have a low KL-divergence, one possible explanation is that these sites have no tolerate mutations, for example, G416 and R457. Another explanation is that some mutations at antibody-contact sites do not directly influence antibody binding.

The synthetic variant sequences share similar mutations with the chronic SARS-COV-2 infections. A reverse mutation, R493Q, for example, was found in a persistently infected, immunocompromised individual[44]. Other mutations found by our model, like E340K[45], E484T[33], G485R[46], and F490L/E484G[47], are also found in immunocompromised patients treated with monoclonal antibodies. Moreover, the unique mutations found in the emerging variants, BA.4/5, the L452R, F486V, and the reverse mutation R493Q, are also captured by our model. (Extended Data Fig. 5, Supplementary Fig. 5) This suggests that our model could be used for finding novel mutations that may occur naturally. A detailed list of the found mutations in compromised patients is available in Supp Table 1.

In silico estimation of variant sequence risk

Having generated the synthetic sequences and their mutational sites, we next evaluated the risk of the variant sequences with Evo-velocity analysis and viral language model risk inference, followed by structure modeling and antibody-antigen docking.

We first visualized the synthetic sequences together with the existing GISAID variant RBD sequences with the Evo-velocity figure (Fig. 5a, b, Supplementary Fig. 6). As the initial sequences for the generation mainly come from the Omicron lineage, our synthetic sequences and Omicron variants locate in a different cluster compared to the wild type and other VOCs. The streamlines suggest that the B.1.1.529 and the BA.2 may develop into different sublineages. Considering that the inferred pseudotime and trajectory have a high correlation on the existing GISAID variant sequences, the newly generated synthetic sequences, especially the ones located at the directional flow after B.1.1.529 and BA.2, need to be monitored as they have a high chance of becoming dominate variants in the coming wave.

The language models trained with the spike protein sequences with a self-supervised masked language modeling objective have been used to assess the risk of a viral variant[20]. The risk is composed of two proxies,

namely grammaticality as a fitness measurement and semantic for antigenic variation. We validated the synthetic and GISAID sequences in terms of the two dimensions. Compared with the existing covid variant sequences, the variants searched by our model have a larger semantic score, suggesting a high immune escape potential (Fig. 5c, Extended Data Fig. 6, 7, Supp Table 3). While for the grammaticality dimension, the synthetic variants have lower variance and are generally high. Taken the two dimensions together, the synthetic variants are more likely to induce immune escape.

We further performed *in silico* antibody-antigen docking experiments to access the immune escape, which includes four antibodies that correspond to the four antibody epitope classes. All of them have been proven to be effective in neutralizing coronavirus[48-51]. We picked 20 synthetic variants RBD sequences with the highest average model prediction score and used the Swiss-model protein structure homology modeling server[52] to get the variants' structure. To calculate the binding score, we used the difference of estimated binding energy between the bound complex and the sum of unbounded isolates. We performed 1000 independent runs with the SnugDock[53] for each antibody-antigen pair (Fig. 5d, Supplementary Fig. 7, Supp Table 4) and used the mean value as a proxy for binding score (Fig. 5e, Methods). We also performed similar experiments for Wuhan-wildtype and the Omicron variants for comparison. The synthetic variants perform generally higher binding scores compared with that of the wild type, while compared with the Omicron variant, the synthetic sequences perform as well as (or even higher immune escape potential than) Omicron. The simulated docking experiments also suggest that the generated variants have a high probability of becoming immune escape.

Discussion

In this paper, we developed a machine learning-guided directed evolution paradigm for accessing the antigenic evolution of SARS-COV-2. We first trained a multi-task deep learning model for predicting ACE2/antibody binding specificity with the variant sequences and binding target structures. Through *in vitro* validation, we showed that the model can predict the antigenic evolutionary potential as the result of high immune pressure. Combined with the genetic algorithm, we performed *in silico* directed evolution using the model scores. The generated synthetic sequences contain high immune escape potential, which is further validated with the viral language model score and *in silico* docking experiments. Top mutational sites in the generated sequences

largely overlap with the epitopes. Moreover, MLAEP captures mutations that also happen in chronic SARS-COV-2 infections and emerging variants like BA.4/5. While we used the genetic algorithm to search for novel variants, other search algorithms like hill-climbing[54], simulated annealing[55], and reinforcement learning[56] could also be combined with our model. The multi-task learning model could be also replaced with other mutation effects prediction models[30].

Deep learning models can learn high-order epistasis relationships among the multiple mutations[28, 29]. Our multi-task model, meanwhile, can capture such relationships and work as a monitor for predicting the escaping potential of newly emerging variants, especially for those heavily mutated variants, like Omicron lineage. Combined with the Evo-velocity analysis, our model helps to reveal the evolution trajectory of existing sequences and enables the discovery of high-risk variants that may appear in the future. Although MLAEP is built upon *in vitro* deep mutagenesis experimental data with *in silico* generation method, we get mutational patterns similar to that of the adaptive evolution in immunocompromised patients and newly emerging variants like BA. 4/5. The results suggest that the *in silico* directed evolution can lead to the prediction of *in vivo* virus evolution. Consequently, MLAEP may enable the support of public health decision-making and guide the development of new vaccines. Besides, our approach could be applied to the rapidly evolving virus and other potential outbreaks, such as antibiotic resistance[57].

One limitation of our model is that we only focused on the RBD sequences, while many mutations occur outside the region. We noted that the mutagenesis assayed data provides semantically meaningful directions for finding “better-than-natural” sequences. An increasing number of experiments characterize the functionality of mutations in other regions, and we will explore these datasets in the future. Another concern is that we only optimized two targets, the ACE2 binding and antibodies escape, while the directionality of evolution is also driven by many other properties, like the epidemiology features and T cell responses. In the future, we will utilize these datasets and incorporate more *in vivo* and *in vitro* experimental data. Specifically, we will combine the *in vivo* antibody-antigen co-evolution data from patients and the assessment of other immune responses to better understand and predict the evolution of SARS-COV-2.

Method

Dataset

We collected and cleaned nine deep mutational scanning datasets, which measure the binding affinity of the SARS-COV-2 RBD variants towards the ACE2 and eight antibodies from four classes. We built a dataset consisting of 19132 RBD sequences, where each sequence has nine labels, corresponding to their binding ability to the nine targets. Most sequences have one or two mutations compared to the Wuhan-wildtype RBD sequence. However, considering the possible batch effect and the physical meaning differences among the measured scores, we normalized each score independently by transforming the continuous variables into semantically meaningful binary labels. For the ACE2 task, we directly compared the binding score of the mutated sequences to the wild type and set the label to “enhanced binding” if the score is larger than the wild type and vice versa. There is no information about the wild type for the eight antibodies tasks, so we cannot set the threshold as described earlier. Instead, we found that the distributions of the binding score's logarithm clearly show two clusters; therefore, for all antibody datasets, we took it as a mixture-of-Gaussian model respectively, defining the one with smaller binding scores as escaped and vice versa. This preprocessing step is consistent with the subsequent work[58].

In summary, there were 1540 (8%) mutated RBD sequences identified as enhanced binding to ACE2, 3482 (18%) mutated RBD sequences identified as escaped to COV2-2096, 1220 (6%) mutated RBD sequences identified as escaped to COV2-2832, 2000 (10%) mutated RBD sequences identified as escaped to COV2-2094, 1473 (8%) mutated RBD sequences identified as escaped to COV2-2050, 1859 (10%) mutated RBD sequences identified as escaped to COV2-2677, 929 (5%) mutated RBD sequences identified as escaped to COV2-2479, 780 (4%) mutated RBD sequences identified as escaped to COV2-2165 and 3347 (17%) mutated RBD sequences identified as escaped to COV2-2499.

We used the pseudovirus neutralization test assay data from Karim et al.[22] to validate our model performance. The dataset measures the immune escaping of 19 HRVs pseudovirus by comparing the observed reduction of neutralizing antibody response against the wildtype pseudovirus. As we were only concerned about the RBD variants, we chose 15 out of 19 RBD sequences of the assay for visualization. The detailed information about chosen sequences and the corresponding reduction rate is available in Supp Table 2.

Overview of the multi-task model

A central feature of SARS-COV-2 is antigenic evolution, that is, under high immune pressure, the newly emerging variants will tend to escape the antibody while do not lose much binding ability to the ACE2.

To accomplish the goal of predicting the antigenic evolution, we need to construct the virtual fitness landscape of the antigenic regions, especially for the RBD protein. We aimed to infer the fitness landscape of the RBD by learning the effects of mutations on ACE2 binding and antibody escape. Specifically, given RBD variant sequences and their labels, together with the binding partner (ACE2/antibody) structures, our model learned the nonlinear mapping function f that can simultaneously predict the binding specificity for ACE2 and antibody. The function f is parameterized by learnable mapping parameters θ composed of three modules: the sequence feature extractor $\mathcal{S}$, the structure feature extractor $\mathcal{G}$ and the sets of nine classification heads $H = \{\mathcal{H}_c\}_{c=1}^{9}$, where all $\mathcal{H}_c$ share the same group of parameters. All three modules are neural networks. The parameters of the three modules are optimized in an end-to-end manner.

1. Sequence feature extractor. The sequence feature extractor takes as input of amino acid sequences of RBD variants $x = (x_1, x_2, \ldots, x_l)$ of length L, where L denotes the length of the RBD sequence and elements x_i belongs to $A = \{\text{all amino acids}\}$. Input is mapped to a dense representation vector (sequence representation). The backbone of the sequences feature extractor is the ESM-1b transformer, which is pretrained on UniRef50 representative sequences with the masked language modeling objective. We chose the ESM-1b as the sequence feature extractor because it outperforms other baselines on a range of downstream tasks. The pretrained weights were used for initializing the neural network, and we fine-tuned the model parameters during training.

2. Representing structure as graph. We first represented the 3D structure as a k-nearest neighbor graph $g = (V, E)$ with the node set $V = \{v_j\}_{j=1}^{N}$ of size N, where each element v_j denotes for the features of representative atoms (we chose N, C, and O atom in the experiment) in the protein

3D structure, N denotes the total number of atoms. For each atom, we got its two nearest neighbors with the following constraints: the ones with the same atom type but belong to different amino acids. We then measured the dihedral angles of the atom and its neighbors to as node features. The edge features $E = \{e_{ij}\}_{i \neq j}$ describes the relationship between the nodes, including the relative distance, direction, and orientation between the two nodes in the three-dimensional space. We set k as 30.

3. Structure feature extractor. The structure feature extractor tasks as input of graphs $g = (V, E)$ describing the spatial feature of the protein structure. The transformed graph is further mapped into a dense representation vector (structure representation). The backbone of the structure feature extractor is a Structured Transformer[37], where the attention for each node is restricted to its k-nearest neighbors in 3D space. We chose the Structured Transformer for the structure feature extraction as it is computationally efficient and performs well in the protein design task. The structure representation works as conditional tags in our multi-task learning.

4. Classification heads. After getting the sequence representation and the structure representation, we concatenated the two vectors into the joint representation, and fed it into the classification heads. The classification heads map the joint representation to the labels. We used nine parallel classification heads for the nine classification tasks, while the neural network parameters are shared. During training, the sequence feature extractor, the structure feature extractor and the classification heads are trained in an end-to-end manner to minimize the average classification loss among the nine tasks.

5. Loss function. Let $\boldsymbol{x} = \{x_i\}_{i=1}^{N}$ be the set of RBD variant amino acid sequences, and $\boldsymbol{y} = \{y_i\}_{i=1}^{N}$ be the set of labels of all sequences, and $\boldsymbol{y} = \{y_i\}_{i=1}^{N}$ denotes for the set of M labels of the i-th RBD variant. Furthermore, let $G = \{g_c(V, E)\}_{c=1}^{M}$consists of M graphs derived from the ACE/antibody structures. We seek to learn a joint embedding for all downstream

classification tasks to better model the fitness landscape of RBD. Therefore, the sequence feature extractor and the structure feature extractor are shared among all tasks. Considering that all the tasks are imbalanced in terms of the positive and negative samples, we added a rescaling weight p_c to all tasks and optimized the following loss function:

$$L = \frac{1}{MN}\sum_{c=1}^{M}\sum_{i=1}^{N} -[p_c y_i^c \cdot \log \sigma\left(\mathcal{H}_c\left(\mathcal{S}(x_i); \mathcal{G}(g_c)\right)\right) + (1 - y_i^c) \cdot \log(1 - \sigma\left(\mathcal{H}_c\left(\mathcal{S}(x_i); \mathcal{G}(g_c)\right)\right))]$$

Where p_c equals to the number of positive samples divided by the number of negative samples, M equals to nine, σ is the sigmoid function. The equation measures the binary cross entropy between the targets and predicted probabilities.

Architecture and hyperparameters

The architecture of the sequences feature extractor is based on the ESM-1b transformer, which consists of 34 layers, we used the outputs of the 33rd layer as the sequence feature representations. For the structured transformer, we only kept the transformer encoder, and used three layers of self-attention and position-wise feedforward modules with a hidden dimension of 128. Finally, we got a 1280-d vector for each sequence as the sequence representation and a 1300-d vector for each 3D structure as the structure representation. For each classification head, we used 1024 neurons in the first layer and two neurons in the second layer. The RELU function is used between the layers as nonlinear activations. We also passed a dropout rate p=0.1 to prevent overfitting. We trained the entire model with AdamW as the optimizer and used a linear schedule with warmup to adjust the learning rate. We set the batch size as 16 and gradient accumulation steps as 10, which means the total train batch size is 160, and the validation is the same. The model was trained for 9500 updates with the initial learning rate of 1e-5 and warmup steps 120, during which the model with the best marco F1 score among all the tasks was kept. The hyperparameters described above were decided through several trials of

experiments and selected the one with the best performance.

Choice of baselines

Our framework follows the machine learning-guided directed evolution paradigm, while the quality of generated sequences largely depends on the sequence-function model. Thus, we validated the generalization ability of the models to newly seen variants with cross-validation. Our multi-task deep learning model was designed as a supervised technique to infer the fitness of variant sequences, while there is no existing method designed for multi-task learning or multi-label learning on this task. So instead, we evaluated the performance of the existing methods separately for all nine tasks. We chose the state-of-the-art supervised-learning-based methods for inferring the effect of mutations, including the augmented Potts model[30], the eUniRep model[26] and the gUniRep model[38].

The augmented Potts model combines the evolutionary information with the one-hot encoded amino acid sequences as input features and trains a linear regression model on top of the features. It outperforms most existing methods in inferring the effects of mutations. We first generated the multiple sequence alignments profile of RBD using the profile HMM homology search tool Jackhmmer. We set the bit score threshold as 0.5 and the number of iterations to 1. We then calculated the evolutionary Potts potential of the RBD variant sequences using the plmc. We replaced the Ridge regression with a Logistic regression head for the classification objective while keeping the rest procedures the same as the original settings.

The gUniRep model was trained on 24 million UniRef50 amino acid sequences with the next amino-acid prediction objective, and the representations extracted from the pretrained model acts as a featurization of the sequences, benefits the downstream protein informatics tasks. With the RBD variant sequences as input, we got the fixed-length vector representations from the pretrained model as sequence embeddings. We added a Logistic regression head for downstream classification.

The eUniRep model was built on top of the gUniRep model. An unsupervised fine-tuning step with sequences related to the target protein (evotuning) was introduced to learn the distinct features of the target family. Previous *in vitro* studies on the GFP and beta-lactamase proved its effectiveness for efficiently modeling the protein fitness landscape. We performed evotuning with the same MSA profile we generated in the augmented

Potts model. After that, we characterized the sequence embeddings with the eUniRep model and train a Logistic regression model for downstream classification. All methods were trained and tested on the same training data and validation data for all five folds.

Ablation study

We performed ablation studies to show the effectiveness of each module. We first explored the effectiveness of the graphical representations. We replaced the structure features with the random Uniform noise $X \sim U(0,1)$ and performed the multi-task learning with the same training details and procedures. Besides, we also explored the importance of fine tuning by freezing the parameters of the ESM-1b model and only optimized the parameters of the classification head.

Performance evaluation

We evaluated the multi-task learning model with the 5-fold cross-validation. We randomly split the dataset into five folds. Each time, we used four folds as the training data and held out the remaining fold for validation. We used the Accuracy, Precision, Recall, and F1 score to evaluate the classification performance across all models. As all the nine classification tasks are imbalanced, we used the Marco-precision, Marco-recall, and Macro-F1-score to get an unbiased evaluation. True positive (TP), true negatives (TN), false positives (FP), and false negatives (FN) were measured by comparison between the prediction results produced by the model and the ground truth in the validation set.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall_c = \frac{TP}{TP + FN}, \quad Recall_{macro} = \frac{Recall_1 + Recall_2}{2}$$

$$Precision_c = \frac{TP}{TP + FP}, \quad Precision_{macro} = \frac{Precision_1 + Precision_2}{2}$$

$$F1_{macro} = 2 \times \frac{Precision_{macro} \times Recall_{macro}}{Precision_{macro} + Recall_{macro}}$$

Generate virtual RBD variant sequences with the genetic algorithm

To forecast the variants that follows the antigenic evolutionary potential, we applied the genetic algorithm for searching the peaks of the fitness landscape described by our model. Inspired by Darwin's theory of natural evolution, the genetic algorithm mimics the evolutionary process in the genome, where mutations, crossover and selection happen, letting candidate solutions of a population with higher fitness scores have a higher probability of surviving and producing the next generation of offspring. For SARS-CoV-2, it has been proven that similar progress happens in immunocompromised infected patients who got treated with the monocle antibodies[33]. Hence, we used the genetic algorithm to model the antigenic evolution process and search for the potential risky variants that might appear in the future. The genetic algorithm we used consists of the following steps:

1. Selection of initial sequences. The initial input sequences were obtained from the March 8, 2022 GISAID release[9]. We selected the RBD sequences from the recent two months (i.e. from January 1, 2022 to March 8, 2022), resulting in 957 distinct RBD sequences. For each sequence, we created a generation P_0 of size S by perturbing the sequences $\boldsymbol{S}$ times to generate a set of distinct modifications to the original sequence.

2. Perturb operation. For a given sequence $X = (x_1, x_2, \dots, x_n)$, we first randomly selected an amino acid x_i, and got the K nearest neighbors of the selected amino acid according to the BLUSUM62 matrix. Secondly, we computed the fitness value when x_i is replaced with its neighbors, while keeping the remaining set of words unchanged. We then picked the mutation with a probability proportional to its fitness value. Finally, the selected mutated amino acid replaced the original one, we got a new sequence. We set K=20 in our experiments.

3. Estimation of the fitness. The fitness score is defined as the average value of target label prediction probabilities for all nine tasks. For ACE2 binding task, the target label is binding,

while for the antibody task, the target label is escape. The probabilities were found by querying the trained multi-task model.

4. Crossover. After getting the perturbed population and the fitness values for each individual sequence, we performed the crossover operation. Pairs of sequences are randomly selected with the probabilities proportional to its fitness value. A child sequence is then generated by independently sampling from the two parents. The newly generated sequences form the new generation. If the fitness value of a population member in the generation is higher than the high-risk threshold, the optimization is done. Otherwise, the perturbation, selection and crossover operation will be applied on the new generation.

We performed *in silico* evolution for each initial sequence from GISAID subset for 100 times independently, and finally got 38870 unique RBD variant sequences.

Evo-velocity analysis

The Evo-velocity analysis follows the study of Hie et al[39]. They used the pretrained protein language model (e.g. ESM-1b) to predict the local evolution within protein families and used a dynamic "vector field" to visualize it. We first computed the embeddings for each sequence with the ESM-1b model, and then constructed the K-nearest-neighbor graph based on the embeddings, in which node represents the sequences and edges connected similar sequences. Further, the edges were assigned with directions based on the language model pseudolikelihoods, with flow-in node meaning evolutionarily favorable. Here, we performed the Evo-velocity analysis with their settings and ours. In our setting, we used the joint embeddings extracted from the fine-tuned protein language model and the Structured Transformer model to represent the sequences and set the direction of the edge by comparing the average predicted score among the nine tasks, where vertex with a large value is defined as the tail. We collected 7594 unique RBD sequences from the March 8, 2022 GISAID release. The date of the sequence is defined as the first reported date. After constructing the directed KNN neighborhood graph, we further performed network diffusion analysis to infer the pseudo time. We manually set the root as the Wuhan-wildtype RBD sequences.

Visualization

We visualized the model embeddings using UMAP. The K-nearest-neighbor network was built with the k set to 30, while the resolution is set to 1. We calculated the KNN graph and performed UMAP for both the GISAID sequences and the generated sequences. We further projected the predicted KL-divergence maps onto a RBD structure (PDB id: 6m0j) and visualized the structure with PyMol. We collected the binding epitopes for class 1, 2, and 3 antibodies from Greaney et al.[18] , while for the class 4 antibodies, we used the contact sites of antibody CR3022 to represent the class 4 binding epitope.

We used probability weighted Kullback-Leibler Logo plot for visualizing the generated mutations. Let $M_1 = (\boldsymbol{f_1}, \boldsymbol{f_2}, \boldsymbol{f_3}, \dots, \boldsymbol{f_n})$ denote the position frequency matrix (PFM) of the initial sequences, where the length of the initial sequences are n and each $\boldsymbol{f_i} = (a_1, a_2, \dots, a_{20})^T$, represents the frequency of each amino acid at position i. Further, let $\boldsymbol{M_2} = (\boldsymbol{f'_1}, \boldsymbol{f'_2}, \dots, \boldsymbol{f'_n})$ denotes for the PFM of the generated sequences, each $\boldsymbol{f'_i} = (a'_1, a'_2, \dots, a'_{20})^T$. We computed the KL divergence for each position:

$$D_{KL}(\boldsymbol{f'_i} || \boldsymbol{f_i}) = \sum_{i=1}^{20} a'_i \cdot \ln (a'_i / a_i)$$

The KL divergence denotes for the total heights at each position in the logo plot. We further set the height and the direction of a letter with a probability weighted normalization[43], where the relative height of each individual amino acid is proportional to $a'_i \cdot \ln (a'_i / a_i)$:

$$h(a'_i) = \frac{a'_i \cdot \ln (a'_i / a_i)}{\sum_{i=1}^{20} a'_i \cdot |\ln (a'_i / a_i)|} D_{KL}(\boldsymbol{f'_i} || \boldsymbol{f_i})$$

Antibody antigen docking

We collected the crystal structures of representative and clinical effective antibodies of four classes from PDB database: LY-CoV16 (class1, PDB id: 7c01), LY-CoV555 (class2, PDB id: 7kMG), S309 (class3, PDB id: 7R6W) and CR3022 (class4, PDB id: 6w41). The top 20 risky RBD sequences that our model generated were selected and modeled by the protein structure homology-modeling server SWISS-MODEL. Besides, we also

collected the RBD structure of the Wuhan-wildtype and Omicron variant. After that, the *in silico* docking simulation between RBD structures and antibodies was implemented with the Rosetta antibody-antigen docking protocols. We applied the all-atom relax protocol, docking prepack protocol, and antibody-antigen docking simulation using the SnugDock with the complex structures of the combination of 22 RBD structures and four antibodies. We ran the docking 1000 times independently for each antibody-antigen pair and got 1000 Rosetta interface scores as binding scores. The Rosetta interface score (I_{sc}) is defined as

$$I_{sc} = E_{bound} - E_{unbound}$$

where E_{bound} is the score of the bound complex, while the $E_{unbound}$ is the sum of the scores of individual docking partners. We further filtered out the scores with $I_{sc} > 0$ for getting reasonable scores. The statistical significance was tested using the student's *t-test*.

Data availability

The deep mutational scanning datasets is publicly available at https://jbloomlab.github.io/SARS-CoV-2-RBD_DMS/ and https://github.com/jbloomlab/SARS-CoV-2-RBD_MAP_Crowe_antibodies . The pseudovirus neutralization test assay data is publicly available in its original publications. The GISAID data is publicly available at https://www.gisaid.org/. We retrieved data from the website on 8 March 2022.

Acknowledgment

We acknowledge all members in the Structural and Functional Bioinformatics (SFB) group at KAUST for the fruitful discussions.

Funding

King Abdullah University of Science and Technology (KAUST). [BAS/1/1624-01, FCC/1/1976-23-01,

FCC/1/1976-26-01, REI/1/0018-01-01, REI/1/4216-01-01, REI/1/4437-01-01, REI/1/4473-01-01, URF/1/4098-01-01, REI/1/4742-01-01]

Author contributions

W.H., C.N., and X.G. conceptualized the study and developed the methodology. W.H. and C.N. implemented models and analyzed data. X.G. supervised the research and the entire project. All authors wrote the paper.

Reference

1. Wang, P. et al. Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7. *Nature* **593**, 130-135 (2021).
2. Wang, P. et al. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. *Cell Host & Microbe* **29**, 747-751.e744 (2021).
3. McCormick, K.D., Jacobs, J.L. & Mellors, J.W. The emerging plasticity of SARS-CoV-2. *Science (New York, N.Y.)* **371**, 1306-1308 (2021).
4. Iketani, S. et al. Antibody evasion properties of SARS-CoV-2 Omicron sublineages. *Nature* **604**, 553-556 (2022).
5. Starr, T.N. et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. **182**, 1295-1310. e1220 (2020).
6. Ozono, S. et al. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. **12**, 1-9 (2021).
7. Jangra, S. et al. SARS-CoV-2 spike E484K mutation reduces antibody neutralisation. **2**, e283-e284 (2021).
8. Volz, E. et al. Assessing transmissibility of SARS-CoV-2 lineage B. 1.1. 7 in England. **593**, 266-269 (2021).
9. Charmet, T. et al. Impact of original, B.1.1.7, and B.1.351/P.1 SARS-CoV-2 lineages on vaccine effectiveness of two doses of COVID-19 mRNA vaccines: Results from a nationwide case-control study in France. *The Lancet Regional Health - Europe* **8**, 100171 (2021).
10. Mlcochova, P. et al. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **599**, 114-119 (2021).
11. Ong, S.W.X. et al. Clinical and virological features of SARS-CoV-2 variants of concern: a retrospective cohort study comparing B.1.1.7 (Alpha), B.1.315 (Beta), and B.1.617.2 (Delta). *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* (2021).
12. Andrews, N. et al. Covid-19 vaccine effectiveness against the Omicron (B. 1.1. 529) variant. (2022).
13. Piccoli, L. et al. Mapping neutralizing and immunodominant sites on the SARS-CoV-2 spike receptor-binding domain by structure-guided high-resolution serology. **183**, 1024-1042. e1021 (2020).
14. Dejnirattisai, W. et al. The antigenic anatomy of SARS-CoV-2 receptor binding domain. **184**, 2183-2200. e2122 (2021).
15. Zost, S.J. et al. Potently neutralizing and protective human antibodies against SARS-CoV-2. **584**, 443-449 (2020).
16. Barnes, C.O. et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. **588**, 682-687 (2020).

17. Shrestha, L.B., Tedla, N. & Bull, R.A. Broadly-Neutralizing Antibodies Against Emerging SARS-CoV-2 Variants. *Front Immunol* **12**, 752003-752003 (2021).
18. 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).
19. Cao, Y. et al. Omicron escapes the majority of existing SARS-CoV-2 neutralizing antibodies. *Nature* **602**, 657-663 (2022).
20. Hie, B., Zhong, E.D., Berger, B. & Bryson, B.J.S. Learning the language of viral evolution and escape. **371**, 284-288 (2021).
21. Taft, J.M. et al. Predictive profiling of SARS-CoV-2 variants by deep mutational learning. (2021).
22. Beguir, K. et al. Early Computational Detection of Potential High Risk SARS-CoV-2 Variants. (2021).
23. Maher, M.C. et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. eabk3445 (2021).
24. Markov, P.V., Katzourakis, A. & Stilianakis, N.I. Antigenic evolution will lead to new SARS-CoV-2 variants with unpredictable severity. *Nature Reviews Microbiology* **20**, 251-252 (2022).
25. Starr, T.N. et al. Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. (2022).
26. Biswas, S., Khimulya, G., Alley, E.C., Esvelt, K.M. & Church, G.M.J.N.m. Low-N protein engineering with data-efficient deep learning. **18**, 389-396 (2021).
27. Yang, K.K., Wu, Z. & Arnold, F.H.J.N.m. Machine-learning-guided directed evolution for protein engineering. **16**, 687-694 (2019).
28. Riesselman, A.J., Ingraham, J.B. & Marks, D.S. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15**, 816-822 (2018).
29. Rives, A. et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. **118**, e2016239118 (2021).
30. Hsu, C., Nisonoff, H., Fannjiang, C. & Listgarten, J.J.N.b. Learning protein fitness models from evolutionary and assay-labeled data. 1-9 (2022).
31. Whitley, D.J.S. & computing A genetic algorithm tutorial. **4**, 65-85 (1994).
32. Alzantot, M. et al. Generating natural language adversarial examples. (2018).
33. Halfmann, P. et al. Evolution of a globally unique SARS-CoV-2 Spike E484T monoclonal antibody escape mutation in a persistently infected, immunocompromised individual. (2022).
34. Harari, S. et al. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. (2022).
35. Wilkinson, S.A. et al. Recurrent SARS-CoV-2 mutations in immunodeficient patients. (2022).
36. UniProt: the universal protein knowledgebase in 2021 %J Nucleic acids research. **49**, D480-D489 (2021).
37. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T.J.A.i.N.I.P.S. Generative models for graph-based protein design. **32** (2019).
38. Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G.M.J.N.m. Unified rational protein engineering with sequence-based deep representation learning. **16**, 1315-1322 (2019).
39. 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.e276 (2022).
40. McInnes, L., Healy, J. & Melville, J.J.a.p.a. Umap: Uniform manifold approximation and projection for dimension reduction. (2018).
41. Sandhu, M., Spence, M.A. & Jackson, C.J. Evo-velocity: Protein language modeling accelerates the study of evolution. *Cell Systems* **13**, 271-273 (2022).
42. Cox, M.A. & Cox, T.F. in Handbook of data visualization 315-347 (Springer, 2008).

43. Thomsen, M.C.F. & Nielsen, M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. *Nucleic Acids Research* **40**, W281-W287 (2012).
44. Scherer, E.M. et al. SARS-CoV-2 evolution and immune escape in immunocompromised patients treated with exogenous antibodies. 2022.2004.2012.22273675 (2022).
45. Destras, G., Bal, A., Simon, B., Lina, B. & Josset, L. Sotrovimab drives SARS-CoV-2 Omicron variant evolution in immunocompromised patients. 2022.2004.2008.22273513 (2022).
46. Harari, S. et al. Drivers of adaptive evolution during chronic SARS-CoV-2 infections. 2022.2002.2017.22270829 (2022).
47. Wilkinson, S.A. et al. Recurrent SARS-CoV-2 Mutations in Immunodeficient Patients. 2022.2003.2002.22271697 (2022).
48. Shi, R. et al. A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **584**, 120-124 (2020).
49. Jones, B.E. et al. The neutralizing antibody, LY-CoV555, protects against SARS-CoV-2 infection in nonhuman primates. **13**, eabf1906 (2021).
50. Pinto, D. et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **583**, 290-295 (2020).
51. Yuan, M. et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. **368**, 630-633 (2020).
52. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. **46**, W296-W303 (2018).
53. Weitzner, B.D. et al. Modeling and docking of antibody structures with Rosetta. **12**, 401-416 (2017).
54. Davis, L. in Proc. Intl. Conf. Genetic Algorithm, 1991 18-23 (1991).
55. Van Laarhoven, P.J. & Aarts, E.H. in Simulated annealing: Theory and applications 7-15 (Springer, 1987).
56. Nareyek, A. in Metaheuristics: Computer decision-making 523-544 (Springer, 2003).
57. Li, Y. et al. Hmd-arg: hierarchical multi-task deep learning for annotating antibiotic resistance genes. **9**, 1-12 (2021).
58. Frazer, J. et al. Disease variant prediction with deep generative models of evolutionary data. **599**, 91-95 (2021).

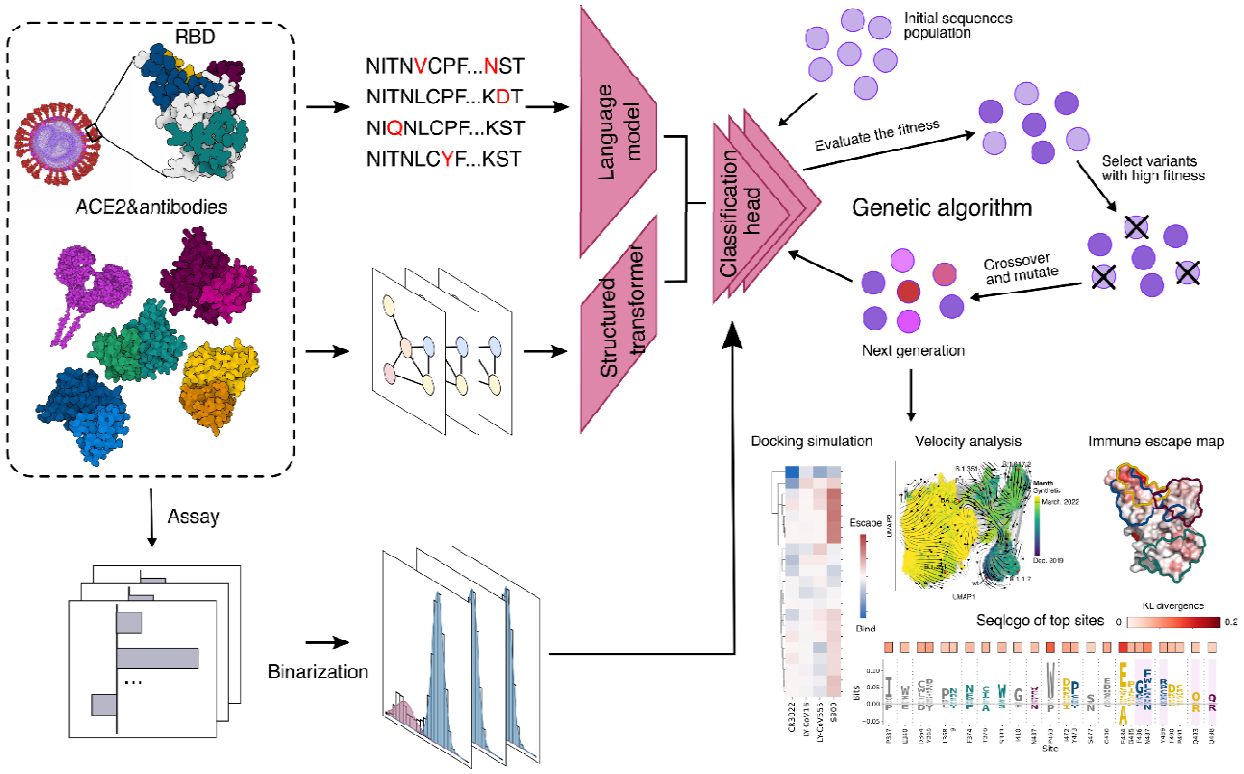


Fig. 1 | Overview of the MLAEP framework. Left: The multi-task learning model. We collected and cleaned the RBD variant sequences and their corresponding binding/escaping specificity to the ACE2 and eight antibodies. Then, the sequences and the structures of their binding partners were fed into the deep learning model with the multi-task learning objective. Right: The genetic algorithm. *In silico* directed evolution was performed to navigate the virtual fitness landscape defined by the nine scores from the multi-task model. The generation loop was repeated multiple times until the desired functionality was reached. These generated sequences were then subjected to *in silico* computational methods for evaluating their functional attributes.

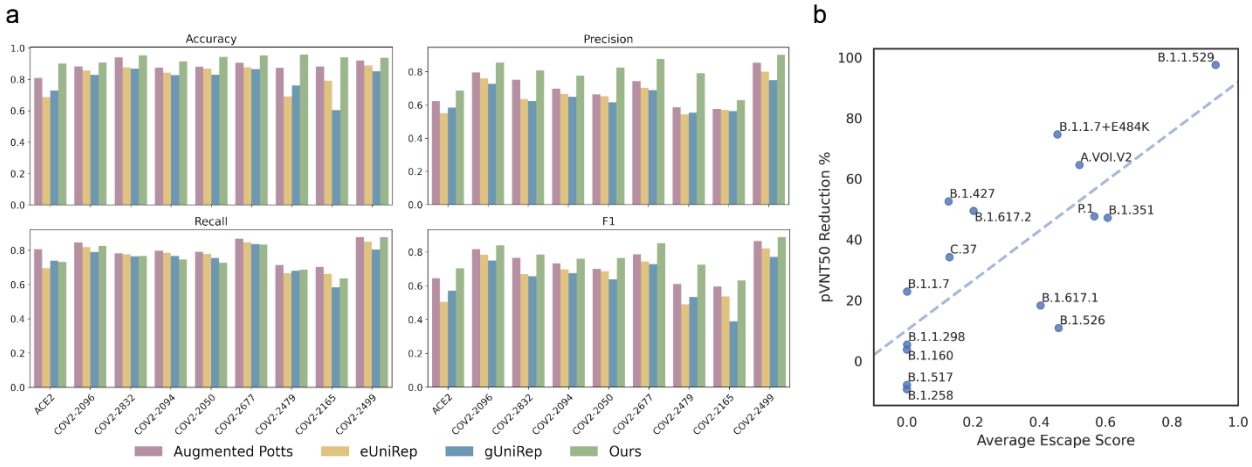


Fig. 2 | Performance evaluation and *in vitro* pVNT experimental data validation. a, Performance of augmented Potts model, eUniRep model, gUniRep model, and our model for predicting ACE2 binding and antibody escape are shown in terms of accuracy, macro precision, macro recall, and macro F1 score. b, Validation of the predicted immune escape potential using the pVNT assay data. The x-axis indicates the model predicted variant escape potential, while the y-axis is the $pVNT_{50}$ reduction of the VOCs compared with the wildtype.

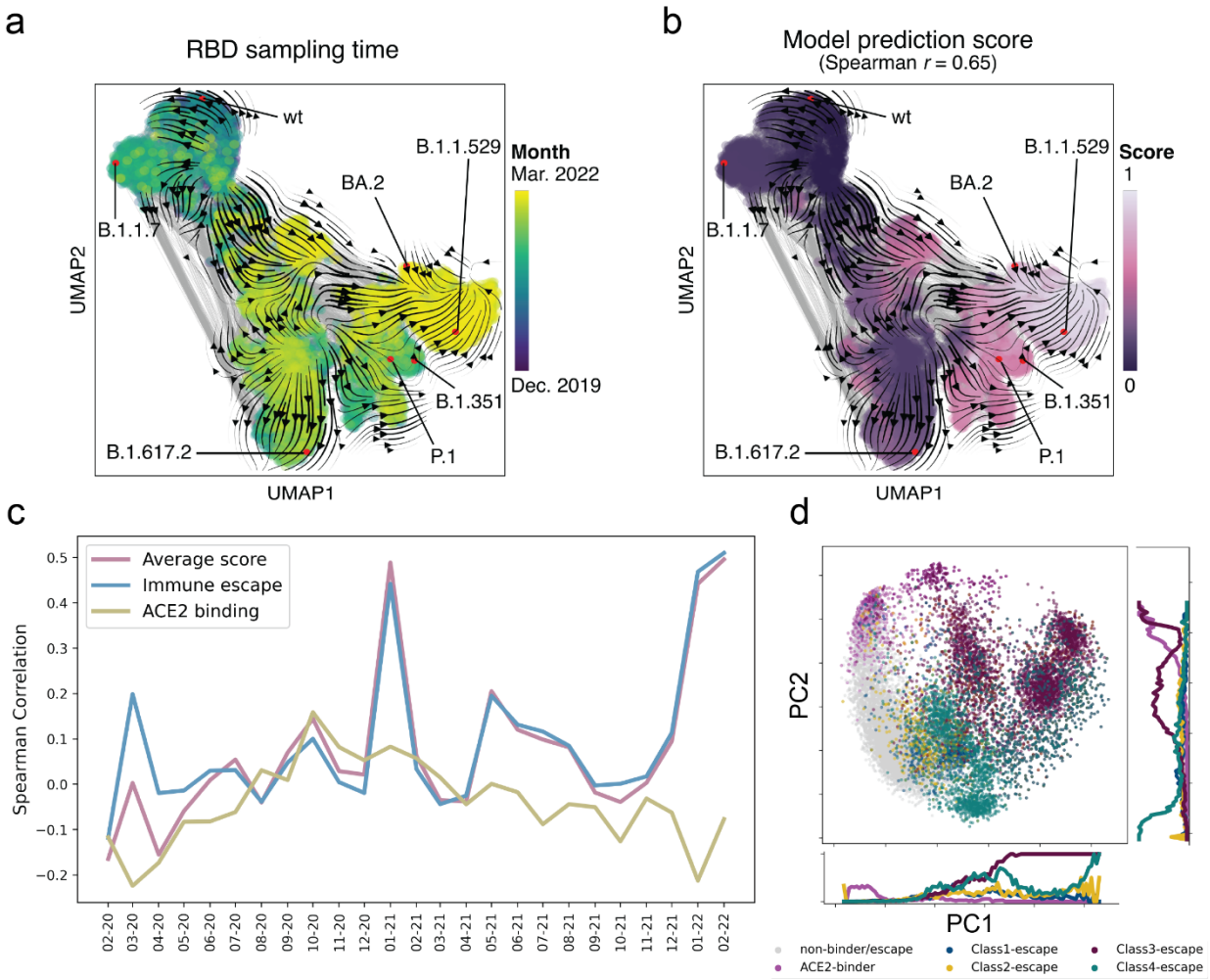


Fig. 3 | Multi-task model captures the antigenic evolutionary potential. a, The landscape of SARS-COV-2 RBD variant sequences (obtained from GISAID), represented as a KNN-similarity graph (with the darker blue region represents less recent date, e.g., 2019, and yellow represents more recent date, e.g., 2022). The gray lines indicate graph edges, while the colored points are sequences with the known sampling time. The streamlines among the points show a visual correlation between model predicted scores and the known sampling time. b, Use the average score of our model to visualize the landscape. The landscape is colored by the model prediction score with darker colors represent lower scores and lighter colors represent higher scores. c, Spearman correlation overtime for the model predictions, including the ACE2 binding score, immune escape potential, and the weighted average of the two in a time window of previous three months for each sampled

date. (From February 2020 to February 2022) d, Principal component analyses of the sequence's representations from our model, colored by the escaping/binding ability towards COV-2832, COV-2165(class 1 antibody), COV-2479, COV-2500 (class 2 antibody), COV-2096, COV-2499 (class 3 antibody), COV-2677, COV-2094 (class 4 antibody) and ACE2.

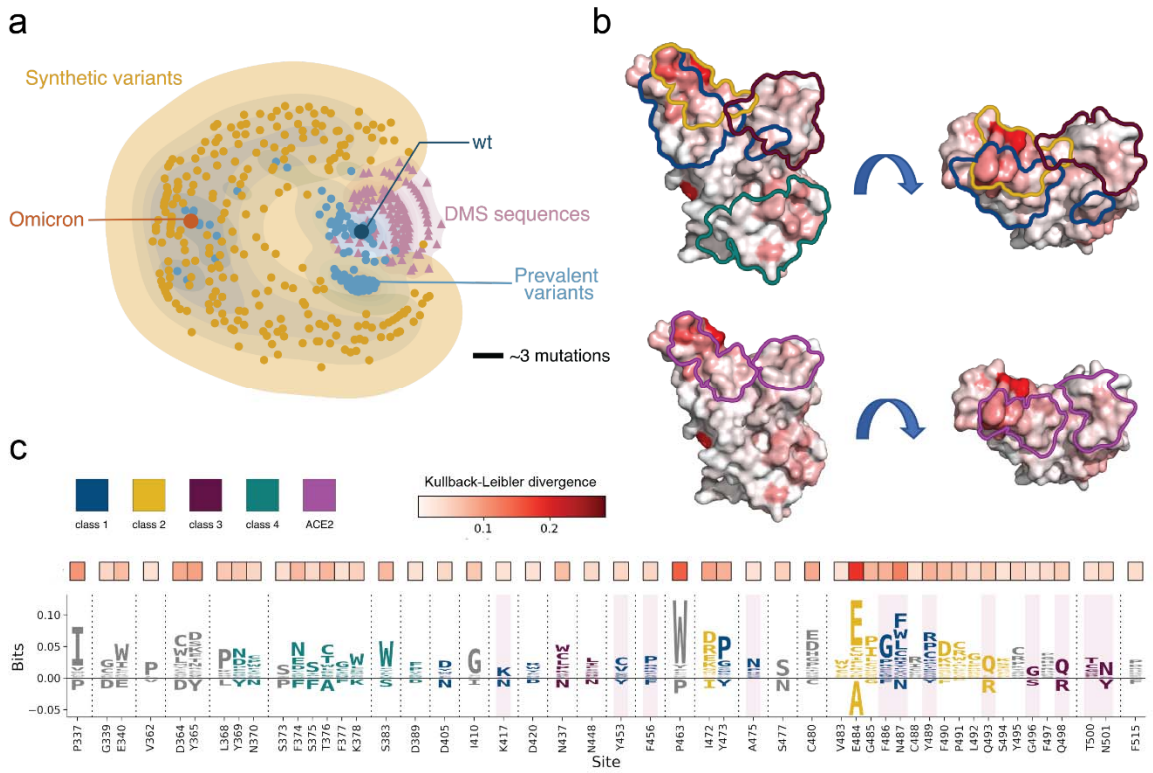


Fig. 4 | Overview of the synthetic sequences. a, Distance-preserving multidimensional scaling plot illustrates synthetic sequences' diversity compared to existing variants and deep mutagenesis sequences. A scale bar of three mutations is shown. b and c, the differences between the initial sequences and the synthetic sequences. b, The surface of the RBD protein, colored by the KL divergence between the initial sequences and the synthetic sequences. Colored outlines indicate the epitope structural footprint. c, The top 50 sites with the highest KL-divergence value are selected for visualizing the difference between the generated sequences and the existing sequences. Enriched amino acids locate at the positive side of the y-axis and depleted amino acids locate at the negative side.

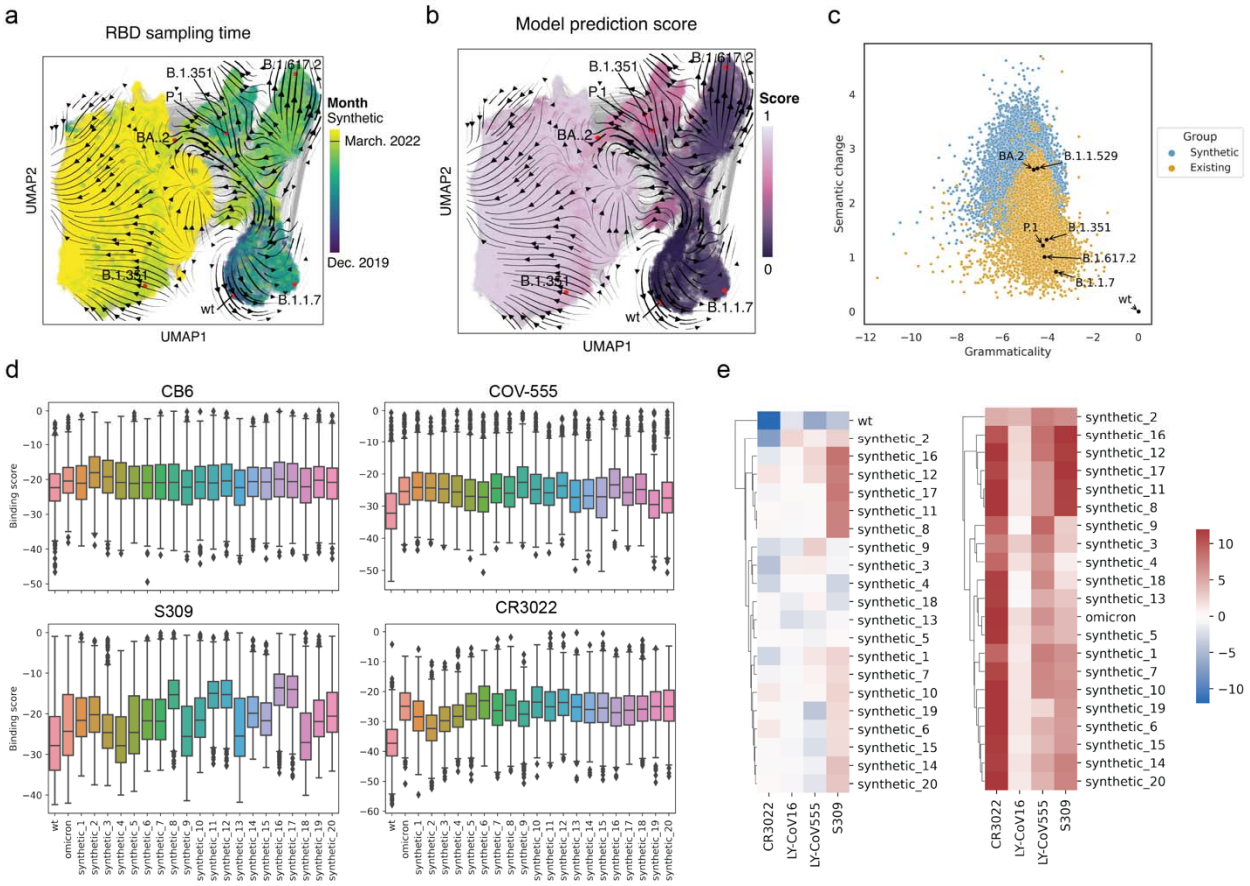
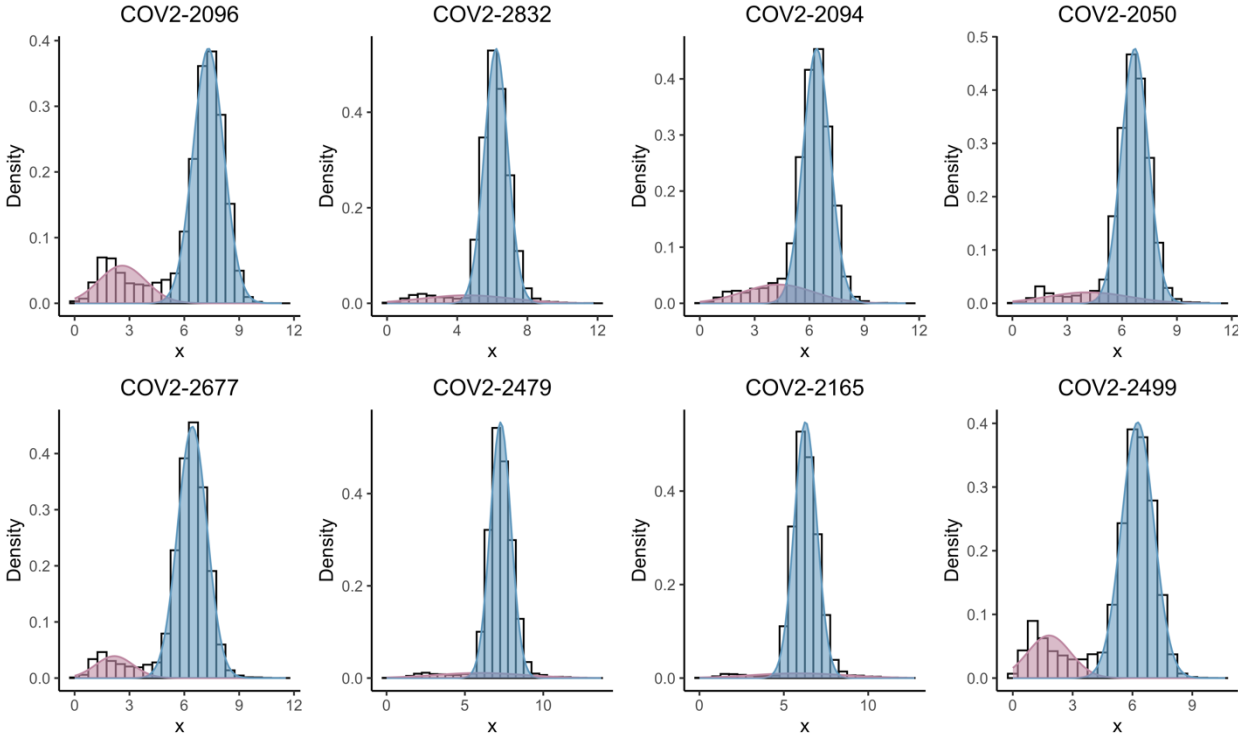
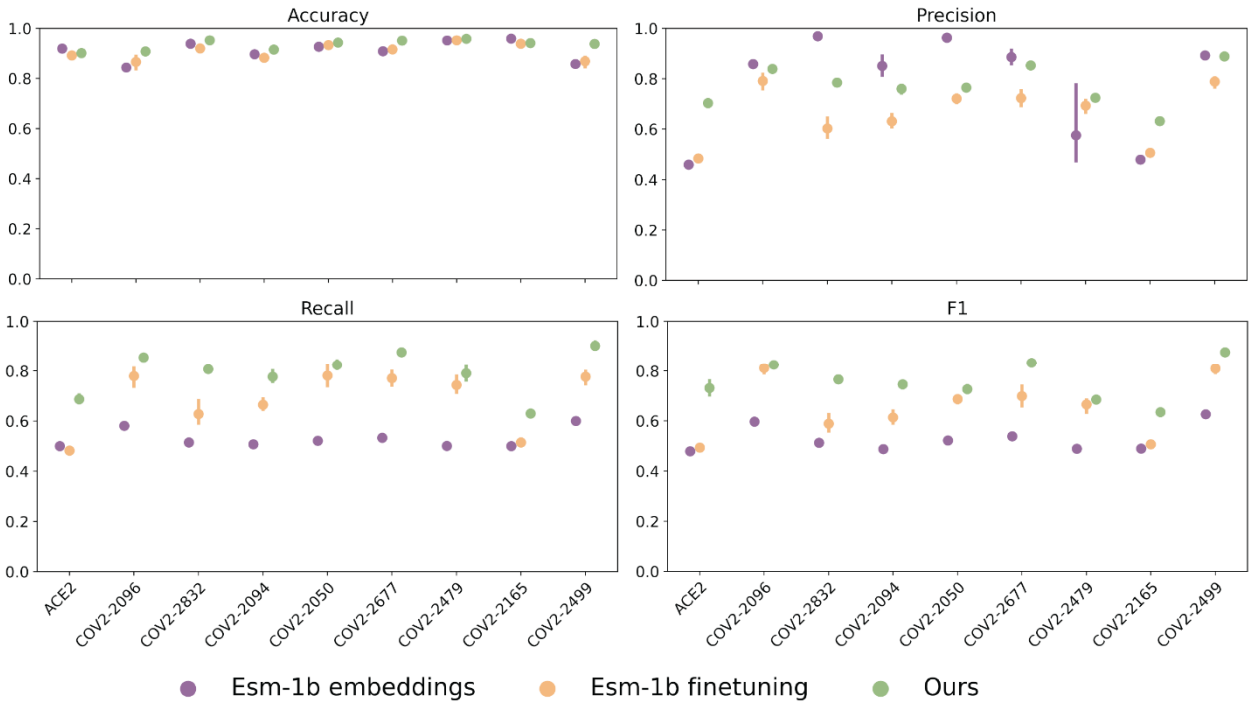


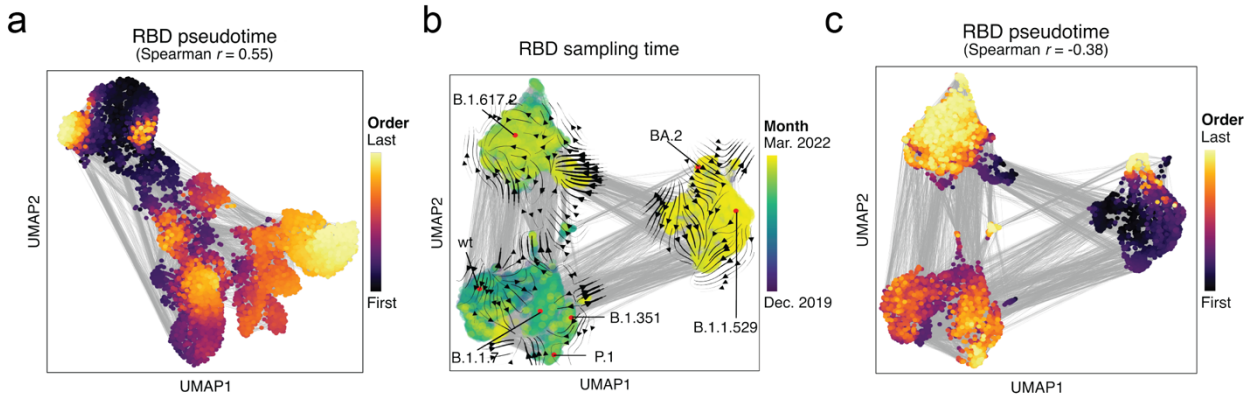
Fig. 5 | Evaluate the risks of the synthetic sequences. a, The landscape of existing SARS-COV-2 RBD variant (obtained from GISAID) and synthetic sequences, represented as a KNN-similarity graph. The gray lines indicate graph edges, the colored points denote the sequences with known sampling time/virtual time. The streamlines among the points show a visual trend from the existing sequences to synthetic sequences. b, The landscape of existing SARS-COV-2 RBD variant sequences (obtained from GISAID) and synthetic sequences, colored by model prediction score. c, The grammaticality and semantic analysis on existing GISAID sequences and synthetic sequences. d, The distribution of docking scores between four antibodies (CB6, S309, COV-555, CR3022) and RBD variants. Higher scores indicate better immune escape ability. The statistical significance is tested using student's *t-test*. e, Heatmap of the mean docking score difference. Red score suggests the variant sequence obtains higher immune escape ability compared with the Omicron (left) and Wuhan-wildtype(right).



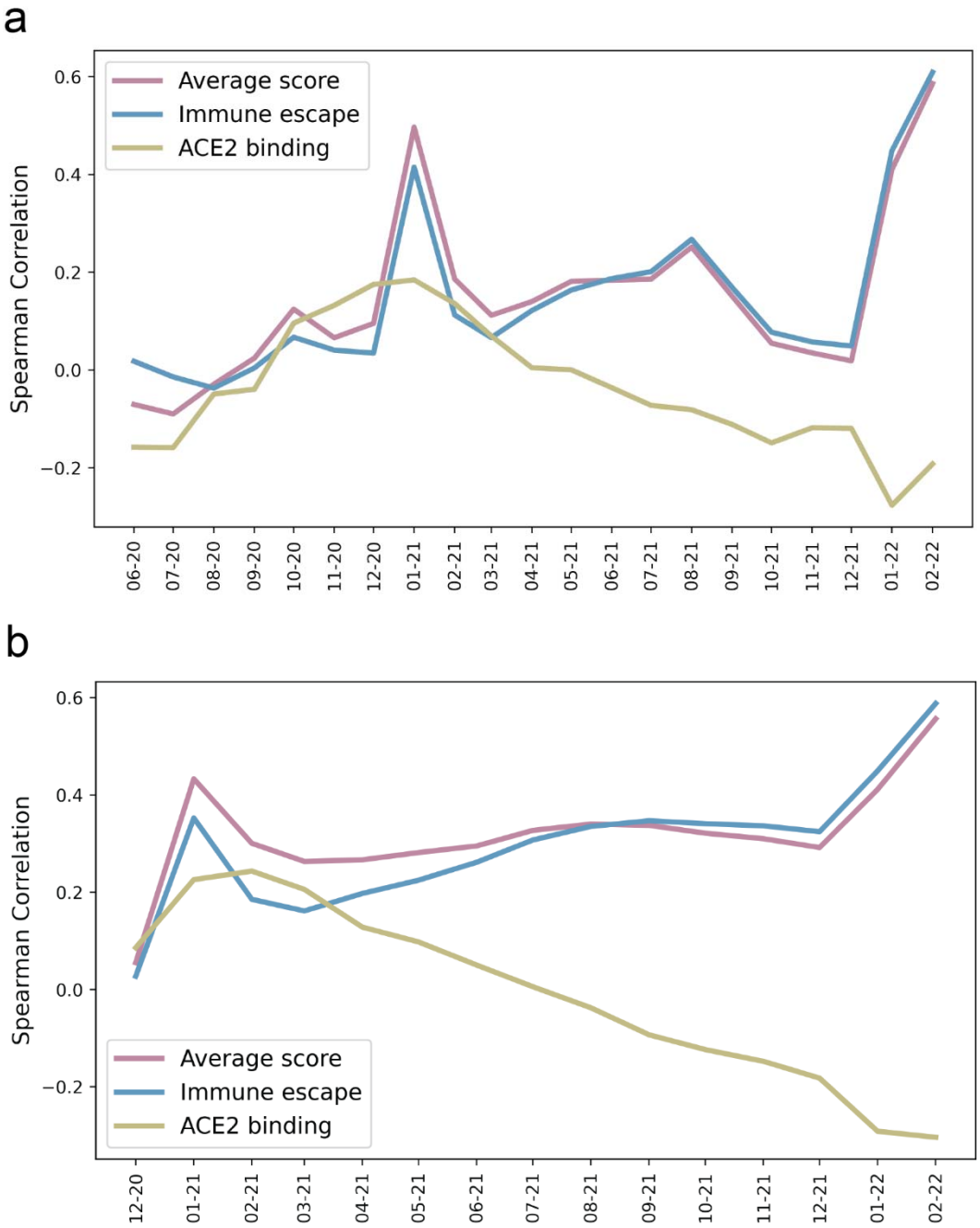
Extended Data Fig. 1 | The distribution of the DMS scores of eight antibodies. We log-transformed the deep mutational scanning scores, and got clearly two clusters for all antibodies. We then used the Gaussian mixture model to split the score into two clusters. Red cluster is defined as non-escape, while the blue clusters is defined as escape.



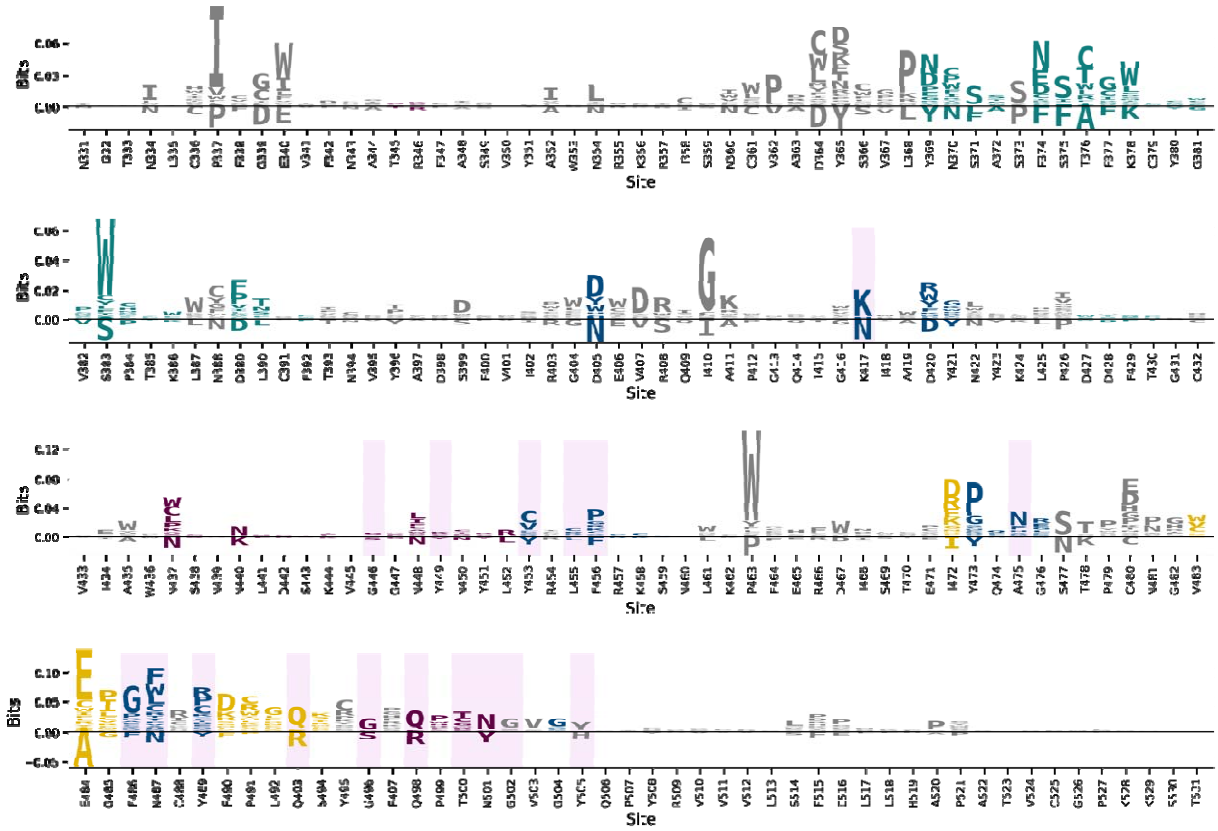
Extended Data Fig. 2 | Ablation studies. Justification of using the fine-tuning, the structure representations in the multi-task learning framework, in terms of Accuracy, macro-Precision, macro-Recall and macro-F1 score. ESM-1b shows the results of the fine-tuning steps' ablation. ESM-1b finetuning shows the results of replacing the Structured Transformer's ablation.



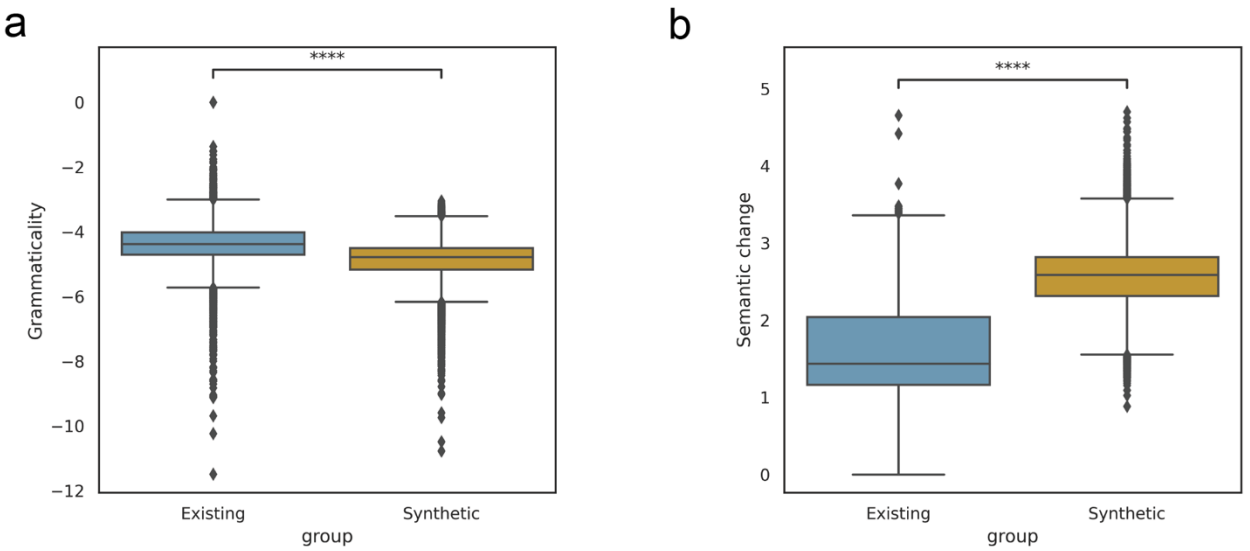
Extended Data Fig. 3 | Pseudo time and ESM-1b model inference. a, The landscape of RBD sequences, represented as a KNN network and visualized use the UMAP, colored with the inferred pseudo time using our model embeddings and scores. Gray lines indicate network edges. b and c, The landscape of the RBD sequences from GISAID, represented as a KNN similarity network and visualized with the UMAP. The sequences embeddings and directions among time are from the ESM-1b model. b, colored with the real-world sampling time. Streamlines show the visual correlation between the ESM-1b inferred velocity and sampling time. c, colored by the inferred pseudo time.



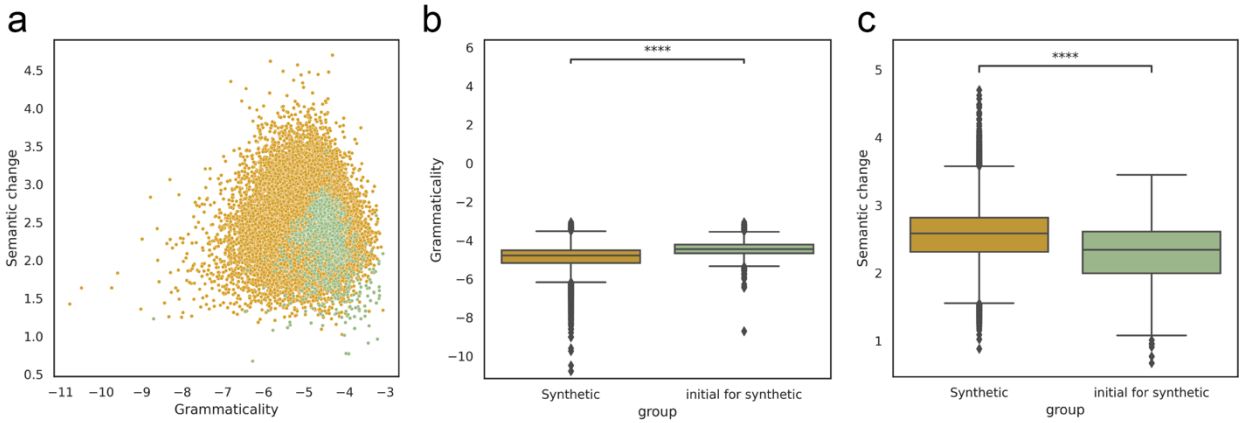
Extended Data Fig. 4 | Correlation between the model scores and sampling date. a, Spearman correlation overtime for the model predictions, including the ACE2 binding score, immune escape potential, and the weighted average of the two in a time window of previous six months for each sampled date. b, In a time window of previous 12 months for each sampled date.



Extended Data Fig. 5 | KL logo plot for the entire RBD region. Sequence logos generated from the differences of the generated sequences and the initial sequences, spanning the entire RBD region. The logos were calculated with probability weighted Kullback-Leibler divergence with a pseudo count of 0.1.



Extended Data Fig. 6 | Box plot for the existing GISAID sequences and the synthetic sequences. Grammatically and semantic change of the existing sequences and the synthetic sequences compared with that of the wild type Spike RBD sequence. Variant sequences with both high semantic change and high grammatically score is more likely to induce immune escape.



Extended Data Fig. 7 | The difference among the initial sequences and the synthetic sequences. a, Scatter plot represented the distribution of the two groups of sequences in terms of the Grammatically and semantic change. b and c, Box plot comparing the Grammatically and semantic change of the two groups. Both high semantic change and high grammatically score is more likely to induce immune escape.