E2VD: a unified evolution-driven framework for virus variation drivers prediction

Zhiwei Nie[1,2†], Xudong Liu[1,2†], Jie Chen[1,2*], Zhennan Wang[2], Yutian Liu[3,2], Haorui Si[4,5,6], Tianyi Dong[4,5,6], Fan Xu[2], Guoli Song[2], Yu Wang[2], Peng Zhou[6], Wen Gao[2,3], Yonghong Tian[1,3,2*]

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

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

Abstract

The increasing frequency of emerging viral infections necessitates a rapid human response, highlighting the cost-effectiveness of computational methods. However, existing computational approaches are limited by their input forms or incomplete functionalities, preventing a unified prediction of diverse viral variation drivers and hindering in-depth applications. To address this issue, we propose a unified evolution-driven framework for predicting virus variation drivers, named E2VD, which is guided by virus evolutionary traits priors. The architecture of E2VD includes two innovative modules: local-global dependence coupling and multi-task focal learning. The local-global dependence coupling module integrates the local dependence on nearby residues with the global dependence over the entire sequence at the motif level of a mutation, thereby amplifying the subtle effects induced by mutations. The multi-task focal learning module addresses the severe imbalance between beneficial and harmful mutations that shape the evolutionary

fitness landscape. E2VD comprehensively and significantly outperforms state-of-the-art methods across key virus variation drivers prediction tasks. Moreover, E2VD effectively captures the fundamental patterns of virus evolution, not only clearly distinguishing different types of mutations but also accurately identifying rare beneficial mutations critical to virus survival. Additionally, E2VD demonstrates superior performance on variation drivers prediction tasks across different viral lineages, exhibiting enhanced generalization capabilities. Importantly, E2VD can be flexibly combined to perceive virus evolutionary trends. It not only reveals the changes in evolutionary fitness caused by single-site mutations but also predicts high-risk mutation sites across various lineages. Overall, E2VD represents a unified, flexible, and interpretable approach for analyzing and predicting viral evolutionary fitness, providing an ideal alternative to costly wet-lab methods.

1 Introduction

Since 2000, numerous emerging viral infections [1, 2], especially COVID-19 [3], have seriously threatened the global public health system. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the causative agent of COVID-19, is a single-stranded RNA virus belonging to the coronaviridae family [4]. It enters cells by targeting the human angiotensin-converting enzyme 2 (ACE2) as the host receptor [5–7]. To combat the ongoing pandemic, various vaccines and drugs have been designed to mitigate the damage caused by SARS-CoV-2 [8, 9]. However, a concerning aspect is that SARS-CoV-2 continues to evolve, with mutations accumulating and recombination processes contributing to increased diversity [10, 11]. The sub-lineages of Omicron, in particular, remain prevalent and continue to cause breakthrough infections [12].

The impact of SARS-CoV-2 is only a microcosm of plenty emerging viral infections, reminding us to always pay attention to the possible emergence of unknown pathogens. More importantly, the increasing frequency of emerging viral infections demands rapider human responses. Uncovering evolutionary drivers to analyze evolutionary trends is an effective means to increase response speed. Some wet-lab methods have been developed to measure the mutation-induced properties of viruses [13–16], which are then adopted to analyze the evolutionary trajectories and mutation hotspots. However, wet-lab measurement is often used for post-hoc analysis because their huge cost of money and time prevents them from being carried out casually as an active-exploration method. Therefore, efficient and low-cost computational methods are urgently needed.

The computational methods currently available for virus evolution analysis, especially deep-learning methods, can be roughly divided into two categories in terms of input form: methods that only use sequences and methods that incorporate structural information. Structure-dependent methods [17–20] require input of the tertiary structure of the protein being studied, such as the structures of antibodies, the ACE2, or various protein complexes. In contrast, sequence-only methods [21–27] adopt protein sequences as input to the model, i.e., in the manner that does not require structures. The above methods can be divided into three categories in terms of the functions they implement, namely specific property prediction, evolutionary trajectory reproduction, and evolutionary trend prediction. Specific property prediction methods [19, 20, 22, 23, 25–27] aim to predict mutation-induced changes in certain properties, such as binding affinity, expression, and antibody escape. The evolutionary trajectory reproduction methods [21] can reproduce the evolutionary route of the observed variants. The evolutionary trend prediction methods [17, 18, 24] are capable of providing predictive analysis of evolutionary trends.

However, current methods have significant limitations. First, we analyze the methods under the classification system of input forms. Structure-dependent approaches [17–20] are severely limited by the rarity or even absence of high-quality protein structures within specific families. Even though deep-learning-based structure prediction methods [28–31] have made significant progress, inherent prediction errors have unpredictable effects on the accurate prediction of slight structural rearrangements caused by mutations of few sites [32–35]. In contrast, one-dimensional protein sequences are

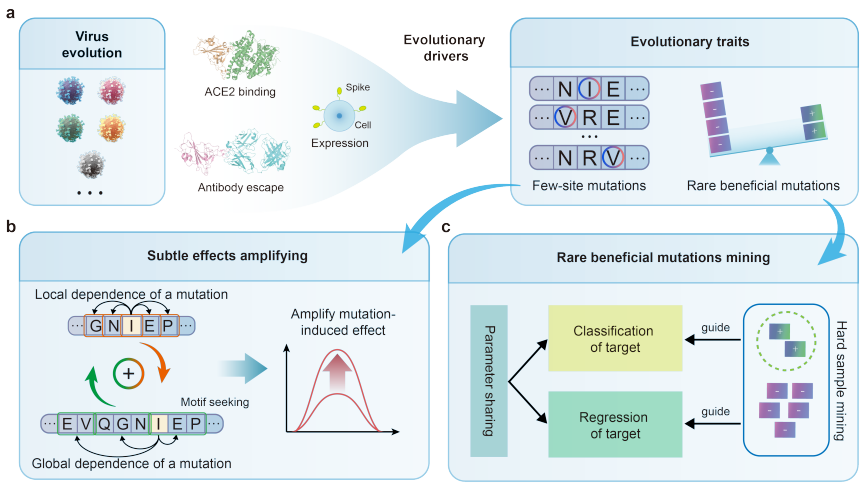

Fig. 1 The motivation and methodology of E2VD. **a**, The evolutionary traits of viruses under diverse evolutionary drivers, including 'few-site mutations' and 'rare beneficial mutations'. **b**, Local-global dependence coupling module is proposed to amplify the subtle effects caused by mutations at few sites through integrating the local dependence of a mutation on nearby amino acids and the global dependence at the motif level over the entire sequence. **c**, Multi-task focal learning module is proposed to assign more focus on rare beneficial mutations to improve prediction performance through multi-task learning mode.

easy to obtain, but their distance alone 'sequence-structure-function' route makes predicting functions based on the sequence uninterpretable, which is the fundamental reason for the performance bottleneck of sequence-only methods [21–27]. Secondly, we analyze the methods under the classification system of the implemented functions. Current specific property prediction methods [19, 20, 22, 23, 25–27] lack comprehensiveness and unification, and often only develop one or two customized predictors to achieve prediction of specific properties. Such methods are not suitable for studying virus evolution because they do not comprehensively consider diverse variation drivers. The evolutionary trajectory reproduction methods [21] can only process observed variants and lacks predictive capabilities, which is essentially a type of post-hoc data modeling and analysis. Although evolutionary trend prediction methods [17, 18, 24] have emerged, they still have obvious shortcomings of incomplete consideration of diverse variation drivers or dependence on structure input. Therefore, it is crucial to develop a unified, structure-free, and interpretable virus variation drivers prediction method and flexibly apply it to the prediction of evolutionary trends.

The prerequisite for achieving unification and interpretability is to design a deep-learning framework under the guidance of the basic rules of virus evolution. Mutation is the cornerstone of viral evolution [36], and RNA viruses have significantly higher mutation rates [37]. While a small proportion of mutations are beneficial and some are nearly neutral, the majority of mutations are harmful to viral fitness [38]. The ratio of harmful to beneficial mutations may vary across organisms and environments, but harmful mutations are thought to always outnumber beneficial mutations across the

evolutionary scale [37, 38]. Correspondingly, most mutations at the amino-acid level (amino acid substitutions, deletions, or insertions) are harmful to the survivability of viral proteins, and only a small part of them are beneficial. In other words, beneficial mutations are a rare subset of the viral protein's evolutionary fitness [39]. A large number of harmful mutations significantly reduces the likelihood of multiple mutations coexisting in a sequence. Generally, only few sites in a viral protein sequence are mutated, that is, a small proportion of the overall residues are changed compared to wild-type. As shown in Fig.1a, we can summarize the above priors as 'few-site mutations' and 'rare beneficial mutations', which bring corresponding challenges. First, mutations at few sites are difficult to significantly affect the reconstruction of molecular interactions, which results in a very subtle effect on function, making it difficult for neural networks to capture. Second, the scarcity of beneficial mutations causes extreme imbalance in sample categories from a data perspective, which poses a huge challenge to training deep learning models that accurately mine rare beneficial mutations.

In order to solve the difficulties caused by the above evolutionary traits, we propose E2VD (**E**volution-driven **V**irus **V**ariation **D**rivers prediction), a unified deep-learning framework that incorporates priors guidance. The first innovative module, local-global dependence coupling, integrates the local dependence on nearby amino acids with the global dependence at the motif level over the entire sequence of a mutation, thereby amplifying subtle effects caused by few-site mutations (Fig.1b). The introduction of local dependence aims to emphasize the impact of a mutation on their neighboring amino acids to highlight the local interaction system between them. The global dependence comes from the fact that amino acids that are not adjacent to a mutation at the sequence level may be adjacent to it at the structural level, thus jointly supporting the local structural micro-environment of the mutation. Therefore, the introduction of explicit global dependence facilitates complete restoration of the complete interaction system of mutations at sequence level. More importantly, considering that biologically meaningful protein fragments, i.e., motifs, carry functional information, we adopt dynamic attention granularity to establish the interaction dependence of mutations with motifs. This allows the effects of mutations to be more directly linked to functions of viral proteins. The second innovative module, multi-task focal learning, is proposed to alleviate the dilemma caused by the severe imbalance in the number of beneficial and harmful mutations. As shown in Fig.1c, accurately predicting rare beneficial mutations is transformed into a problem of hard sample mining, that is, beneficial mutations are treated as hard samples. Prediction for a specific variation driver is converted into two task forms: classification and regression of a target viral protein property. The prediction performance is promoted by learning the above related tasks simultaneously under the framework of multi-task learning [40] with the supervision of a novel 'multi-task focal loss'.

With the support of the above two innovative modules, E2VD predicts key virus variation drivers (including binding affinity, expression, and antibody escape) in a structure-free form, significantly outperforming the state-of-the-art methods across the board. We additionally explore customized strategies of protein language models for variation drivers prediction, demonstrating the effectiveness of designing pre-training

data for specific downstream tasks. Moreover, E2VD effectively captures the fundamental patterns of virus evolution. It not only clearly distinguishes different types of mutations that make up the evolutionary fitness landscape, but also accurately mines rare beneficial mutations that are critical to virus survival. In addition, E2VD exhibits enhanced generalization than its competitors across different lineages, indicating its potential to cope with intricate evolutionary trends. E2VD demonstrates excellent capabilities in the above three aspects: prediction performance, mutation-induced evolution pattern capturing, and model generalization, laying a solid foundation for virus evolutionary trends perception. The binding affinity, expression and antibody escape predictors developed based on E2VD are flexibly combined to comprehensively restore the evolutionary drivers faced by the virus, thereby screening for potential high-risk mutations in the form of multi-round screening. As expected, the predicted viral evolutionary trends are generally consistent with the distribution of mutation hot-spots measured by wet-lab experiments and E2VD-core pipeline significantly outperforms the state-of-the-art evolutionary trends prediction method. In summary, E2VD is a unified, structure-free, and priors-driven interpretable framework for virus variation drivers prediction, which captures the basic pattern of viral evolution while achieving excellent prediction performance. E2VD offers a cost-effective *in silico* alternative means for analyzing and predicting viral evolutionary fitness, accelerating responses to emerging viral infections.

2 Results

2.1 The architecture of E2VD is driven by evolutionary traits

The overall architecture of E2VD is shown in Fig.2a, and its core is to solve the modeling problems caused by evolutionary traits. First, protein language models (PLMs) with frozen parameters are adopted to extract amino-acid-level embeddings of mutated RBDs or antibody variable-regions. PLMs have made impressive progress on functional and structural prediction tasks [30, 39, 41–44], which is the motivation for their use in extracting embeddings of protein sequences in E2VD. Next, the above amino-acid-level embeddings are fed into the **L**ocal-**G**lobal dependence coupling (LG) module. In this module, Convolutional Neural Network (CNN) [45] is used to build local dependencies of mutations, while ContextPool-attention [46] is used to build global dependencies of mutations at the motif level. With its excellent ability to integrate local dependencies, CNN has been widely used in protein-protein interaction predictions [47–49] or protein-ligand interaction predictions [50–52]. The purpose of using CNN here is to establish the interaction dependence between a mutation and its neighboring amino acids, thereby extracting the latent information of the interaction system around the mutation. However, local dependence at the sequence level alone is not sufficient to predict the reconfiguration of the real interaction system caused by mutations, because residues that are sequenced distant in the tertiary structure may become proximate and thus interact significantly with the mutation. In addition, motifs [53] carrying specific functions are considered to be fragment-level sequence patterns [54], which have been proven to be beneficial to the prediction of protein function by multiple past

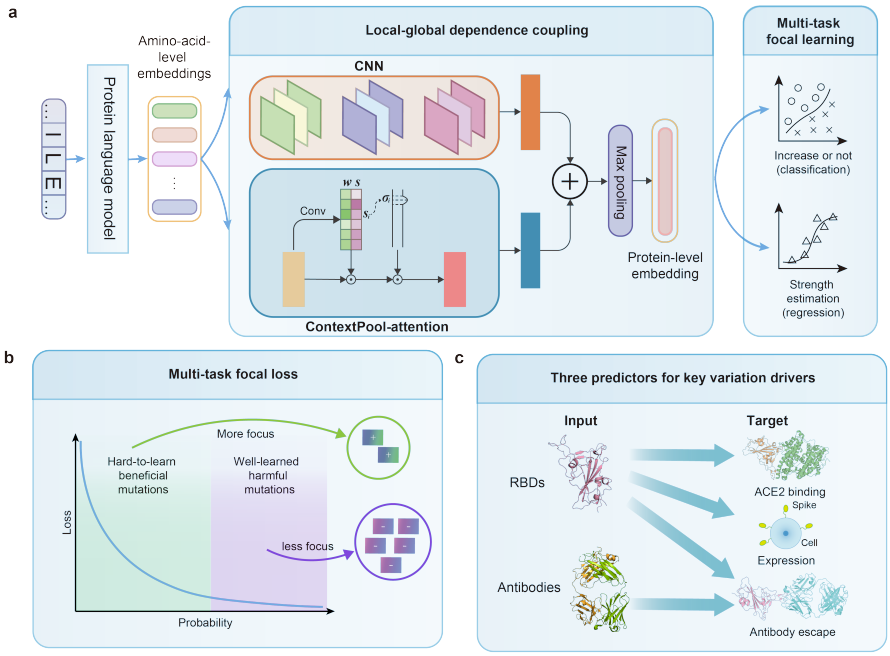

Fig. 2 **a**, The overall architecture of E2VD. A protein sequence is first extracted as amino-acid-level embeddings, and then converted into a protein-level embedding through the local-global dependence coupling module consisting of CNN and ContextPool-attention. Finally, the protein-level embedding flows into the multi-task focal learning module for multi-task learning of classification and regression. **b**, Multi-task focal loss is designed to help the model better learn hard-to-learn samples. Beneficial mutations are assigned more attention due to their rarity, while harmful mutations that are already well-learned in large numbers do not require much attention. **c**, Three types of prediction tasks for E2VD, including binding affinity, expression and antibody escape. The binding affinity and expression prediction tasks require the input of the mutated RBDs, while the antibody escape prediction task requires the input of both the mutated RBDs and the antibodies.

works [55–57]. Therefore, we further adopt ContextPool-attention with dynamic granularity to establish the global dependence of mutations based on seeking for motifs, thereby complementing the local dependencies of mutations. Specifically, the original self-attention [58] only constructs query and key vectors at a fixed granularity for each individual token. In the scenario of protein sequences, only residue-level global dependencies can be established. To solve this problem, we set a learnable granularity to realize phrase mining, where the phrases correspond to motifs in protein sequences. The global dependencies established on this basis take into account motifs carrying specific spatial conformation or specific functional information, and naturally introduce more explicit sequence patterns corresponding to functions. Finally, the local dependencies and motif-level global dependencies are extracted simultaneously, and integrated to get refined protein-level embedding.

The protein-level embedding described above then flow into the **M**ulti-**T**ask focal learning (MT) module. In this module, the prediction task of specific variation driver

is transformed into two closely related tasks, classification and regression, so that they can be trained simultaneously in the form of multi-task learning. More importantly, hard-to-learn samples, i.e. rare beneficial mutations, require customized loss functions to make the neural network pay more attention to (Fig.2b). Focal loss for object detection [59] performs well in classification tasks with imbalanced categories, which inspired us to design customized focal loss for regression tasks. The focal loss for classification and designed focal loss for regression are combined into our proposed 'multi-task focal loss', which is a compatible loss function that simultaneously utilizes task correlation and hard sample mining strategies.

The drivers for virus variation are diverse, which requires E2VD to be unified to adapt to multiple scenarios. In this work, we consider viral protein binding affinity, expression, and antibody escape, three key drivers [13, 60, 61] that significantly influence the variation process of SARS-CoV-2 (Fig.2c). For binding affinity prediction, only mutated RBDs are input to predict their binding to human ACE2 because the binding target is fixed in this task. There is no interaction involved in expression prediction, so only mutated RBDs are input. Different from the above two, antibody escape prediction involves interactions between variants and multiple antibodies, in which mutated RBDs and variable regions of multiple antibodies are input together.

2.2 Prediction performance of E2VD

There are currently a variety of off-the-shelf PLMs to choose from, such as SeqVec [44], ProtTrans series [41] and ESM-2 series [30]. Previous work ESM-1v [42] compared the impact of pre-training datasets with different sequence identities on zero-shot inference performance, highlighting the importance of pre-training data in the design of PLMs, which inspires us to explore how to train dedicated PLMs suitable for prediction of viral variation drivers. To improve adaptability, we customize the pre-training dataset selection and the truncation length in this work. First of all, the evolution of viruses is accompanied by the continuous accumulation of novel mutations [37, 38, 63], which requires the protein language model to have strong enough zero-shot generalization to maintain the quality of the extracted viral protein sequence embedding. Therefore, we adopt the Uniref90 dataset for pre-training that ESM-1v has proven to achieve the best zero-shot inference performance [42]. The advantage of Uniref90 dataset is that it contains richer evolutionary information and does not cause performance degradation in the early stage of training like the Uniref100 [42], facilitating zero-shot generalization performance by helping the protein language model see as many sequence examples from protein families as possible. Secondly, considering that the sequence length of the input RBDs (length 201) or antibody variable regions (length 136) is short, we set the maximum truncation length of the protein sequence to 256 for pre-training to facilitate the capturing of protein fragment patterns smaller than this length.

To evaluate the effectiveness of customized pre-training configuration on the performance of viral variation drivers prediction, the widely recognized pre-trained PLMs are compared with our customized one. As shown in Fig.3abc, we conduct comprehensive prediction performance comparisons of different PLMs on the three key variation drivers prediction tasks: binding affinity, expression, and antibody escape. To ensure

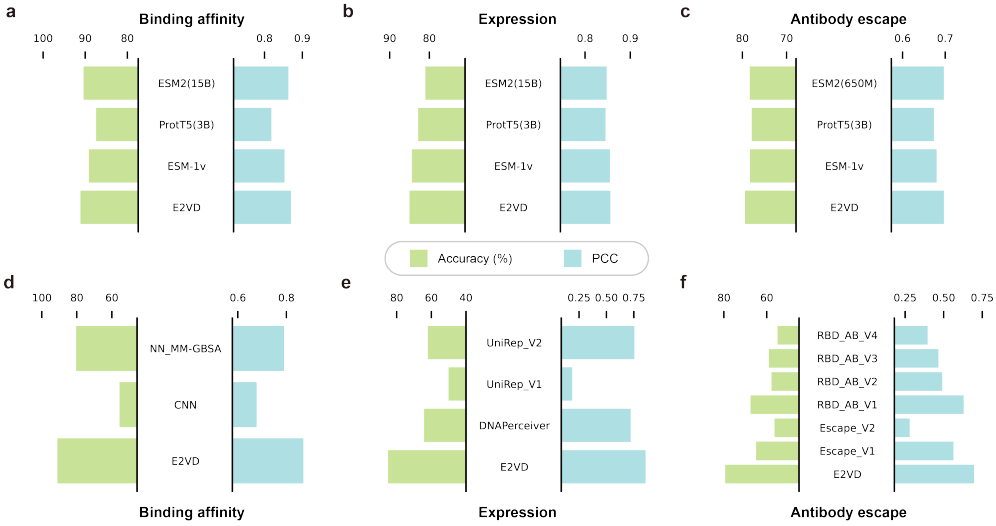

Fig. 3 **abc**, Prediction performance of off-the-shelf PLMs and our customized one on three key variation drivers prediction tasks, including binding affinity (a), expression (b), and antibody escape (c). To ensure a fair comparison, we only replace different pre-trained PLMs on the original architecture of E2VD, and the remaining modules remain unchanged. **def**, E2VD is benchmarked on binding affinity (d), expression (e), and antibody escape (f) prediction tasks respectively. For expression prediction, two models of a previous study [26] trained with UniRep embeddings [62] for Random Forest (RF) and Artificial Neural Network (ANN) training are referred to as UniRep_V1 and UniRep_V2. For antibody escape prediction, two antibody escape prediction models of a previous study [23] based on pre-trained protein language models are referred to as Escape_V1 and Escape_V2. In addition, four models of the previous method RBD_AB [27] with different masking ratios are referred to as RBD_AB_V1, RBD_AB_V2, RBD_AB_V3, and RBD_AB_V4 respectively.

a fair comparison, we only replace different pre-trained PLMs on the original architecture of E2VD, and the remaining modules remain unchanged. Taken together, E2VD with our customized PLM achieves the best prediction performance with the smallest parameter size (340M), which proves the effectiveness of our customized pre-training configuration Remarkably, ESM-1v (650M) performs well on three tasks with its design for zero-shot prediction, tying or even surpassing ESM-2 with 15B parameters, proving the effectiveness of using Uniref90 for pre-training. In addition, the prediction performance powered by our customized PLM (340M) comprehensively surpasses ESM-1v (650M), which demonstrates that the pre-training truncation length tailored to the downstream task improves the adaptability of language models for virus variation drivers prediction.

As shown in Fig.3def, E2VD is benchmarked on binding affinity, expression, and antibody escape prediction tasks respectively. First, for binding affinity prediction, the widely used CNN and a previous method NN_MM-GBSA [20] are adopted for comparison. The accuracy of the state-of-the-art method NN_MM-GBSA [20] in the classification task is improved by 10.7% by E2VD, along with a 8% improvement of the Pearson's Correlation Coefficient (PCC) in the regression task. Second, for expression prediction, three models from two previous studies [25, 26] are used for

comparison. The accuracy of the state-of-the-art method DNAPerceiver [25] in the classification task is improved by E2VD by 20.8%, and its PCC in the regression task is improved by 13.5%. Finally, for antibody escape prediction, six models from two previous studies [23, 27] are used for comparison. The accuracy of the state-of-the-art method RBD_AB_V1 [27] in the classification task is improved by more than 11.8% by E2VD, while at the same time, its PCC in the regression task is improved by about 6.8%. Overall, E2VD comprehensively and significantly surpasses the state-of-the-art methods in the above three tasks, demonstrating the universality and advancement for virus evolutionary drivers prediction.

2.3 E2VD captures fundamental patterns of virus evolution

As the core modules of E2VD, **L**ocal-**G**lobal dependence coupling (LG) and **M**ulti-**T**ask focal learning (MT) need to be further explored for their contribution to prediction performance. As shown in Fig.4a, we first conduct ablation study of modules on binding affinity prediction task. Four sets of experiments are designed for comparison, including the original E2VD (referred to as Ours), E2VD without the LG module (referred to as w/o LG), E2VD without the MT module (referred to as w/o MT), and E2VD without both modules (referred to as w/o MT&LG). We find that the introduction of the LG module is beneficial to both classification and regression tasks, with the performance of the regression task being significantly improved (PCC: 0.385 to 0.871). In contrast, the MT module significantly improves the performance of classification task (Accuracy: 50% to 68.52%). It is worth noting that the Recall is sharply improved (Recall: 0 to 69.63%), which means that MT is effective in mining rare beneficial mutations for virus fitness. The combined use of the above two modules further improves the performance of classification and regression tasks, reaching the Accuracy of 91.11%, the Recall of 96.3%, and the PCC of 0.87. We also conduct module ablation experiments based on ESM-2 (650M) and ProtT5 (3B) (Supplementary information, Table.S11, Table.S12), and the results are consistent with the above conclusions.

As shown in Fig.4b, we perform dimensionality reduction visualization of different types of mutations in the binding affinity prediction task using Principal Component Analysis (PCA). Mutations that are beneficial for viral fitness are risky for humans, so we describe such mutations as 'risky'. Similarly, mutations that are harmful to viral fitness are not risky for humans, so we describe such mutations as 'risk-free'. Considering that risk-free mutations account for the majority in the deep mutational scanning (DMS) dataset [13] of the binding affinity prediction task, we further divide this type of mutations into two categories. Therefore, three categories of mutations are presented in total, including Risk-free (K_a $ratio$<0.5), Risk-free (0.5<K_a $ratio$<1), and Risky (K_a $ratio$>1). K_a $ratio$ represents the ratio of the binding strength between the mutated receptor-binding domain (RBD) and human ACE2 relative to the wild-type. A value greater than 1 represents an increase in binding affinity, and a value less than 1 represents a decrease in binding affinity. We can see that the three types of mutations cannot be distinguished in the feature space before being processed by the LG module, showing a very high degree of overlap. In sharp contrast, the three types of mutations processed by the LG module are clearly distinguished, with discernible boundaries.

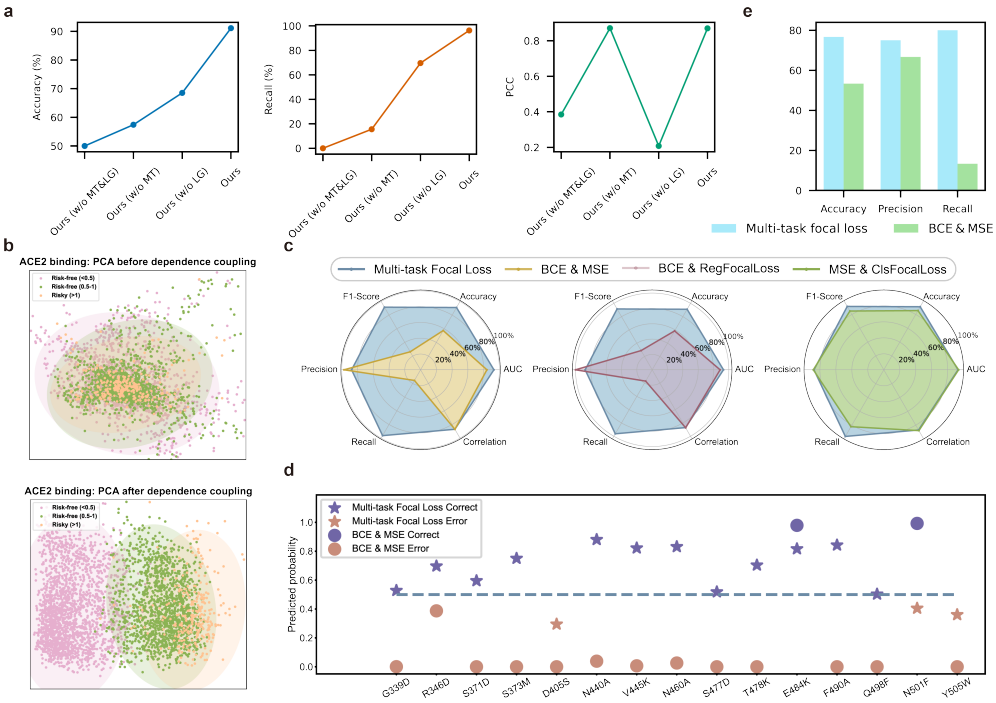

Fig. 4 **a**, The module ablation study on the binding affinity prediction task. LG refers to Local-global dependence coupling module, and MT refers to Multi-task focal learning module. **b**, Dimensionality reduction visualization of different types of mutations on the binding affinity prediction task. Three categories of mutations are presented, including Risk-free (K_a *ratio*<0.5), Risk-free (0.5<K_a *ratio*<1), and Risky (K_a *ratio*>1). **c**, The ablation study on multi-task focal loss to analyze the impact of specific focal loss on prediction performance. **d**, To test the ability to capture rare mutations that are beneficial to virus fitness (i.e. beneficial mutations), our proposed multi-task focal loss and BCE&MSE are used for comparison. **e**, Comparison of the overall prediction performance of the two loss functions after taking into account beneficial mutations and harmful mutations.

We can conclude that LG makes E2VD sensitive to different types of mutations by restoring the interaction system faced by mutations, and thus perceives the evolutionary fitness of the virus. We also perform dimensionality reduction visualization on the datasets of the expression prediction task and antibody escape prediction task, and the results further support the above conclusion (Supplementary information, Fig.S1).

We further conduct ablation experiments on multi-task focal loss to analyze the impact of specific focal loss on prediction performance. As shown in Fig.4c, we compare four different loss functions for multi-task learning, including the combination of Mean Squared Error (MSE) loss and Binary Cross Entropy (BCE) loss, BCE paired with focal loss for regression (referred to as RegFocalLoss), MSE paired with focal loss for classification (referred to as ClsFocalLoss), and the proposed multi-task focal loss. We find that the introduction of RegFocalLoss or ClsFocalLoss alone improves the overall prediction performance. The combined use of RegFocalLoss and ClsFocalLoss, that

is, multi-task focal loss, further improves the overall performance, with the Accuracy increasing from 57.41% to 91.11% and the Recall increasing from 15.56% to 96.30%.

In order to further test the ability of multi-task focal loss to capture rare mutations that are beneficial to virus fitness, we select XBB.1.5 [64] that has spread widely as the reference lineage. Specifically, we randomly select 15 beneficial (increased binding affinity) mutations based on DMS data of 22 real-world mutation sites within its RBD [1]. To ensure comprehensiveness, we also randomly select a harmful (reduced binding affinity) mutation at each beneficial mutation site to form a test set with balanced positive and negative samples. As shown in Fig.4d, E2VD with multi-task focal loss achieves excellent prediction performance on beneficial mutations, correctly predicting 12 out of 15 beneficial (risky for humans) mutations, i.e. 80% of the total. In contrast, the prediction performance of E2VD with BCE&MSE lags far behind, only correctly predicting 2 mutations, i.e. about 13% of the total. For harmful mutations, multi-task focal loss and BCE&MSE achieve close prediction performance, correctly predicting 11 and 14 respectively (Supplementary information, Fig.S2). As we expected, since BCE&MSE cannot help the model learn rare beneficial mutations, the model tends to predict all mutations as harmful (14 out of 15). Taking into account beneficial and harmful mutations in Fig.4e, three indicators, including Accuracy, Precision, and Recall, are adopted to evaluate overall prediction performance. The E2VD with multi-task focal loss outperforms the BCE&MSE version comprehensively and significantly, with the Recall increasing by a factor of six.

In summary, E2VD sensitively distinguishes different types of mutations by amplifying the subtle effects caused by a few mutations through local-global dependence coupling module, and accurately mines rare risky (beneficial) mutations through the difficult sample mining strategy of multi-task focal learning module, thereby capturing the fundamental patterns of virus evolution.

2.4 Generalization performance of E2VD on different lineages

Viruses continue to evolve under selective pressure, in which SARS-CoV-2 has evolved into multiple lineages [63]. Therefore, the generalizability of virus variation drivers prediction methods is crucial to cope with intricate evolutionary trends. Considering that the wet-lab experiment data used for training is severely affected by batch effects [65], predicting specific values of variation driver is difficult to promote the efficiency of discovering high-risk variants. On the contrary, predicting the relative order can help humans sort variants according to evolutionary fitness, which is more instructive. At this point, the inverse pairs reflecting the degree of disorder of the sequence [66, 67] catches our attention. Specifically, the greater the number of inverse pairs in a sequence, the higher the degree of disorder of the sequence. If the number of inverse pairs is small, the sequence is relatively ordered. Inspired by this, we propose **O**rdinal **P**air **P**roportion (OPP) to evaluate the generalization performance of different models. As shown in Fig.5a, if the fitness of pairwise mutations in the ground-truth fitness landscape still maintains the original ordering in the predicted fitness landscape, this is considered to be a correctly predicted pair. If the ground-truth ordering of the fitness of pairwise mutations is inverse in the predicted fitness landscape, this is considered

[1] https://outbreak.info/

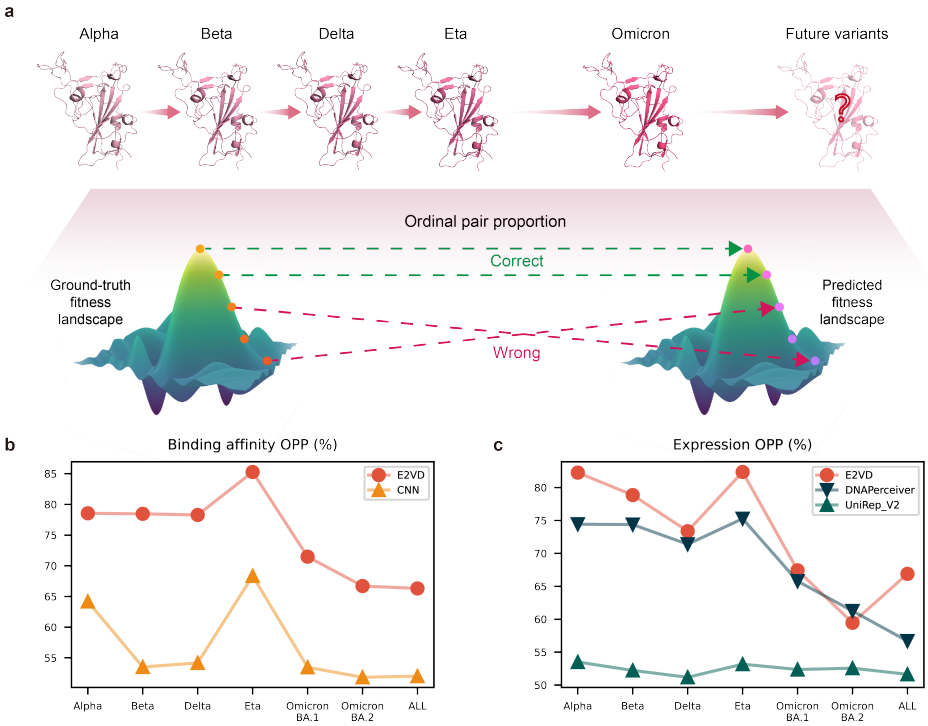

Fig. 5 **a**, To reasonably evaluate the generalization of different models, we propose **O**rdinal **P**air **P**roportion (OPP) based on the ground-truth fitness landscape and the predicted fitness landscape. It represents the proportion of correctly predicted pairwise mutations in all mutation combinations, reflecting the degree of chaos of the predicted fitness landscape. **bc**, In the binding affinity and expression prediction tasks, OPPs are calculated for six different lineages as well as a mixture of all lineages (referred to as All) to evaluate the generalization performance of different methods.

a mis-predicted pair, i.e., an inverse pair. OPP represents the proportion of correctly predicted pairwise mutations in all mutation combinations, reflecting the degree of chaos of the predicted fitness landscape. The larger the value of OPP, the less chaotic the predicted fitness landscape, which means that the model is suitable for predicting the relative order of virus variation drivers. Correspondingly, a small OPP value indicates that the model is difficult to predict the relative order.

Next, we compare the generalization capabilities of different models on virus variation drivers prediction tasks. We transfer the model trained with DMS dataset of wild-type RBD [13] to perform zero-shot inference on different lineages [14]. As different lineages contain multiple mutations relative to the wild-type, this can be considered a task of out-of-distribution generalization [68]. It is worth noting that we conduct generalization assessment on virus variation drivers prediction tasks except for antibody escape task. This is because the DMS dataset of antibody escape [15] only involves the wild-type RBD, resulting in a lack of data of other lineages.

As shown in Fig.5b, in the binding affinity prediction task, OPPs of CNN and E2VD are calculated for six different lineages as well as a mixture of all lineages

(referred to as All) to evaluate the generalization performance. We can see that E2VD comprehensively and significantly surpasses CNN, whether OPPs for a single lineage or the mixture of all lineages. Specifically, E2VD's prediction performance remains stable above 78% on the four lineages (Alpha, Beta, Delta, Eta) before Omicron (B.1.1.529), but there is a clear performance decay on the two sub-lineages (BA.1, BA.2) of Omicron. Omicron's previous lineages contain only few mutations of the RBD, but Omicron contains 15 mutations of the RBD [60], making generalization to the Omicron sub-lineages more challenging, resulting in the expected sharp drop in prediction performance. Taking all lineages together, CNN's OPP is improved by 14.3% by E2VD. The situation is similar in expression prediction task (Fig.5c), where E2VD outperforms the other two competitors on most lineages, with the OPP achieving an 10.2% improvement over DNAPerceiver [25] for the mixture of all lineages. Overall, E2VD surpasses state-of-the-art methods across the board on out-of-distribution lineages, demonstrating enhanced generalization performance.

2.5 E2VD is flexibly combined to perceive evolutionary trends

E2VD has demonstrated superior prediction performance and generalization in multiple virus variation drivers prediction tasks, making it an ideal tool for studying viral evolutionary trends. In this work, as shown in Fig.6a, we demonstrate a flexible combination scheme that assembles *in silico* deep mutational scanning with three different variation drivers predictors developed by E2VD into a E2VD-core pipeline for viral evolutionary trends prediction. Specifically, referring to the practice of wet-lab deep mutational scanning [13, 69], we first sequentially mutate a single-site amino acid in the RBD to other 19 types, and complete the *in silico* mutational scanning in the entire region site by site. Subsequently, the binding affinity, expression, and antibody escape induced by each single-site mutation are predicted by the corresponding predictors. Finally, we set different thresholds based on the importance of different variation drivers to viral viability to screen single-site variants for subsequent analysis.

We first demonstrate the ability of our E2VD-core pipeline to reveal the evolutionary fitness changes caused by single-site mutations. XBB.1.5 showed a significant growth advantage over XBB.1 during its rapid rise to become the dominant lineage [64, 70], even though they only differed by one S486P mutation. In order to explore the essential connection between this mutation and transmissibility, we perform *in silico* deep mutational scanning on 486 site of XBB.1 and subsequently predict the three key variation drivers (binding affinity, expression, and antibody escape) for all mutations (normalized regression predictions are presented in Fig.5b). We find that when the Ser on 486 site in XBB.1 is mutated to Pro (i.e. XBB.1.5), its binding affinity with human ACE2 surges and its expression is also improved to a certain extent. Beyond this, antibody escape exhibits a slight attenuation. The above prediction results are basically consistent with previous wet-lab measurements [13, 14, 70], allowing us to draw the following conclusion. XBB.1.5 achieves a surge in binding affinity while maintaining qualified antibody escape, which likely contributes to its significant growth advantage.

We next demonstrate how our E2VD-core pipeline recommends the high-risk mutation sites of a lineage. Previous studies [15, 16] estimated the high-risk mutation sites

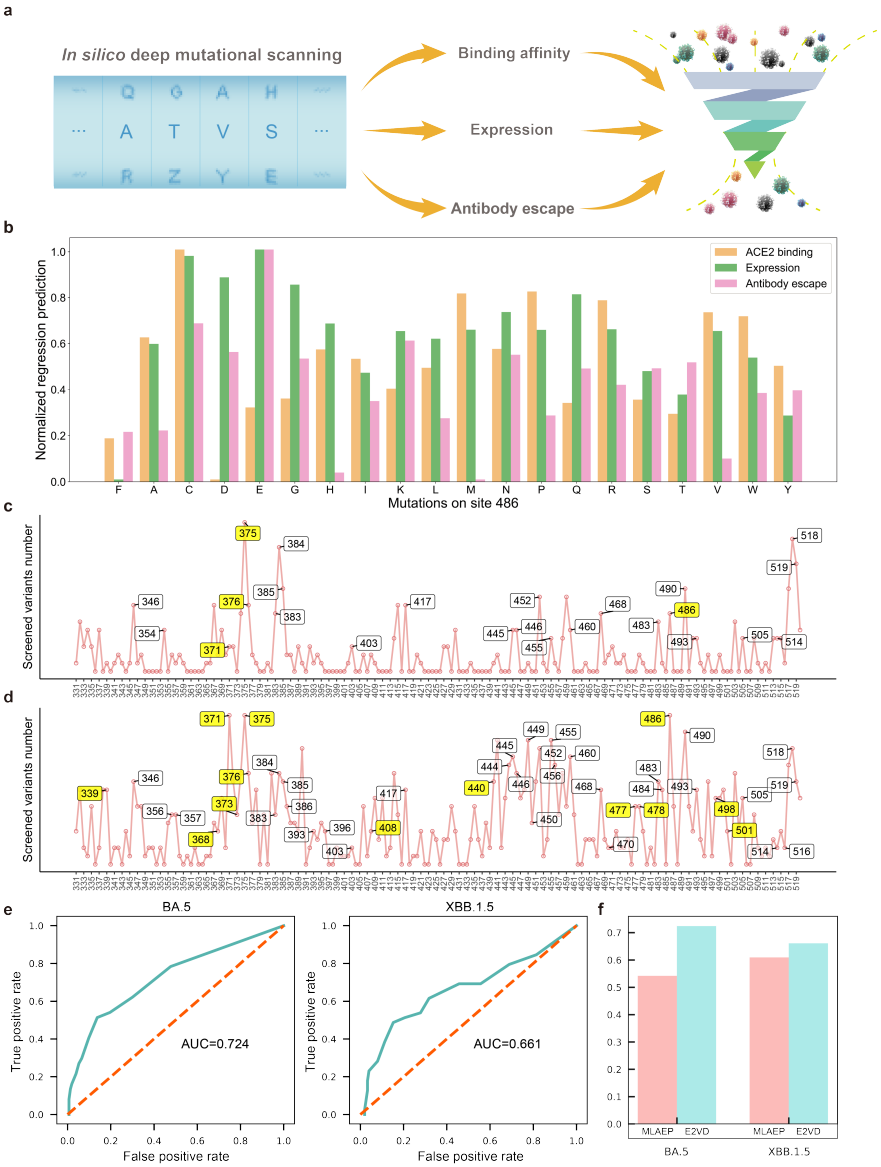

Fig. 6 **a**, A flexible combination scheme that assembles *in silico* deep mutational scanning with three different variation drivers predictors into a E2VD-core pipeline for virus evolutionary trends prediction. **b**, *In silico* deep mutational scanning on 486 site of XBB.1 is performed and subsequently the three key variation drivers (ACE2 binding, expression, and antibody escape) for all mutations are predicted. **cd**, The predicted high-risk mutation sites of BA.5 (c) and XBB.1.5 (d), where the "screened variants number" represents the remaining mutations at a site after screening. The higher the number of remaining mutations at a site, the higher the risk of this site. The sites marked in yellow are the ones that have been mutated in BA.5 and XBB.1.5 lineages in the real world. **e**, The Recriver Operating Characteristic (ROC) curves of BA.5 and XBB.1.5 lineages, in which the corresponding Area Under Curve (AUC) is calculated respectively. **f**, For the high-risk mutation site prediction task, the state-of-the-art method MLAEP is adopted for comparison on both BA.5 and XBB.1.5 lineages.

for RBD of Omicron BA.5 and XBB.1.5 lineages based on deep mutational scanning method, which systematically observed evolutionary trends through a statistical method for wet-lab measurements (Supplementary information, Fig.S3). Since these results are statistics on the results of wet-lab experiments, they can be considered as credible ground truth. Correspondingly, we systematically observe the mutational risk of each site through the proposed E2VD-core pipeline. Specifically, we first perform *in silico* deep mutational scanning on BA.5 and XBB.1.5, and then all single-mutation variants are separately predicted for ACE2 binding affinity, expression, and antibody escape, which are subsequently screened by different predictors. The high-risk mutation site distributions of screened single-mutation variants of BA.5 and XBB.1.5 are shown in Fig.6cd, where the "screened variants number" represents the remaining mutations (i.e. single-mutation variants) at a site after screening. The higher the number of remaining mutations at a site, the higher the risk of this site. The sites marked in yellow are the ones that have been mutated in BA.5 and XBB.1.5 lineages in the real world, while the remaining sites with high mutation tendency that are not marked are potential high-risk sites recommended by our pipeline. Taking the wet-lab measurements as ground truth, we plot the Receiver Operating Characteristic (ROC) curves for the two lineages and calculate the corresponding Area Under Curve (AUC) (Fig.6e). Han et al. proposed MLAEP [18] as a scoring function to predict the antigenic evolution trend of SARS-CoV-2 through *in silico* directed evolution. As shown in Fig.6f, the above-mentioned state-of-the-art method are used for comparison with our pipeline. We found that E2VD-core pipeline comprehensively surpasses MLAEP-core pipeline in two lineages, in which the AUC of MLAEP for BA.5 is improved by approximately 18.2% by E2VD, while the AUC of MLAEP for XBB.1.5 is improved by approximately 5.2%.

In summary, when E2VD is flexibly combined, it not only reveals the evolutionary fitness changes caused by single-site mutations, but also recommends high-risk mutation sites in different lineages, demonstrating the perception of virus evolutionary trends.

3 Discussion

Under the evolution-driven methodology, E2VD comprehensively and significantly surpasses state-of-the-art methods on three key virus variation drivers prediction tasks. Furthermore, we show that E2VD captures fundamental patterns of virus evolution. We first demonstrate the benefits of our proposed innovative modules and multi-task focal loss through extensive ablation studies. Second, dimensionality reduction visualizations are performed to demonstrate the discriminative ability of E2VD on different types of mutations. Finally, the experiments of rare beneficial mutations mining are designed to demonstrate that E2VD sensitively captures rare beneficial mutations that are critical for virus evolution.

The generalization ability of computational methods for predicting virus variation drivers is crucial to cope with intricate evolutionary trends. In order to get rid of the influence of the batch effects of wet-lab experimental data, we propose an indicator OPP that reflects the degree of chaos of the predicted fitness landscape to evaluate the

generalization performance of different models. When spanning to lineages carrying more mutations, E2VD outperforms state-of-the-art methods on different variation drivers prediction tasks, demonstrating enhanced generalization.

In this work, we demonstrate a flexible combination scheme in which *in silico* deep mutational scanning and three different variation drivers predictors are assembled into an E2VD-core pipeline. The above pipeline not only reveals the evolutionary fitness changes caused by single-site mutations, but also recommends high-risk mutation sites of different lineages, accompanied by comprehensively surpassing the prediction performance of state-of-the-art methods. It is worth noting that the predictors based on E2VD for different tasks are all retrained in this pipeline to update the model. This is to avoid the significant decrease in model generalization caused by the notable difference in lineages before and after Omicron.

E2VD provides an alternative to costly wet-lab experimental approaches for the analysis and prediction of virus evolutionary fitness in a unified, flexible, and interpretable way, holding great potential for improving the response speed to emerging viral infections. It should be emphasized that SARS-CoV-2 is studied as a typical case in this work because there are abundant deep mutational scanning data for it, but E2VD is not limited to specific viruses. The "few-site mutations" and "rare beneficial mutations" we summarized are general rules faced by all viruses in the process of evolution. Therefore, our evolution-driven model designed on this basis can potentially be extended to a variety of viruses when the corresponding data is ready, including HIV, MERS-CoV, etc. However, E2VD still has some limitations at this stage, including not supporting non-equal length mutation prediction and a notable attenuation of prediction performance when generalizing to Omicron lineages. Existing deep mutational scanning data of viral proteins often only involve amino acid substitutions, that is, data on amino acid insertions or deletions are very scarce, which makes it difficult to train deep learning models that support non-equal length mutation prediction. Facing the future, we can further improve model generalization performance from both data and model architecture. As the deep mutational scanning data of specific lineages with multiple mutations becomes increasingly abundant, integrating it into training dataset can naturally improve the generalization of the model. At the same time, the reasonable introduction of modules such as domain adaptation [71] is also expected to help the model generalize better. In addition, in this work, we predict the evolutionary trends of viruses based on the evolutionary fitness of viral proteins, without integrating higher-level viral properties directly, including pathogenicity and transmissibility, etc. Incorporating the above-mentioned high-level viral properties into the model is a potential next step to connect selective pressures within and outside the host.

4 Methods

4.1 Data

4.1.1 Benchmark dataset of binding affinity prediction task

For binding affinity prediction task, we adopt the deep mutational scanning (DMS) dataset of binding affinity from previous study [13]. The above work reported the

difference in log10K_d, i.e. apparent dissociation constant, relative to wild-type SARS-CoV-2 RBD. The raw data form is that one mutation corresponds to one $\Delta \log_{10} K_d$, so we preprocess the raw dataset to facilitate model training. First, based on the mutations, we substitute residues at corresponding positions on wild-type RBD sequence to obtain the mutated RBD sequences. Second, we convert $\Delta \log_{10} K_d$ to K_a *ratio* to get the training labels representing the binding strength as follows:

$$K_a \ ratio = 10^{\Delta \log_{10} K_d} \tag{1}$$

where K_a *ratio* represents the ratio of the binding strength between the mutated RBDs and human ACE2 relative to the wild-type. A value greater than 1 represents an increase in binding affinity, and a value less than 1 represents a decrease in binding affinity.

Finally, there are 3702 samples, including 3,388 negative samples (risk-free) with label less than 1, and 314 positive samples (risky) with label greater than 1. To evaluate the performance of E2VD on binding affinity task, 5-fold cross-validation training is adopted. The blind test set (n=54) keeps the same as previous study [20], which includes the same number of positive and negative samples. The prediction performance on the blind test set is the average of five trained models. The validation set (n=110) of each fold randomly samples the same number of positive and negative samples to avoid the imbalance of sample categories affecting the verification results, and the remaining samples form the training set (n=3,538) of each fold.

4.1.2 Benchmark dataset of expression prediction task

For expression prediction task, we adopt the DMS dataset of expression from previous study [13]. This work represented expression as $\Delta \log(MFI)$ relative to wild-type SARS-CoV-2 RBD. Similarly, we substitute residues at corresponding positions on wild-type RBD sequence to obtain the mutated RBD sequences, and calculate the training label as follows:

$$MFI \ ratio = e^{\Delta \log(MFI)} \tag{2}$$

where MFI *ratio* represents the ratio of the expression of the mutated RBDs relative to the wild-type. A value greater than 1 represents an increase in expression, and a value less than 1 represents a decrease in expression.

Finally, there are 3,781 samples, including 436 positive samples (risky) with labels greater than 1 and 3,345 negative samples (risk-free) with labels less than 1. To evaluate the performance of E2VD on expression prediction task, 5-fold cross-validation training is adopted. The blind test set (n=100) randomly sample the same number of positive and negative samples, and the prediction performance on it is the average of five trained models. The validation set (n=120) of each fold randomly samples the same number of positive and negative samples, and the remaining samples form the training set (n=3,561) of each fold.

4.1.3 Benchmark dataset of antibody escape prediction task

For antibody escape prediction task, we adopt the DMS dataset of wild-type RBD with different antibodies from previous study [15]. This work conducted deep mutational

scanning between 3,051 antibodies and the wild-type RBD, reporting antibody names, mutations on wild-type RBD, and the corresponding escape scores. In this work, we use the escape scores as training labels directly. According to the previous study [72], 0.4 is a relatively high escape score, so we assume that it is an appropriate threshold to differentiate between high and low antibody escape risk. Taking 0.4 as the threshold of escape risk, samples with scores greater than 0.4 are considered to have a high escape risk and the rest are considered to have low escape risk.

Finally, there are 57,740 positive samples (high risk) and 284,032 negative samples (low risk) in the dataset. To evaluate the performance of E2VD on antibody escape task, we randomly sample the same number of positive and negative samples as test set (n=9,600) and the remaining samples form the training set (n=332,172).

4.2 Protein language model customization

UniRef90 dataset consisting of 144 million raw protein sequences is adopted by us. The architecture of customized protein language model is similar to $BERT_{LARGE}$ [73], and it is pre-trained with only the objective of masked language modeling with the original masking probability of 15% [73]. We follow the widely recognized configurations and two-stage pre-training strategy of mainstream protein language models ProtTrans [41, 74]. The efficiency gain of two-stage pre-training strategy comes from optimizing resource utilization using shorter sequences, thereby speeding up the convergence of longer sequences [41, 74]. This strategy is adopted by us to shorten pre-training time, which is consistent with our motivation to rapidly deploy highly adaptable and customized PLMs to rapidly respond to emerging viral infections. During preprocessing, non-generic or unresolved amino acids are mapped to unknown (represented by X). In the first pre-training stage, the raw protein sequences are truncated with a length of 128, resulting in a training dataset of 459.3 million fragments with a maximum sequence length of 128. The protein language model is pre-trained with a learning rate of 0.0004, warm-up steps of 40,000, and epochs of 65. In the second pre-training stage, the raw protein sequences are truncated with a length of 256, resulting in a training dataset of 265.5 million fragments with a maximum sequence length of 256. The protein language model is continuously pre-trained with a learning rate of 0.0001 and epochs of 35. It is worth noting that we retain all truncated protein fragments for pre-training during the sequence truncation process, that is, no fragments are discarded. During pre-training, the Lamb optimizer and weight decay of 0.01 are adopted as ProtBert [41]. The detailed configuration is as follows: 24 layers, 1,024 hidden dimension, 16 attention heads, 64 local batch size, and 16,384 global batch size.

4.3 Architecture overview

The architecture of E2VD includes three parts: protein language model, local-global dependence coupling and multi-task focal learning. The local-global dependence coupling module consists of CNN for establishing local dependencies, and ContextPool-attention for establishing global dependencies. In the multi-task focal learning module, our model performs multi-task learning including classification and regression tasks under the supervision of the proposed multi-task focal loss.

Specifically, we denote $\mathcal{P}$ as protein language model, $\mathcal{L}$ as local dependence establishing, $\mathcal{G}$ as global dependence establishing, $\mathcal{C}$ as focal learning under classification task, and $\mathcal{R}$ as focal learning under regression task. Given an input protein sequence S, the protein language model extracts the residue-level features E of S as $E = \mathcal{P}(S)$. The extracted features are further refined by local-global dependence coupling module, concatenated and pooled into sequence-level feature F as follows:

$$\begin{cases} E_g = \mathcal{G}(E), \\ E_l = \mathcal{L}(E), \\ E = \mathrm{CONCAT}(E_g, E_l), \\ F = \mathrm{MaxPool}(E), \end{cases} \tag{3}$$

where E denotes residue-level features, E_g and E_l denote global and local residue-level features, F denotes sequence-level features, CONCAT represents concatenate, and MaxPool represents max pooling.

Subsequently, the focal learning under classification task predicts whether the variation drivers increase or not, and the focal learning under regression task predicts the specific values of variation drivers as follows:

$$\begin{cases} Cls = \mathcal{C}(F), \\ Reg = \mathcal{R}(F). \end{cases} \tag{4}$$

where Cls denotes the output of classification task, Reg denotes the output of regression task.

4.4 Sequence tokenizing and embedding

An input protein sequence is treated as a sequence of tokens (residues). The tokenizer maps every token to its corresponding index. For protein sequences with different lengths, the tokenizer uses the reserved token '[PAD]' to fill the sequence to the longest, and set the value of the corresponding attention mask to 0. The features of amino acids are extracted in the form of embeddings, i.e., vector representations. Specifically, let l be the longest length of sequences and d be the dimension of the embedding. Embeddings extracted by different protein language models have different dimension d. For our customized protein language model and ProtT5 (3B), d are 1024. For ESM2-650M, ESM2-3B and ESM2-15B, d are 1280, 2560 and 5120, respectively. The length of the tokenized sequence is l, and the shape of the embedding for each sequence is $d \times l$. After extracting embeddings, we conduct standardization to enlarge the difference between different sequences. In order to distinguish the impact of standardization and our proposed local-global dependence coupling module on prediction performance, we perform ablation experiments (Supplementary information, Table.S8).

Specifically, let $X_{d\times l}$ be the extracted embeddings, and N be the number of sequences. First, the embeddings are flattened into one dimension $X_{d\times l} \longrightarrow X^f_{1\times dl}$. For feature dimension $i \in [1, dl]$, the standardization is as follows:

$$\begin{cases} std_i = \mathrm{STD}_N(X_i^f), \\ mean_i = \mathrm{MEAN}_N(X_i^f), \\ \tilde{X}_i^f = \frac{X_i^f - mean_i}{std_i}, \end{cases} \tag{5}$$

where $\mathrm{STD}_N(\cdot)$ and $\mathrm{MEAN}_N(\cdot)$ denote the standard deviation calculation and the mean calculation among all the N embeddings respectively. After standardization, the flattened embedding $\tilde{X}^f_{1\times dl}$ is reshaped into the original dimension $\tilde{X}^f_{1\times dl} \longrightarrow \tilde{X}_{d\times l}$.

For binding affinity and expression prediction, the input is only the mutated RBD. For antibody escape prediction, the input involves RBD and antibody sequences. Considering that the antibody variable region includes heavy and light chains, we conduct standardization separately to the extracted embeddings of RBD, antibody heavy chain and antibody light chain as described above.

4.5 Local-global dependence coupling

We use ContextPool-attention and CNN to capture the global dependencies and the local dependencies of extracted embeddings, respectively.

4.5.1 ContextPool-attention

Context pooling [46] adopts an explicit way to learn context-aware token features by learning to pool neighboring features for each token. After that, self-attention between such pooled features can thus be context-aware and model high-level dependencies. Inspired by this, we set the learnable granularity to establish global dependencies of mutations based on seeking for motifs in protein sequences.

Specifically, a mapping function $m(\cdot)$ is learned to predict both the pooling weights $w \in \mathbb{R}^{n\times 1}$ and sizes $s \in \mathbb{R}^{n\times 1}$ for l input token features $\tilde{X}_{d\times l}$, i.e. $\{w, s\} = m(\tilde{X})$. Here, we implement $m(\cdot)$ using three convolutional layers with softmax function behind, and the output channel size of the last layer is set to 2 to generate the vectors w and s simultaneously. The normalized pooling size s_i is transformed into standard deviation of Gaussian mask $g^i \sim N(i, \sigma_i^2)$, where $\sigma_i = rn * s_i$, and r is a scalar empirically set as 0.1.

The final context pooling function multiplies the learned pooling weights w with Gaussian mask g^i for token $\tilde{x}_i$. Specifically, the context pooling function is calculated as follows:

$$y_i = f_{ave}(\tilde{X} \odot \gamma(w) \odot \gamma(g^i)) = \sum_{j=1}^{n} \tilde{x}_j \cdot w_j \cdot g_j^i, \tag{6}$$

where $y_i \in Y$ represents the context pooling features, f_{ave} denotes average pooling function, and $\gamma(\cdot)$ is broadcasting function for element-wise multiplication $\odot$.

4.5.2 CNN

We adopt a one-dimension convolution neural network to establish the local dependencies between the mutation and its neighboring residues. The CNN here consists of 3 convolution layers with Layer Normalization and uses leaky rectified linear unit

(Leaky ReLU) as activation function. We set the kernel size to 3 and the stride to 1 to ensure the shape of the output is the same as the input.

4.5.3 Feature coupling

The output of ContextPool-attention and CNN both have the same shape as the input, so we add a shortcut on both modules. The outputs of both scales are concatenated together to get the final extracted features of residues, and the feature of the whole sequence is calculated through max pooling on the residue dimension.

It is worth noting that, for antibody escape prediction, the embeddings of RBD, antibody heavy chain, and antibody light chain are input into independent local-global dependence coupling channel respectively to capture corresponding global dependencies and local dependencies. The features of each kinds of sequence are calculated through max pooling on the residue dimension, and the results are concatenated as the the final feature to predict antibody escape.

4.6 Multi-task focal learning

We propose multi-task focal learning module to alleviate the imbalance between positive and negative samples through multi-task learning manner, in which the fitting strength to hard-to-learn samples is increased. We adopt two focal losses for different training tasks, i.e. the focal loss for classification and the focal loss for regression.

For classification task, the training intensity of a certain category can be adjusted by assigning different weights to samples with different labels in focal loss for classification. Specifically, the calculation of the focal loss for classification is as follows:

$$p_t = \begin{cases} pred, label = 1, \\ 1 - pred, label = 0. \end{cases} \tag{7}$$

$$\alpha_t = \begin{cases} \alpha, label = 1, \\ 1 - \alpha, label = 0. \end{cases} \tag{8}$$

$$ClsFocalLoss = -\alpha_t (1 - p_t)^\gamma \log p_t, \tag{9}$$

where *pred* stands for the prediction of the classification model, *label* stands for the classification label. The penalty index γ indicates the punishment intensity for large error samples. The larger γ is, the lower the tolerance to wrong samples will be. The weight coefficient α can adjust the attention of loss function to positive and negative samples.

For regression task, we propose the focal loss for regression. Specifically, the calculation of the focal loss for regression is as follows.

$$\begin{cases} a_i = (score - target)^\gamma, \gamma = 1, 2, 3, ..., \\ \hat{a}_l = \frac{a_i}{\sum a_i}, \\ l_i = \hat{a}_l (score - target)^2, \\ RegFocalLoss = \sum_{i=1}^{B} l_i, \end{cases} \tag{10}$$

where *score* stands for the prediction of the model, *target* stands for the regression label, and $\hat{a}_l$ is the scaled error in case the gradient value is too small. The penalty index γ is the same as that in the focal loss for classification.

The final multi-task focal loss function is as follows:

$$FocalLoss = \mu * RegFocalLoss + ClsFocalLoss, \tag{11}$$

where μ denotes the weight of the regression loss. We set $\mu = 1$ in this work.

4.7 Evaluation metrics

4.7.1 Metrics for regression tasks

For regression tasks, we use Mean Squared Error (MSE) and Pearson's Correlation Coefficient (PCC). Specifically, let y_{pred} be the predicted K_a *ratio* and y_{label} be the experimental K_a *ratio*, the MSE and PCC are calculated as follows

$$\begin{cases} MSE = \frac{1}{N}\sum_{i=1}^{N}(y_{pred_i} - y_{label_i})^2, \\ PCC = \frac{\sum_{i=1}^{N}(y_{pred_i} - \overline{y_{pred}})(y_{label_i} - \overline{y_{label}})}{\sqrt{\sum_{i=1}^{N}(y_{pred_i} - \overline{y_{pred}})^2}\sqrt{\sum_{i=1}^{N}(y_{label_i} - \overline{y_{label}})^2}}, \end{cases} \tag{12}$$

where N is the number of samples.

4.7.2 Metrics for classification tasks

The metrics for classification tasks includes AUC, $Accuracy$, $F1 - Score$, $Precision$ and $Recall$ between label and prediction.

AUC, i.e. $AU - ROC$, indicates the probability that the predicted score of positive samples is higher than the one of negative samples. Specifically, let m^+ and m^- be the number of positive and negative samples respectively. D^+ and D^- denote the set of positive and negative samples respectively. AUC is calculated as follows:

$$AUC = 1 - \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}(\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))), \tag{13}$$

where the $f(\cdot)$ denotes the prediction function, and $\mathbb{I}(\cdot)$ values 1 if the input is $True$, otherwise it values 0.

For binary classification tasks, the predictions can be divided into four cases: TP if the prediction and label are both positive, TN if the prediction and label are both negative, FP if the prediction is positive and the label is negative, FN if the prediction is negative and the label is positive. Therefore, the other four metrics are calculated as follows:

$$\begin{cases} Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \\ Precision = \frac{TP}{TP+FP}, \\ Recall = \frac{TP}{TP+FN}, \\ F1 - Score = \frac{2*Precision*Recall}{Precision+Recall}. \end{cases} \tag{14}$$

4.7.3 Ordinal Pair Proportion

We propose a new metric, *Ordinal Pair Proportion*(OPP), to evaluate generalization performance of different models. Specifically, we first combine all the prediction results into binary pairs. For each pair, if the label order of the two predictions is consistent with the order of prediction, we call the pair *ordinal*, and the OPP is the proportion of the *ordinal* pairs in all pairs. Then, for the prediction set N, $n = |N|$, OPP is calculated as follows:

$$\begin{cases} OPP = \frac{|M|}{C_n^2}, \\ M = \{(i,j) | \forall i,j \in N, i < j, (label_i - label_j)(pred_i - pred_j) > 0\}. \end{cases} \tag{15}$$

4.8 Ablation studies

We conduct ablation studies to explore how our innovations specifically contribute to prediction performance, including protein language model ablation study, module ablation study, and loss function ablation study.

4.8.1 Protein language model ablation experiments

To evaluate the effectiveness of customized pre-training configuration on the prediction performance, the widely recognized pre-trained PLMs are adopted to compare with our customized one. Specifically, ProtT5 (3B) in the ProtTrans series [41], which achieved excellent performance at both the residue-level and protein-level, is selected. Among the ESM series, ESM-1v (650M) [42], which has excellent zero-shot prediction ability, and ESM-2 (15B) [30], which owns the largest parameter size, are selected. The antibody escape prediction task differs from the other two tasks in that ESM-2 (15B) is replaced by ESM-2 (650M) [30]. The training data of this prediction task is two orders of magnitude larger than the other two tasks, and the larger parameter size of the ESM model brings higher embedding dimension (greater than 2560). The above two facts results in terrible memory usage and training time, which prompts us to adopt ESM-2 (650M) with a smaller number of parameters.

4.8.2 Module ablation experiments

We conduct ablation experiments of local-global dependence coupling module and multi-task focal learning module using three different protein language models to demonstrate the generalizability of the results, including ESM-2(650M), ProtT5(3B) and our customized protein language model. The benchmark dataset of binding affinity prediction task is adopted and the experiment results can be found in Supplementary information (Table.S10,Table.S11, and Table.S12).

4.8.3 Loss ablation experiments

We conduct ablation experiments on multi-task focal loss to analyze the impact of specific focal loss on prediction performance. Specifically, we adopt the benchmark dataset of binding affinity prediction task and set four different combinations of loss

functions, including binary cross entropy loss paired with mean squared error loss, binary cross entropy loss paired with focal loss for regression, mean squared error loss paired with focal loss for classification, and the multi-task focal loss. The experiment results can be found in Supplementary information (Table.S13).

4.9 Dimensionality reduction visualization of different types of mutations

We adopt Principal Component Analysis (PCA) for feature dimensionality reduction, in which the features before and after local-global dependence coupling module are visualized. For binding affinity and expression prediction tasks, we choose the RBD embedding from protein language model as the feature before dependence coupling. For antibody escape prediction task, we concatenate the RBD embedding, antibody heavy chain embedding and antibody light chain embedding from protein language model as the feature before dependence coupling. The feature that passes through local-global dependence coupling module are used as the features after dependence coupling. Finally, we color the samples according to the their risk level.

4.10 Experiments to mine rare beneficial mutations

We select XBB.1.5 as the reference lineage, and randomly select 15 mutations with increased binding affinity based on DMS data of 22 real-world mutation sites within its RBD region. Besides, we randomly select 15 mutations with reduced binding affinity at each beneficial mutation site to form a test set with balanced positive and negative samples. The remaining samples are used as training data. We train two models with multi-task focal loss and BCE&MSE loss respectively and compare their prediction performance.

4.11 Generalization performance experiments

We calculate *Ordinal Pair Proportion*(OPP) metrics on binding affinity and protein expression prediction tasks to evaluate the generalization performance of different models. All models are trained using DMS data of wild-type RBD [13] and subsequently perform zero-shot inference on DMS data of multiple lineages [14].

We calculate OPP metrics on all samples from different lineages, as well as all samples within each lineage, including Alpha, Beta, Delta, Eta, Omicron BA.1 and Omicron BA.2. Specifically, samples within each lineage V is denoted as Seq_V. The samples of multiple lineages datasets Seq includes $Seq = \{Seq_{Alpha}, Seq_{Beta}, Seq_{Delta}, Seq_{Eta}, Seq_{Omicron\ BA1}, Seq_{Omicron\ BA2}\}$.

The OPP of all samples from different lineages can be calculated as follows:

$$\begin{cases} n = |Seq|, \\ M = \{(i,j) | \forall i,j \in Seq, i < j, (label_i - label_j)(pred_i - pred_j) > 0\}, \\ OPP = \frac{|M|}{C_n^2}, \end{cases} \tag{16}$$

where $pred_i$ and $label_i$ are the predictions and labels of binding affinity and protein expression respectively.

The OPP of all samples within each lineage V can be calculated as follows:

$$\begin{cases} n_V = |Seq_V|, \\ M_V = \{(i,j)|\forall i,j \in Seq_V, i<j, (label_i - label_j)(pred_i - pred_j) > 0\}, \\ OPP_V = \frac{|M_V|}{C_{n_V}^2}. \end{cases} \tag{17}$$

4.12 E2VD-core pipeline for virus evolutionary trends prediction

4.12.1 Training datasets of three virus variation drivers predictors

Since the selected lineages to be studied are all branches of Omicron, three DMS datasets are selected to avoid a significant decrease in model generalization caused by the notable difference in lineages before and after Omicron.

The binding affinity predictor is trained with the DMS dataset of Omicron BA.1 and Omicron BA.2 [14]. We normalize the provided $\log_{10} K_d$ values with that of Omicron BA.2 to get the training labels. Samples with labels that are greater than 1 are considered as positive samples, while those with labels that are less than 1 are considered as negative samples. Finally, there are 7,193 negative samples and 628 positive samples in the training dataset of binding affinity predictor.

The expression predictor is also trained with the samples of Omicron BA.1 and Omicron BA.2 in the DMS dataset from [14]. We normalize the provided expression scores with that of Omicron BA.2 to get the training labels. There are 7,182 negative samples and 639 positive samples in the training dataset of expression predictor.

The antibody escape predictor is trained with the DMS dataset of wild-type RBD with 3,051 antibodies from [15], i.e., the benchmark dataset of antibody escape prediction task, since the DMS dataset currently available for antibody escape only involves the wild-type RBD. There are 57,740 positive samples and 284,032 negative samples in the training dataset.

4.12.2 Evolutionary fitness changes of 486 site of XBB.1

To explore the difference between XBB.1 and XBB.1.5 in evolutionary fitness, we mutate the residue on site 486 of XBB.1 into other 19 residues. Subsequently, the binding affinity predictor, expression predictor, and antibody escape predictor are adopted to predict properties of each mutation. We adopt the regression output and conduct normalization within each kinds of property due to the difference of output range between predicted properties.

4.12.3 High-risk mutation sites prediction of BA.5 and XBB.1.5

To predict high-risk mutation sites of Omicron BA.5 and XBB.1.5, we take them as reference sequences, and enumerate all the variants with single-mutation through

in silico deep mutational scanning. Subsequently, the ACE2 binding affinity, expression and antibody escape of each variant are predicted. As to the antibody used during antibody escape prediction, we adopt BD57-0129 from previous work [15] as it achieves relatively strong neutralizing capability. Based on the predicted properties of variants, we set different thresholds of the probability of the classification output to perform variants screening. Based on previous studies [75–77], partial sacrifice in ACE2 affinity is acceptable, while antibody escape is more important for SARS-CoV-2 evolution. Therefore, we set the screening threshold for binding affinity low (0.25 here) and antibody escape moderately (0.5 here). Considering the importance of RBD expression level for the survival of variants, we set the screening threshold for expression higher (0.7 here). For the screened single-mutation variants, the number of remaining mutations at each site is considered as indicator of risk level. In other words, the greater the number of remaining mutations, the higher the risk of the site.

In this work, we use AUC to evaluate the ability to predict the potential high-risk mutation sites of different models. Specifically, we consider the wet-lab deep mutational scanning statistic results from [15, 16] as ground truth, and compare the predictions of E2VD-core pipeline or MLAEP-core pipeline with them. We set different threshold from the minimum to the maximum of evolutionary scores, and calculate the true positive rate (TPR) and false positive rate (FPR) with each threshold to draw ROC curve, and AUC is the area under the resulting ROC curve. For our model, we use the number of screened mutations at each site as evolutionary scores. For MLAEP, we use the difference of KL-divergence of each site before and after *in silico* evolution as evolutionary scores.

Data availability

UniRef90 for protein language model pre-training can be downloaded from https://www.uniprot.org/. The raw deep mutational scanning data of binding affinity, expression, and antibody escape prediction tasks are freely available from previous studies [13–15]. The processed data are publicly available at https://github.com/ZhiweiNiepku/E2VD.

Code availability

Relevant code and models are available at https://github.com/ZhiweiNiepku/E2VD.

References

[1] Kuiken, T., Fouchier, R., Rimmelzwaan, G. & Osterhaus, A. Emerging viral infections in a rapidly changing world. *Current opinion in biotechnology* **14**, 641–646 (2003).

[2] Luo, G. G. & Gao, S.-J. Global health concerns stirred by emerging viral infections. *Journal of medical virology* **92**, 399 (2020).

[3] Prokunina-Olsson, L. *et al.* Covid-19 and emerging viral infections: The case for interferon lambda. *Journal of Experimental Medicine* **217** (2020).

[4] Lu, R. *et al.* Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The lancet* **395**, 565–574 (2020).

[5] Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of sars-cov-2 entry into cells. *Nature reviews Molecular cell biology* **23**, 3–20 (2022).

[6] Shang, J. *et al.* Structural basis of receptor recognition by sars-cov-2. *Nature* **581**, 221–224 (2020).

[7] Walls, A. C. *et al.* Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell* **181**, 281–292 (2020).

[8] Krammer, F. Sars-cov-2 vaccines in development. *Nature* **586**, 516–527 (2020).

[9] Dong, Y. *et al.* A systematic review of sars-cov-2 vaccine candidates. *Signal transduction and targeted therapy* **5**, 237 (2020).

[10] Tao, K. *et al.* The biological and clinical significance of emerging sars-cov-2 variants. *Nature Reviews Genetics* **22**, 757–773 (2021).

[11] Graham, R. L. & Baric, R. S. Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission. *Journal of virology* **84**, 3134–3146 (2010).

[12] Yang, S. *et al.* Fast evolution of sars-cov-2 ba. 2.86 to jn. 1 under heavy immune pressure. *The Lancet Infectious Diseases* **24**, e70–e72 (2024).

[13] 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).

[14] 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).

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

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

[17] 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).

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

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

[20] 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).

[21] 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).

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

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

[24] 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).

[25] Stefanini, M., Lovino, M., Cucchiara, R. & Ficarra, E. Predicting gene and protein expression levels from dna and protein sequences with perceiver. *Computer Methods and Programs in Biomedicine* **234**, 107504 (2023).

[26] Martiny, H.-M., Armenteros, J. J. A., Johansen, A. R., Salomon, J. & Nielsen, H. Deep protein representations enable recombinant protein expression prediction. *Computational Biology and Chemistry* **95**, 107596 (2021).

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

[28] Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

[29] Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

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

[31] Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* 1–3 (2024).

[32] Pak, M. A. *et al.* Using alphafold to predict the impact of single mutations on protein stability and function. *Plos one* **18**, e0282689 (2023).

[33] Yin, R., Feng, B. Y., Varshney, A. & Pierce, B. G. Benchmarking alphafold for protein complex modeling reveals accuracy determinants. *Protein Science* **31**, e4379 (2022).

[34] Stevens, A. O. & He, Y. Benchmarking the accuracy of alphafold 2 in loop structure prediction. *Biomolecules* **12**, 985 (2022).

[35] McDonald, E. F., Jones, T., Plate, L., Meiler, J. & Gulsevin, A. Benchmarking alphafold2 on peptide structure prediction. *Structure* **31**, 111–119 (2023).

[36] Baer, C. F. Does mutation rate depend on itself. *PLoS biology* **6**, e52 (2008).

[37] Duffy, S. Why are rna virus mutation rates so damn high? *PLoS biology* **16**, e3000003 (2018).

[38] Loewe, L. & Hill, W. G. The population genetics of mutations: good, bad and indifferent (2010).

[39] Hie, B. L. *et al.* Efficient evolution of human antibodies from general protein language models. *Nature Biotechnology* **42**, 275–283 (2024).

[40] Zhang, Y. & Yang, Q. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* **34**, 5586–5609 (2021).

[41] 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).

[42] Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems* **34**, 29287–29303 (2021).

[43] Unsal, S. *et al.* Learning functional properties of proteins with language models. *Nature Machine Intelligence* **4**, 227–245 (2022).

[44] Heinzinger, M. *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics* **20**, 1–17 (2019).

[45] Gu, J. *et al.* Recent advances in convolutional neural networks. *Pattern recognition* **77**, 354–377 (2018).

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

[47] Chen, M. *et al.* Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics* **35**, i305–i314 (2019).

[48] Xu, W., Gao, Y., Wang, Y. & Guan, J. Protein–protein interaction prediction based on ordinal regression and recurrent convolutional neural networks. *BMC bioinformatics* **22**, 485 (2021).

[49] Zhu, H., Du, X. & Yao, Y. Convsppis: identifying protein-protein interaction sites by an ensemble convolutional neural network with feature graph. *Current Bioinformatics* **15**, 368–378 (2020).

[50] Zheng, L., Fan, J. & Mu, Y. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega* **4**, 15956–15965 (2019).

[51] Ahmed, A., Mam, B. & Sowdhamini, R. Deelig: A deep learning approach to predict protein-ligand binding affinity. *Bioinformatics and Biology Insights* **15**, 11779322211030364 (2021).

[52] Jones, D. *et al.* Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *Journal of chemical information and modeling* **61**, 1583–1592 (2021).

[53] Bork, P. & Koonin, E. V. Protein sequence motifs. *Current opinion in structural biology* **6**, 366–376 (1996).

[54] Bailey, T. L., Williams, N., Misleh, C. & Li, W. W. Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic acids research* **34**, W369–W373 (2006).

[55] Kulmanov, M. & Hoehndorf, R. Deepgoplus: improved protein function prediction from sequence. *Bioinformatics* **36**, 422–429 (2020).

[56] Wang, X., Schroeder, D., Dobbs, D. & Honavar, V. Automated data-driven discovery of motif-based protein function classifiers. *Information Sciences* **155**, 1–18 (2003).

[57] Zhang, F. *et al.* Deepfunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions. *Proteomics* **19**, 1900019 (2019).

[58] Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

[59] Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. *Focal loss for dense object detection*, 2980–2988 (2017).

[60] Cao, Y. *et al.* Omicron escapes the majority of existing sars-cov-2 neutralizing antibodies. *Nature* **602**, 657–663 (2022).

[61] Ma, W., Fu, H., Jian, F., Cao, Y. & Li, M. Immune evasion and ace2 binding affinity contribute to sars-cov-2 evolution. *Nature Ecology & Evolution* **7**, 1457–1466 (2023).

[62] Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods* **16**, 1315–1322 (2019).

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

[64] Parums, D. V. The xbb. 1.5 ('kraken') subvariant of omicron sars-cov-2 and its rapid global spread. *Medical science monitor: international medical journal of experimental and clinical research* **29**, e939580–1 (2023).

[65] Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics* **11**, 733–739 (2010).

[66] Chan, T. M. & Pătraşcu, M. *Counting inversions, offline orthogonal range counting, and related problems*, 161–173 (SIAM, 2010).

[67] Ajtai, M., Jayram, T., Kumar, R. & Sivakumar, D. *Approximate counting of inversions in a data stream*, 370–379 (2002).

[68] Hendrycks, D. *et al. The many faces of robustness: A critical analysis of out-of-distribution generalization*, 8340–8349 (2021).

[69] Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nature methods* **11**, 801–807 (2014).

[70] Yue, C. *et al.* Ace2 binding and antibody evasion in enhanced transmissibility of xbb. 1.5. *The Lancet Infectious Diseases* **23**, 278–280 (2023).

[71] Farahani, A., Voghoei, S., Rasheed, K. & Arabnia, H. R. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020* 877–894 (2021).

[72] Greaney, A. J., Starr, T. N. & Bloom, J. D. An antibody-escape estimator for mutations to the sars-cov-2 receptor-binding domain. *Virus evolution* **8**, veac021 (2022).

[73] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. *Bert: Pre-training of deep bidirectional transformers for language understanding* (2019).

[74] You, Y. *et al.* *Large batch optimization for deep learning: Training bert in 76 minutes* (2019).

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

[76] 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).

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

Acknowledgements

We thank Ming Li for his discussion of the protocol and 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), and R&D Program of Guangzhou Laboratory (Grant No.SRPG22-001).

Competing interests

The authors declare no competing interests.

Supplementary information

S1 Baseline models

Here, we reimplement several baseline models [18, 20, 23, 25–27] on binding affinity, protein expression, antibody escape, and high-risk mutation sites prediction tasks. For high-risk mutation sites prediction baseline model MLAEP [18], we conduct *in silico* directed evolution experiment using MLAEP as the scoring function. MLAEP adopt genetic algorithm to perform variants population evolution, and the difference of KL-divergence of each site between the original variant population and the evolved variant population can be considered as an indicator of the potential high-risk mutation sites. Therefore, we adopt the difference of KL-divergence before and after *in silico* evolution of each site as evolutionary scores to calculate AUC.

S2 Protein language model pre-training hyperparameters

Hyperparameter	Value
Dataset	UniRef90
Number of layers	24
Hidden layers size	1024
Number of heads	16
Masking probability	15%
Local batch size	64
Global batch size	16384
Optimizer	Lamb
Learning rate	0.0004/0.0001
Weight decay	0.01
Number of parameters	340 million

Table S1 Hyperparameters of customized protein language model

S3 Ablation experiments of different protein language models

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
ESM-2(15B)	0.9383	0.9037	0.9066	0.8769	0.9407	0.059	0.863
ProtT5(3B)	0.9281	0.8741	0.8746	0.8631	0.8889	0.079	0.818
ESM-1v	0.939	0.8911	0.8927	0.8806	0.9081	0.063	0.853
E2VD	0.9298	0.9111	0.9158	0.8743	0.963	0.058	0.87

Table S2 Results of different protein language models on binding affinity prediction task.

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
ESM-2(15B)	0.8932	0.81	0.8021	0.8276	0.784	0.0535	0.8479
ProtT5(3B)	0.9174	0.828	0.8188	0.8628	0.78	0.0576	0.8453
ESM-1v	0.9149	0.844	0.8479	0.8259	0.8728	0.0526	0.8551
E2VD	0.9264	0.85	0.8504	0.8428	0.86	0.053	0.8561

Table S3 Results of different protein language models on expression prediction task.

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
ESM-2(650M)	0.8737	0.783	0.8061	0.7287	0.9019	0.0425	0.6969
ProtT5(3B)	0.8596	0.7785	0.7862	0.7598	0.8146	0.0451	0.6735
ESM-1v	0.8641	0.7826	0.7867	0.7726	0.8048	0.0447	0.6798
E2VD	0.874	0.794	0.796	0.7881	0.8042	0.0432	0.697

Table S4 Results of different protein language models on antibody escape prediction task.

S4 Prediction performance of different models

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
NN_MM-GBSA	-	0.8041	-	-	-	0.2	0.79
CNN	0.6952	0.5556	0.206	0.96	0.1185	0.1361	0.6771
E2VD	0.9298	0.9111	0.9158	0.8743	0.963	0.058	0.87

Table S5 Results of different models on binding affinity prediction task.

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
UniRep_V2	0.62	0.62	0.3866	1	0.24	0.1169	0.7546
UniRep_V1	0.5496	0.50	0	0	0	0.2368	0.1875
DNAPreceiver	0.8254	0.642	0.4492	0.6491	0.368	0.0944	0.7209
E2VD	0.9264	0.85	0.8504	0.8428	0.86	0.053	0.8561

Table S6 Results of different models on expression prediction task.

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
RBD_AB_V4	0.7112	0.5494	0.2065	0.8635	0.1173	0.0864	0.3956
RBD_AB_V3	0.7644	0.59	0.3414	0.8673	0.2125	0.0756	0.4654
RBD_AB_V2	0.7772	0.5775	0.3017	0.869	0.1825	0.073	0.49
RBD_AB_V1	0.8391	0.6757	0.5625	0.8644	0.4169	0.052	0.6292
Escape_V2	0.6564	0.5633	0.3386	0.6977	0.2235	0.1005	0.2795
Escape_V1	0.8225	0.6497	0.5041	0.8627	0.356	0.0683	0.5629
E2VD	0.874	0.794	0.796	0.7881	0.8042	0.0432	0.697

Table S7 Results of different models on antibody escape prediction task.

S5 Ablation experiments for standardization

The feature standardization we perform amplifies the differences between the extracted sequence embeddings, which serves a similar purpose to the proposed local-global dependence coupling module. In order to distinguish the impact of feature standardization and the proposed local-global dependence coupling module on prediction performance, we perform ablation experiments on binding affinity benchmark datasets. It can be noticed that E2VD with only local-global dependence coupling module achieves performance close to the original E2VD. In contrast, the performance of E2VD with only feature standardization significantly decreases. The above results indicate that the local-global dependence coupling module is the main contributor to amplifying the differences of similar sequences, and feature standardization only contributes slightly to it.

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
w/o LG&Std	0.5885	0.50	0.6667	0.50	1	0.2356	0.1302
w/o LG[1]	0.7007	0.6852	0.6887	0.6832	0.6963	0.2349	0.2077
w/o Std[2]	0.9248	0.8556	0.8516	0.8734	0.837	0.093	0.7787
E2VD	0.9298	0.9111	0.9158	0.8743	0.963	0.058	0.87

[1]LG represents local-global dependence coupling.

[2]Std represents standardization.

Table S8 Ablation results for standardization on binding affinity prediction task.

S6 Convolution kernel size ablation experiments

To explore the best kernel size for convolution networks in our model, we train models with different kernel sizes on the binding affinity prediction task, including 1, 3, 5 and 7. Overall, the prediction performance of our model is close under different convolution kernel sizes. Considering that our model achieved the highest classification accuracy at kernel size of 3, we adopt this size in this work. We don't conduct further ablation studies on convolution hyperparameters as this is not the key point of our work.

Size	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
1	0.932	0.90	0.9083	0.8434	0.9852	0.0618	0.8652
3	0.9298	0.9111	0.9158	0.8743	0.963	0.058	0.87
5	0.938	0.90	0.9057	0.8559	0.963	0.0551	0.8756
7	0.9391	0.8926	0.8951	0.8746	0.9185	0.0532	0.8885

Table S9 Ablation results of convolution kernel sizes.

S7 Module ablation experiments

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
w/o MT&LG	0.7328	0.50	0	0	0	0.2219	0.385
w/o MT	0.8406	0.5741	0.2633	0.9714	0.1556	0.06	0.871
w/o LG	0.7007	0.6852	0.6887	0.6832	0.6963	0.2349	0.2077
Ours(E2VD)	0.9298	0.9111	0.9158	0.8743	0.963	0.058	0.87

Table S10 Module ablation results based on our customized protein language model.

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
w/o MT&LG	0.6829	0.50	0	0	0	0.2132	0.4535
w/o MT	0.7237	0.5926	0.3185	0.975	0.1926	0.0819	0.7935
w/o LG	0.6787	0.6593	0.6634	0.6596	0.6741	0.2269	0.3074
Ours(E2VD)	0.8826	0.8519	0.8544	0.8371	0.8889	0.08	0.802

Table S11 Module ablation results based on ESM-2(650M).

Model	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
w/o MT&LG	0.808	0.50	0	0	0	0.2433	0.2215
w/o MT	0.828	0.5667	0.2611	0.8714	0.1556	0.0759	0.8194
w/o LG	0.7396	0.6741	0.6962	0.6531	0.763	0.256	-0.0988
Ours(E2VD)	0.9281	0.8741	0.8746	0.8631	0.8889	0.079	0.818

Table S12 Module ablation results based on ProtT5(3B).

S8 Dimensionality reduction visualization

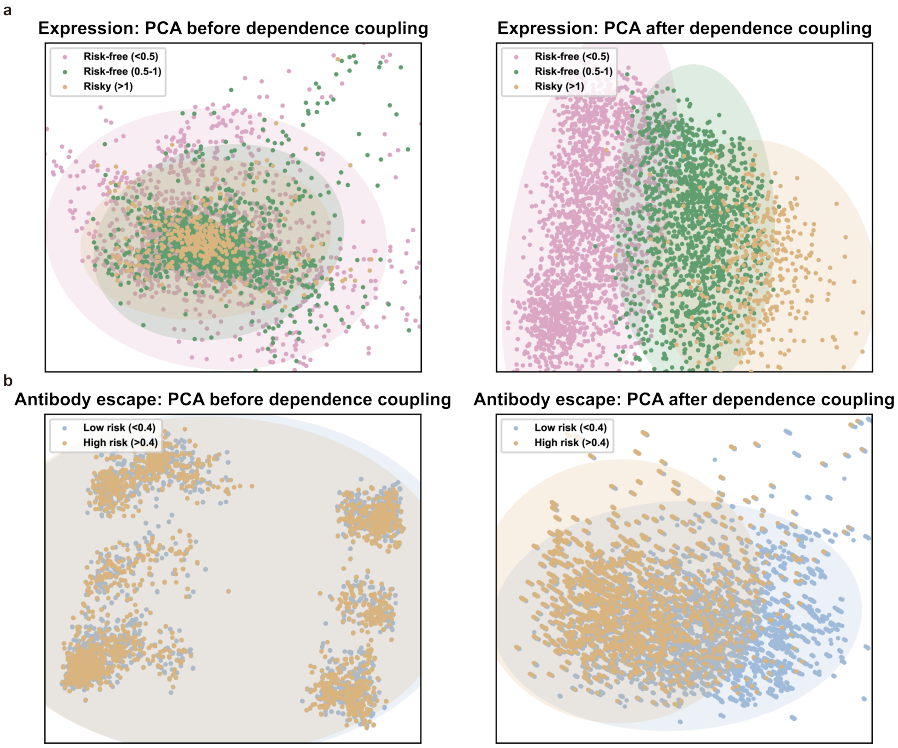

Fig. S1 Dimensionality reduction visualization of expression (a) and antibody escape (b) prediction tasks. In expression prediction tasks, three categories of mutations are presented, including Risk-free ($MFI\ ratio$<0.5), Risk-free (0.5<$MFI\ ratio$<1), and Risky ($MFI\ ratio$>1). In antibody escape prediction tasks, two categories of mutations are presented, including low-risk ($Escape\ score$<0.4) and high-risk ($Escape\ score$>0.4).

S9 Loss ablation experiments

Loss	AUC	Accuracy	F1 score	Precision	Recall	MSE	PCC
BCE&MSE	0.8406	0.5741	0.2633	0.9714	0.1556	0.060	0.871
BCE&RegFocalLoss	0.8859	0.5852	0.2893	1	0.1704	0.059	0.873
MSE&ClsFocalLoss	0.9369	0.8556	0.8492	0.8828	0.8222	0.055	0.879
Multi-task focal loss	0.9298	0.9111	0.9158	0.8743	0.9630	0.058	0.870

Table S13 Ablation results of different losses.

S10 Prediction performance of different losses on beneficial and harmful mutations

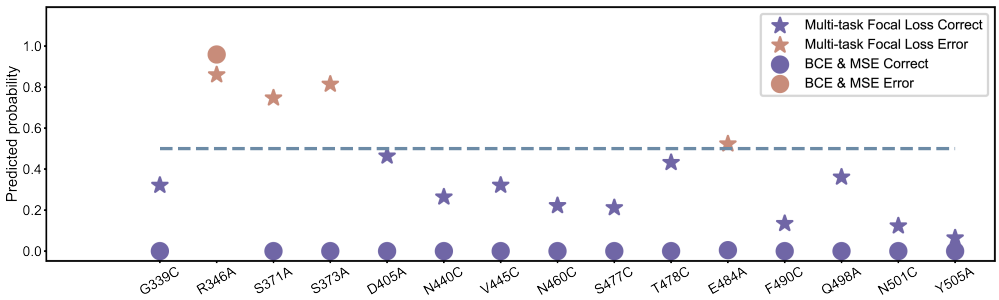

Fig. S2 Comparison of different losses on harmful mutations

Loss	AUC	Accuracy	F1 score	Precision	Recall
BCE&MSE	0.7467	0.5333	0.2222	0.6667	0.1333
Multi-task focal loss	0.7867	0.7667	0.7742	0.75	0.80

Table S14 Overall performance of different losses of beneficial and harmful mutations.

S11 Generalization performance of different models

Model	Alpha	Beta	Delta	Eta	Omicron BA.1	Omicron BA.2	All
CNN	0.6419	0.5351	0.5418	0.6835	0.5347	0.5184	0.5202
E2VD	0.7855	0.7846	0.7828	0.8528	0.7148	0.667	0.6631

Table S15 OPPs of six lineages and the mixture of them on binding affinity prediction task.

Model	Alpha	Beta	Delta	Eta	Omicron BA.1	Omicron BA.2	All
DNAPreceiver	0.7443	0.7438	0.7139	0.7527	0.6578	0.6125	0.5668
UniRep_V2	0.535	0.5222	0.5118	0.5317	0.5236	0.5256	0.5163
E2VD	0.8227	0.7886	0.7337	0.8237	0.6743	0.5946	0.6689

Table S16 OPPs of six lineages and the mixture of them on expression prediction task.

S12 Prediction performance of high-risk mutation sites

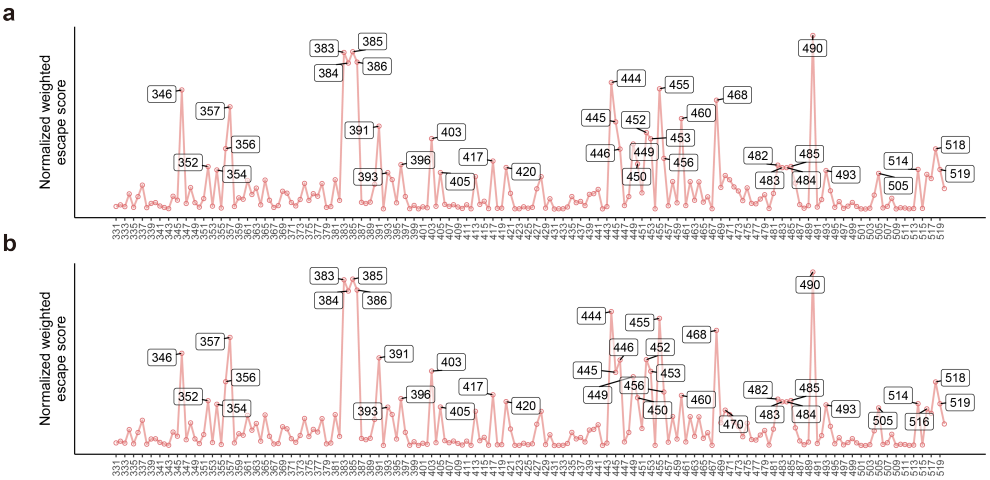

Fig. S3 Ground truth high-risk mutation sites of BA.5 and XBB.1.5.

Model	BA.5	XBB.1.5
MLAEP	0.542	0.6094
E2VD	0.724	0.661

Table S17 AUCs of predicted high-risk mutation sites on BA.5 and XBB.1.5.

S13 Model hyperparameters

Hyperparameter	Value
Learning rate	2e-5
Number of epochs	150
γ	1
β	0.998
Dropout rate	0.6

Table S18 Hyperparameters of E2VD for binding affinity prediction

Hyperparameter	Value
Learning rate	2e-5
Number of epochs	150
γ	1
β	0.99
Dropout rate	0.6

Table S19 Hyperparameters of E2VD for expression prediction

Hyperparameter	Value
Learning rate	1e-4
Number of epochs	70
γ	1.2
β	0.88
Dropout rate	0.6

Table S20 Hyperparameters of E2VD for antibody escape prediction