Evolution-guided large language model is a predictor of virus mutation trends

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, Shenzhen, 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

Emerging viral infections, especially the global pandemic COVID-19, have had catastrophic impacts on public health worldwide. The culprit of this pandemic, SARS-CoV-2, continues to evolve, giving rise to numerous sublineages with distinct characteristics. The traditional post-hoc wet-lab approach is lagging behind, and it cannot quickly predict the evolutionary trends of the virus while consuming high costs. Capturing the evolutionary drivers of virus and predicting potential high-risk mutations has become an urgent and critical problem to address. To tackle this challenge, we introduce ProtFound-V, an evolution-inspired deep-learning framework designed to explore the mutational trajectory of virus. Take SARS-CoV-2 as an example, ProtFound-V accurately identifies the evolutionary advantage of Omicron and proposes evolutionary trends consistent with wet-lab experiments through *in silico* deep mutational scanning. This showcases the potential of deep learning predictions to replace traditional wet-lab experimental measurements. With the evolution-guided large language model, ProtFound-V

presents a new state-of-the-art performance in key property predictions. Despite the challenge posed by epistasis to model generalization, ProtFound-V remains robust when extrapolating to lineages with different genetic backgrounds. Overall, this work paves the way for rapid responses to emerging viral infections, allowing for a plug-and-play approach to understanding and predicting virus evolution.

1 Main

Since 2000, numerous emerging viral infections [1, 2], especially COVID-19 [3], have had a profound global impact. 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. However, a concerning aspect is that SARS-CoV-2 continues to evolve, with mutations accumulating and recombination processes contributing to increased diversity [8, 9]. The sub-lineages of Omicron, in particular, remain prevalent and continue to cause breakthrough infections.

The increasing frequency of emerging viral infections demands rapid human responses. Understanding the evolution mechanism and predicting the direction of viral evolution is a potent approach to tackle these infections at their root. Artificial intelligence technology, particularly large language models (LLMs) based on Transformer [10], has shown promising applications in various life science fields, including proteins [11, 12] and molecules [13–15]. Inspired by this progress, our aim is to create a dedicated large language model tailored specifically for studying virus evolution. However, direct transfer of LLMs to investigate viral evolutionary mechanisms faces significant challenges due to the unique traits of virus evolution. Despite viruses having a high mutation rate during their evolution, only a small number of offspring acquire mutations that are beneficial to themselves. Consequently, virus variants typically exhibit few-site mutations, and beneficial mutations are far fewer than harmful mutations in the final mutational outcomes.

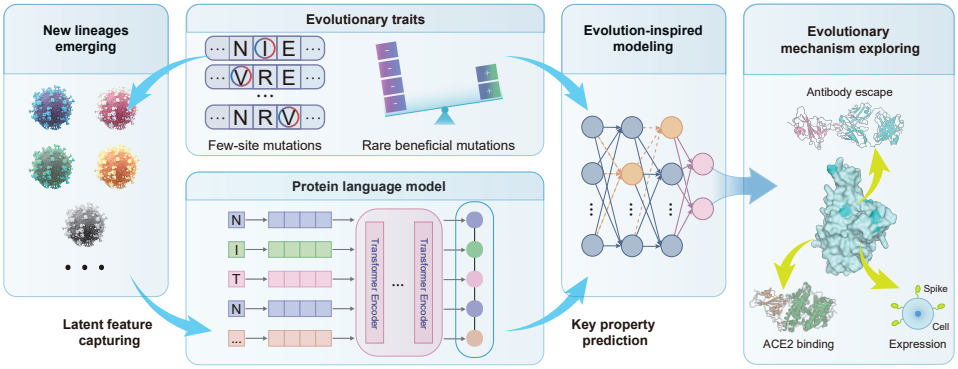

Fig. 1 The proposed framework, ProtFound-V, for exploring the evolutionary mechanisms of viruses. Viral evolution has resulted in the emergence of various lineages, accompanied by unique evolutionary traits. The protein language model is pretrained to capture the latent features of virus. Under evolution-inspired modeling, the key properties are comprehensively considered for the exploration of evolutionary mechanisms, including the identification of evolutionary advantages and the prediction of evolutionary trends. In this work, we take SARS-CoV-2 as the study case, so the three important properties are adopted, namely ACE2 binding, expression, and antibody escape.

In our previous work [16], we proposed a pipeline for SARS-CoV-2 mutation simulation that approximates a lineage through high-throughput variant generation and screening. In this work, we go one step further and develop an evolution-inspired framework, ProtFound-V (**Prot**ein **Found**ation Model for **V**irus), for viral property prediction (Fig.1). With the modeling challenges caused by evolutionary traits greatly alleviated, the prediction of evolutionary trends consistent with wet-lab experiments is achieved. Specifically, ProtFound-V addresses two challenges posed by the evolutionary nature of viruses. Firstly, we introduce a local-global feature coupling mechanism that captures amino acid interactions at different scales, enabling the detection of the subtle effects of few-site mutations. Secondly, we design a multi-task focal loss function to identify rare beneficial mutations. By combining evolution-inspired modeling with large language models, ProtFound-V effectively learns the evolutionary rules of viruses, enabling the prediction of key properties and evolutionary trends. We demonstrate the efficacy of ProtFound-V using SARS-CoV-2 as an example, analyzing its evolutionary advantages and predicting evolutionary trends based on three key properties: ACE2 binding, expression, and antibody escape. ProtFound-V provides not only insights into the leap-forward variation of SARS-CoV-2, from its wild type to Omicron, but also accurately predicts its evolution trends. As a powerful tool to study virus evolution, ProtFound-V can facilitate rapid responses to emerging viral infections by aiding in vaccine and drug design.

2 Results

2.1 Evolution-inspired architecture of ProtFound-V

The evolution-inspired architecture of ProtFound-V consists of four modules as shown in Fig.2a, including protein language model pretraining, feature extracting, local-global feature coupling, and multi-task learning. Specifically, the first module is the pretraining of protein language model, in which two-step training strategy is adopted to promote convergence and improve efficiency of the training process (Fig.2b). After the self-supervised learning on approximately 144 million raw protein sequences, predicted amino acid probabilities of each position of protein sequences are mapped. The second module is feature extracting of SARS-CoV-2 receptor-binding domain (RBD) or antibody sequences. For ACE2 binding and RBD expression prediction, only the mutated RBDs are taken as input. However, for antibody escape prediction, both mutated RBDs and antibodies are input because there are multiple antibodies. The embeddings of each amino acid in the mutated RBD or antibody are extracted in this module, and they are pooled through length to obtain fixed-size protein-level embedding in the next module, called local-global feature coupling. The few-site mutations lead to an extremely high degree of similarity between virus variants, which makes it difficult to learn the effects of few-site mutations. While a mutation affects nearby amino acids, it may also interact with distant amino acids. Therefore, we design the local-global feature coupling module to fuse the effects of mutations on amino acid interactions at different scales, thereby amplifying the subtle effects of few-site mutations. Convolutional Neural Network (CNN) [17] and ContextPool-attention [18] are used to mine the interactions of nearby amino acids (local feature) and distant amino

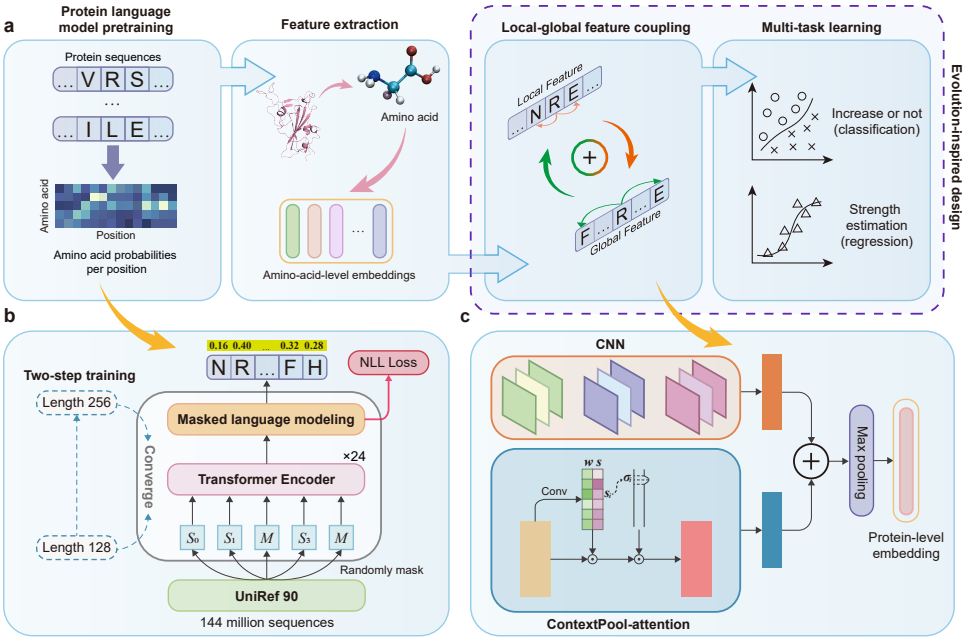

Fig. 2 The evolution-inspired architecture of ProtFound-V. a, The four modules of ProtFound-V, including Protein language model pretraining, Feature extraction, Local-global feature coupling, and Multi-task learning. Inspired by evolutionary traits, local-global feature coupling mechanism and multi-task focal loss are designed to fit the characteristics of virus mutations. b, The pretraining process of our protein language model on 144 million raw protein sequences, in which a two-step training strategy is adopted to achieve better convergence. NLL Loss represents negative log-likehood loss function, S_n represents the split token, and M represents masked token. Through the masked language modeling, the probability distribution of amino acids at each position are mapped. c, The coupling mechanism of local feature and global feature, in which CNN is used to extract the interactions of near-range amino acids, and ContextPool-attention is adopted to extract the interactions of long-range amino acids (see methods). The interaction fusion at different scales amplifies subtle effects of few-site mutations, thereby facilitating the identification of high-risk variants. Conv represents convolution, and "+" represents concatenation.

acids (global feature), respectively. The fused protein-level embeddings are fed into the fourth module, multi-task learning. This module aims to solve the hard sample mining problem caused by the rarity of beneficial mutations, which is a difficult problem often encountered in the field of deep learning [19, 20]. Focal Loss for object detection [21] achieves better results in classification under class imbalance, but a similar loss function is lacking in regression. To fill this gap, the regression focal loss function is proposed, which puts more focus on difficult-to-learn samples. Furthermore, considering the high correlation between classification and regression tasks of a certain property, we introduce multi-task learning [22, 23] to utilize useful information between these two related tasks to improve prediction performance. In summary, ProtFound-V solves the virus-evolution-induced problem of capturing the subtle effects of few-site

mutations and rare beneficial mutations by leveraging large protein language models, which supports the exploration of virus evolution mechanism.

2.2 Local-global feature coupling reveals the subtle effects of few-site mutations

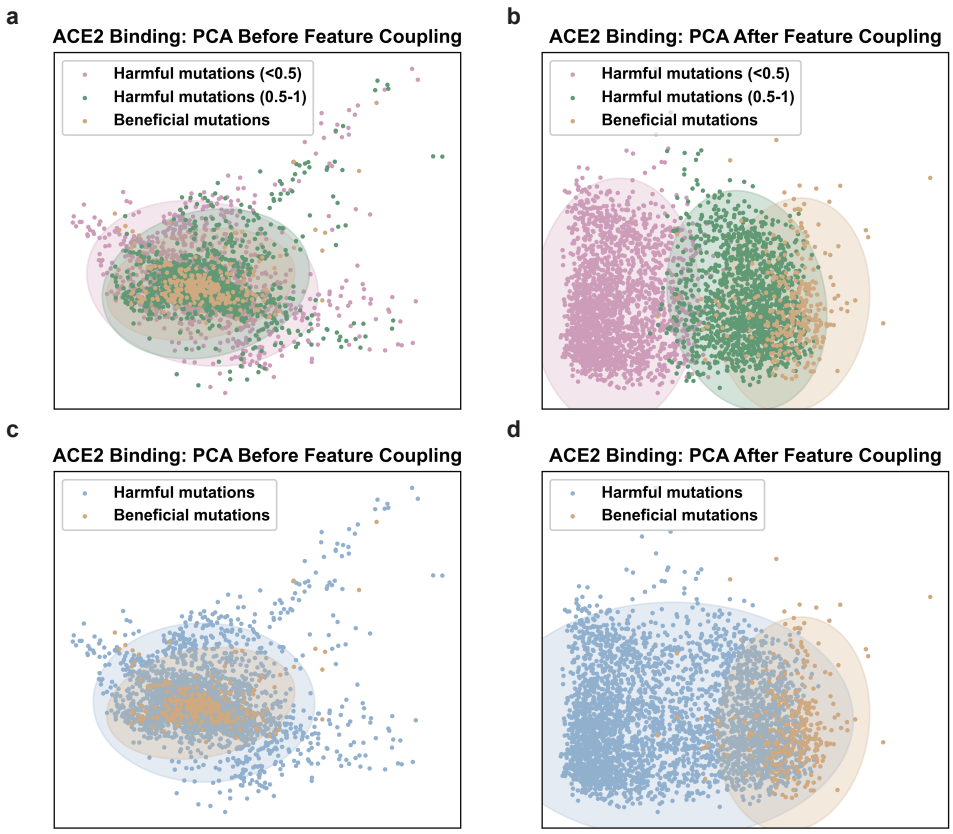

Fig. 3 The comparison of dimension reduction visualization on the embeddings of mutated RBDs with or without local-global feature coupling. When the variants are divided into two categories, i.e. beneficial mutations and harmful mutations, the visualization before feature coupling (a) and after feature coupling (b). When the variants are divided into three categories, i.e. beneficial mutations, harmful mutations with further division, the visualization before feature coupling (c) and after feature coupling (d).

As the target of virus entry into human cells, the binding of SARS-CoV-2 RBD to human ACE2 is crucial, which determines the course of infection. To quantify the ACE2 binding strength, most of the existing methods [24–28] need to preform time-consuming molecular dynamics simulations to obtain trajectories between the mutated

RBDs and ACE2, and then evaluate the binding strength. In contrast, ProtFound-V can rapidly assess and predict the ACE2 binding strength of variants in an efficient manner. Here, due to the importance of ACE2 binding for infection, we take ACE2 binding prediction as an example to demonstrate the advantage of ProtFound-V in mining evolutionary trajectories.

Although viruses usually have a high mutation rate during evolution, only a few offspring acquire beneficial mutations. This leads to a very similarity between the mutated RBD sequences, that is, there are only a few residue differences. Therefore, the initial embeddings extracted by ProtFound-V are too resembling to be distinguished, which makes it difficult for the neural network to learn the subtle effects of few-site mutations. The proposed local-global feature coupling is capable of exploit the interactions between amino acids at different scales (Fig.2c). CNN extracts the local dependencies, i.e., contextual features between residues of mutated RBDs. The ContextPool-attention, an improved version of standard self-attention[10] for efficient representation learning, is adopted to capture long-range dependencies of amino acids. It models long-range interactions with adapting attention granularity for each token instead of assuming a fixed granularity to construct the query and key vectors for individual tokens as in self-attention mechanism, thus better models complex dependencies at higher levels. Followed by the feature coupling, the local and global information can be fused to represent subtle difference among mutated RBDs. The max-pooling operations are adopted on the amino-acid-dimension to get the protein-level embedding of the whole RBD sequence, which is subsequently used for multi-task learning.

To demonstrate the impact of local-global feature coupling mechanism, we perform the dimensional reduction visualization on the embeddings of mutated RBDs using Principal Component Analysis (PCA). Previous work [29] used deep mutational scanning (DMS) to systematically observe the effects of single-site mutations to the SARS-CoV-2 RBD on ACE2 binding and expression. Therefore, we obtain the sequences of mutated RBDs with single-site mutations based on this wet-lab data for dimensional reduction. It can be seen in Fig.3 that positive and negative samples, i.e., binding-affinity-increased variants and binding-affinity-decreased variants, are clearly distinguished whether in the form of three-category or two-category after fusion of global and local dependencies. The results demonstrate that the subtle effects of single-site mutations are accurately identified by mining and fusing interactions between nearby and distant amino acids. We can conclude that local-global feature coupling reveals subtle effects of few-site mutations through fusion of interactions at different scales.

2.3 Multi-task focal loss captures rare beneficial mutations

As an RNA virus, SARS-CoV-2 has a high mutation rate, but most of the mutations are harmful, and only a small number of mutations can give the virus a stronger selection advantage. Therefore, this can be translated into a severe class imbalance problem for deep learning. The effects of mutations on key properties of viruses can be defined as two types of related tasks for prediction, i.e. classification (increase or not)and regression (the specific value). Leveraging useful information between the two related tasks to improve model generalization, the multi-task learning architecture

is designed to improve prediction performance. Inspired by Focal Loss [21] designed for classification, we propose a hard sample mining loss function for regression. The classification-oriented focal loss and regression-oriented focal loss are weighted summed as a whole, called multi-task focal loss, to improve the performance of hard sample mining.

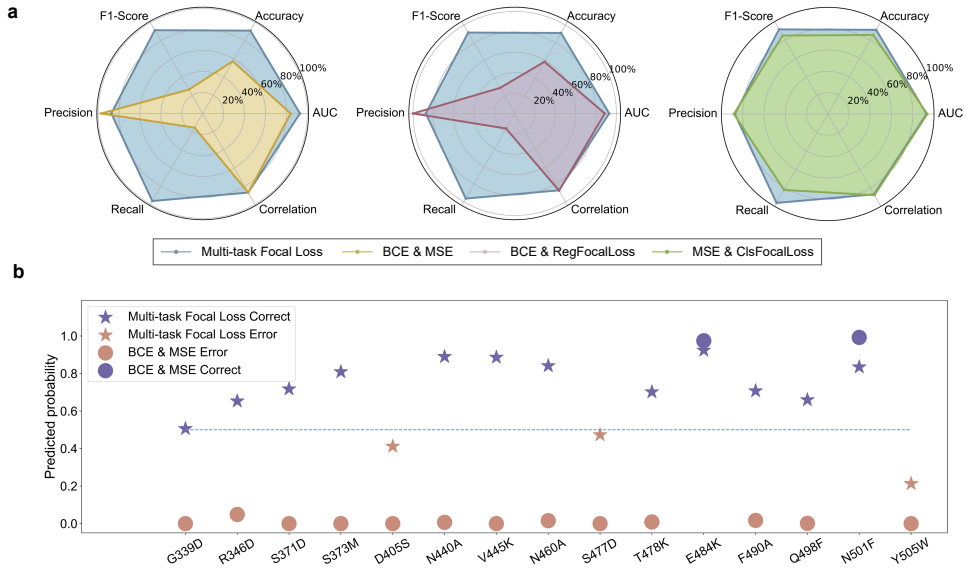

Fig. 4 a, Comprehensive ablation experiments between our multi-task focal loss and other losses. Six indicators are adopted, including AUC (Area Under Curve), Accuracy, F1-Score, Precision, Recall, and Correlation. b, Prediction performance of our multi-task focal loss in capturing high-risk mutations. We select 15 mutations that lead to increased binding strength from deep mutational scanning dataset. Multi-task focal loss with 80% accuracy (12 out of 15 mutations are predicted correctly) is significantly better than the original loss (BCE&MSE) with about 13% accuracy (13 out of 15 mutations are predicted wrongly).

To demonstrate the improvement of our multi-task focal loss in hard sample mining, we use ACE2 binding prediction task for ablation experiments and case studies. Here, deep mutational scanning dataset [29] is adopted for ACE2 binding prediction, which has the obvious imbalance between positive and negative samples, i.e., the binding-affinity-increased mutations and binding-affinity-decreased mutations are usually differ by nearly an order of magnitude. For binding affinity prediction between mutated RBDs and human ACE2, multilayer perceptron (MLP) is adopted here to predict the association constant (K_a) of variants relative to wild type RBD, i.e. K_a *ratio*, to judge whether the mutation has increased the binding strength. K_a *ratio* values greater than one represent increased binding affinity and values less than one represent decreased binding affinity. Therefore, a classification task (increase or not) and a regression task (the K_a *ratio*) are combined here to drive the multi-task learning.

For classification tasks and regression tasks, the most common loss functions are Mean Squared Error (MSE) Loss and Binary Cross Entropy (BCE) Loss. In order to explore the respective roles of Classification Focal Loss (ClsFocalLoss) and Regression Focal Loss (RegFocalLoss), we construct three types of multi-task loss, including BCE&MSE, BCE&RegFocalLoss, and MSE&ClsFocalLoss. As shown in Fig.4a and Table S1, we compare our multi-task focal loss with the three multi-task losses on various indicators. Due to the lack of ability to mine hard samples, BCE&MSE performs very poorly on the three key indicators, Accuracy, F1-Score, and Recall. After replacing MSE with RegFocalLoss, all six metrics for classification and regression tasks improve slightly. If we replace BCE with ClsFocalLoss, all six metrics improve significantly, especially on Accuracy, F1-Score, and Recall, i.e. the three key indicators. Our multi-task focal loss has the best performance on the three key metrics, achieving 91.11% for Accuracy, 91.58% for F1-Score, and 96.30% for Recall. The ablation experiment proves that the synergistic use of ClsFocalLoss and RegFocalLoss can effectively improve the performance of ProtFound-V in mining hard samples, i.e. accurately capturing rare beneficial mutations.

As shown in Fig.4b, referring to the mutation sites of the recently popular sublineage XBB.1.5, we select 15 mutations with increased binding strength at corresponding sites in the deep mutational scanning dataset to test the ability of our multi-task focal loss to identify high-risk mutations. Multi-task focal loss achieves excellent prediction performance, correctly predicting 12 out of 15 high-risk mutations, i.e. 80% accuracy. In contrast, the prediction performance of BCE&MSE lags far behind, only correctly predicting 2 out of 15 high-risk mutations, i.e. about 13% accuracy. This case study demonstrates that our multi-task focal loss efficiently captures rare beneficial mutations, i.e. high-risk mutations.

2.4 ProtFound-V showcases a new state-of-the-art and robust performance

2.4.1 Binding affinity prediction on two benchmark datasets

With the evolution-inspired architecture, ProtFound-V is competitive in high-risk mutation prediction. Here, ProtFound-V is tested on a single-site mutation benchmark dataset and a multi-site mutation benchmark dataset to demonstrate its advantages over other models. The deep mutational scanning dataset of wild type [29] is adopted as the single-site mutation benchmark dataset. The deep mutational scanning dataset of wild type [29], four VOCs (Alpha, Beta, Delta, Omicron), and one VOI (Eta) [30] are integrated to form the multi-site mutation benchmark dataset. The models involved are NN_MM-GBSA [28], BiLSTM [31], CNN [32], ProtFound-V with one-hot as encoder, and ProtFound-V with other six protein language models as encoders. Specifically, the compared pretrained protein language models here are SeqVec [33], ProtTrans T5 [34], ESM-1b [35], ESM-1v [36], ESM-2 with 650 million and 15 billion parameters [11].

As shown in Table S2 and Fig.5a, ProtFound-V achieves state-of-the-art performance on single-site mutation benchmark dataset and excels in key indicators for classification and regression. ProtFound-V with the protein language model as the

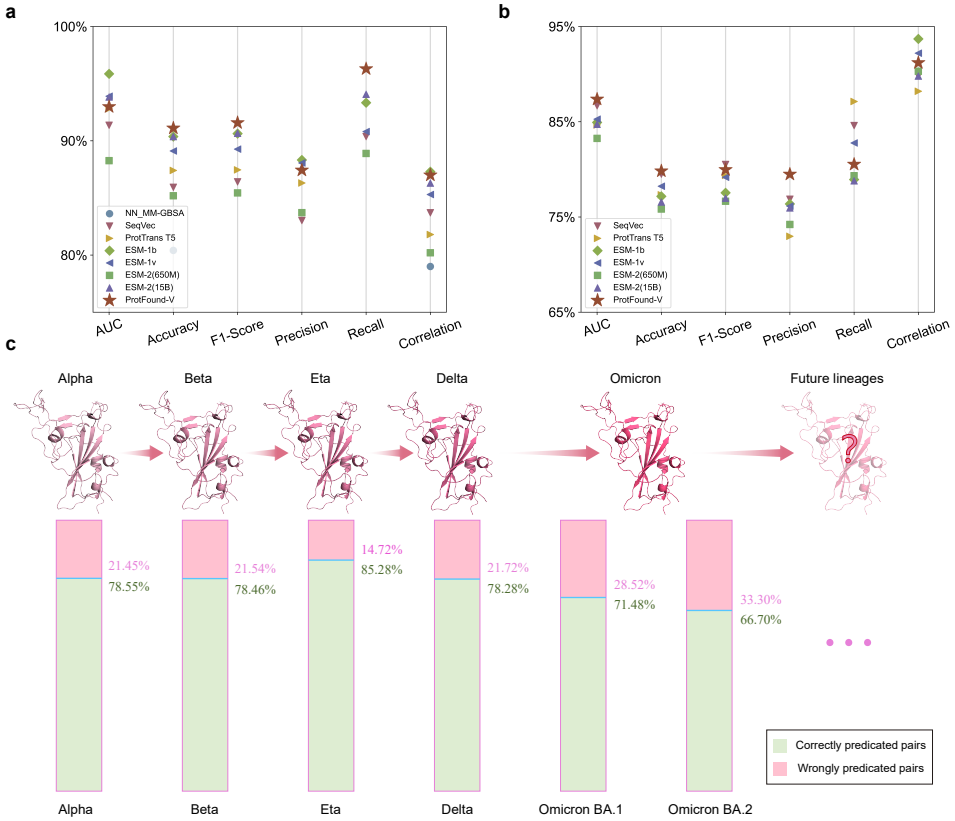

Fig. 5 a, The prediction performance on the single-site mutation benchmark dataset. b, The prediction performance on the multi-site mutation benchmark dataset. Six indicators are adopted, including AUC (Area Under Curve), Accuracy, F1-Score, Precision, Recall, and Correlation. c, Robustness testing on multiple lineages under different genetic backgrounds. The proposed metric *Ordinal Pair Proportion* (*OPP*) here aims to demonstrate the proportion of any pair of mutated RBDs that is correctly predicted.

encoder is overall superior to other methods, which means that the protein language models have strong abilities to extract protein sequence information. Compared with other protein language models, our protein language model achieves the best performance even when the model parameters are much smaller than them, which proves that our model pretrained from scratch has ideal feature extraction ability. Overall, ProtFound-V using our protein language model as an encoder has significantly improved compared to the baseline method (NN_MM-GBSA), in which the *Accuracy* is boosted by more than 10% and the correlation exceeds baseline method by 8%.

Previous works [37, 38] have implicated that epistasis drives ongoing major evolution of SARS-CoV-2 where the effect of a mutation is determined by its genetic background. New variants are constantly emerging, creating an ever-changing genetic background. Compared with ancestral Wuhan-Hu-1 background, the same mutations

of subsequent variants may cause different effects. Therefore, it is necessary to incorporate more variants under different genetic backgrounds for model training, which helps ProtFound-V to update the effects of multi-site mutations. As shown in Table S3 and Fig.5b, ProtFound-V outperforms other pretrained protein language models on multi-site mutation benchmark dataset in key indicators, including AUC, Accuracy, Precision, and the proposed metric *Ordinal Pair Proportion* (*OPP*). The remanufactured multi-site benchmark dataset breaks the limit of ProtFound-V to more variants, bringing in the constantly updated effects of mutations under different genetic backgrounds.

2.4.2 Transfer ProtFound-V to different genetic backgrounds

As the genetic background continues to change, the effects of mutations are constantly updating. Whether the model can generalize to predict high-risk mutations in unseen genetic backgrounds is crucial. Here, we adopt the deep mutational scanning data of four VOCs (Alpha, Beta, Delta, Omicron) and one VOI (Eta) [30] to test the robustness of ProtFound-V under different backgrounds. As shown in Fig.5c, ProtFound-V trained with single-site mutation dataset achieves robust prediction performance under different backgrounds, including Alpha, Beta, Eta, Delta, Omicron BA.1, and Omicron BA.2. Although only trained on a single-site mutation dataset, ProtFound-V is robust in predicting multi-site mutations. Due to possible random errors in wet-lab experiments, i.e. batch effect, the absolute values of the measured properties are not stable. As a result of this, the relative ranking among different variants has more reliable reference value. To this end, we propose a new evaluation metric, *Ordinal Pair Proportion* (*OPP*), to represent the correctly predicted proportion of any pair of mutated RBDs. In the lineages before Omicron, the *OPPs* are maintained above 78%, and that of Eta even reaches above 85%. *OPPs* of the two Omicron sublineages (Omicron BA.1, Omicron BA.2) are smaller than those of previous four lineages (Alpha, Beta, Eta, Delta), which is caused by the larger genetic background differences between Omicron and wild type. Even so, the *OPP* of Omicron BA.1 exceeds 71%, and that of Omicron BA.2 exceeds 66%. In conclusion, ProtFound-V remains robust in generalizing to new lineages in unseen genetic backgrounds, a benefit of evolution-inspired design.

2.5 ProtFound-V perceives evolutionary trends

2.5.1 ProtFound-V identifies the evolutionary advantage of Omicron

Since the outbreak of COVID-19, five VOCs have emerged, of which Omicron has developed many sublineages. Hence, the evolutionary advantages of Omicron are worth exploring, which can provide an important reference for analyzing future evolutionary trends. ACE2 binding is a crucial step for virus infection of host cells, so we started with it. ProtFound-V is applied to predict the binding affinity between human ACE2 and five VOCs, including Alpha, Beta, Gamma, Delta, and Omicron. As shown in Table S4, the predicted K_a *ratios* (greater than 1 means higher binding affinity and vice versa) of Omicron are close to 1, which means that its binding strengths to human ACE2 is similar to SARS-CoV-2 wild type RBD. We can deduce that the evolutionary

advantage of Omicron may lie not in binding affinity, but in immune escape and transmissibility. In fact, Han et al [39] measured and analyzed Omicron from the aspects of binding affinity, cell infection efficiency, and structural interaction, and came to a consistent conclusion: Omicron has similar binding strength with wild type RBD, which may be due to compensation in terms of immune escape and transmissibility. However, even if the enhancement of immune escape or transmissibility is the future direction of Omicron's mutations, we should be clear that maintaining qualified affinity is the basis for the survival of variants. ProtFound-V recognizes the evolutionary advantage of Omicron, which is a powerful tool for understanding the great leap from wild type to Omicron.

2.5.2 ProtFound-V predicts evolutionary trends consistent with wet-lab experiments

Through comprehensive validation of ProtFound-V on the binding affinity prediction, we demonstrate its ability to reveal few-site mutation effects and capture rare beneficial mutations, which is expected to be a powerful tool to predict the evolution trends of viruses. Evolutionary trends can facilitate the identification of potential high-risk mutations and prompt the prospective deployment of drug design. Therefore, as shown in Fig.6a, we construct a pipeline for predicting evolutionary trends based on ProtFound-V, which comprehensively evaluates single-site mutations through *in silico* deep mutational scanning. Specifically, referring to the practice of wet-lab deep mutational scanning, we sequentially mutate a amino acid of a single site in the RBD region to another 19 amino acids, and complete the mutational scanning in the entire region site by site. Subsequently, the binding affinity with ACE2, expression and antibody escape induced by each single-site mutation are predicted by ProtFound-V, and we comprehensively consider these properties to determine high-risk mutation sites.

The proposed pipeline is capable of exploring evolutionary differences in transmissibility between different lineages. More recently, XBB.1.5 is a lineage of popular dominance [1]. It has a significant growth advantage over XBB.1, although there is only one-site difference between them. To explore the mutational effects of this site, we perform a mutation scanning on site 486 and predict three key properties (ACE2 binding, expression, and antibody escape) for each mutation (Fig.6b). Compared with F486 in wild type, S486 in XBB.1 and P486 in XBB.1.5 have improved in all the three key properties. It is worth noting that P486 has a significant increase in ACE2-binding affinity and a certain degree of expression increase compared with S486, which is consistent with the results of previous deep mutational scanning studies [29, 30] and wet-lab measurements [40]. XBB.1.5 achieves a surge in binding affinity through F486P mutation while maintaining the qualified antibody escape, which likely contributes to its higher transmissibility, thus resulting in a significant growth advantage.

The proposed pipeline is competitive for wet-lab experiments in predicting evolutionary trends. Previous work [41, 42] estimated the mutational preference of Omicron BA.5 and XBB.1.5 in RBD region based on deep mutational scanning, which systematically observed evolutionary trends through a statistical method for wet-lab measurements. Using this statistical method for wet-lab deep mutational scanning, we

[1] https://outbreak.info/

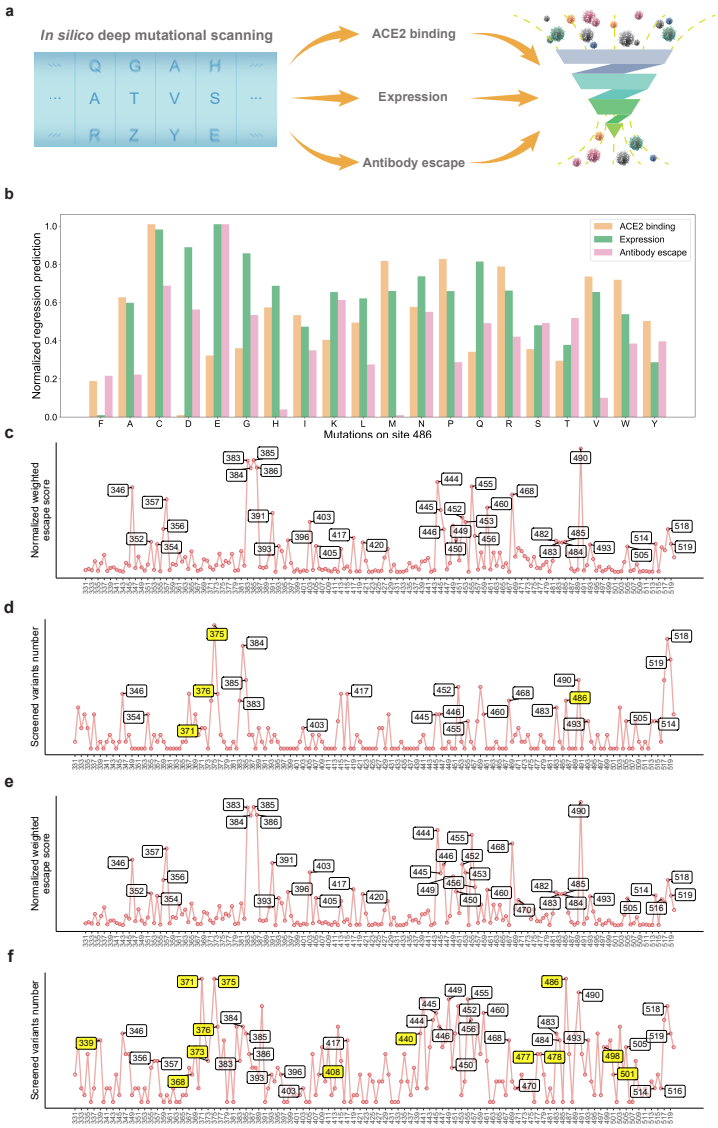

Fig. 6 a, A pipeline based on ProtFound-V for revealing evolutionary trends through *in silico* deep mutational scanning. The amino acid on each site of the RBD region is mutated to the remaining 19 types, which is predicted by ProtFound-V for binding affinity with ACE2, expression, and antibody escape. These predicted properties are in turn used to screen for high-risk mutations. b, *In silico* deep mutation scanning at site 486 to explore the difference in growth advantage between XBB.1.5 and XBB.1. c, The statistical evolutionary trends on wet-lab deep mutational scanning results of Omicron BA.5, i.e. the ground truth. d, The predicted evolution trends of Omicron BA.5 through the proposed pipeline. e, The statistical evolutionary trends on wet-lab deep mutational scanning results of XBB.1.5, i.e. the ground truth. f, The predicted evolution trends of XBB.1.5 through the proposed pipeline. The sites marked in yellow represent those where BA.5 or XBB.1.5 has been mutated.

incorporate two types of weights, ACE2 binding and RBD expression, on the basis of the antibody escape score to produce Fig.6c and Fig.6e. Since these results are statistics on the results of wet-lab experiments [41, 42], they can be considered as credible ground truth. However, virus mutations occur all the time, making expensive wet-lab measurements unable to keep up with the emergence of new variants. To address this issue, we systematically evaluate single-site mutations by *in silico* deep mutational scanning, i.e., predict three key properties for each mutation. The three key properties are then sequentially used to screen for high-risk mutations. As shown in Fig.6d and Fig.6f, the predicted evolutionary trends of BA.5 and XBB.1.5 have a high degree of coincidence with those based on a statistical method, i.e. ground truth. Consistency of predictions with ground truth demonstrates the effectiveness of ProtFound-V in predicting evolutionary trends by leveraging protein large language models. Sites marked in yellow (representing sites where BA.5 and XBB.1.5 have been mutated) need continuous attention because they still have a certain degree of threat. The emerging new sites indicate the potential evolutionary trends in the future, which is the possible mutation direction of SARS-CoV-2. Previous work [41] clustered antibodies based on deep mutational scanning experiments and proposed 30 corresponding escape hotspots, 19 of which are included in the predicted evolutionary trends of BA.5 and XBB.1.5. This indicates that the evolution of SARS-CoV-2 in the future tends to acquire stronger antibody escape capabilities in order to obtain stronger transmissibility.

3 Discussion

Viruses are ubiquitous and constantly evolving, demanding constant vigilance against sudden outbreaks of emerging viral infections. Traditional wet-lab experimental methods or conventional computing techniques fall short in meeting the throughput and speed requirements, making it difficult for humans to keep up with virus evolution. On the other hand, the development of artificial intelligence technology has led to the widespread use of large language models across various fields. Applications like ChatGPT have sparked curiosity about the untapped potential of these language models. Consequently, it is natural for us to consider constructing a customized large language model for the field of viruses to explore their evolutionary mechanisms. In recent years, significant progress has been made in modeling biological sequences using natural language processing techniques, particularly in the domain of proteins. This advancement prompts us to pretrain a large-scale language model on extensive protein sequences and employ it for predicting evolutionary trends of viruses. However, directly applying large language models to the virus domain proves impractical due to the unique trajectories of viruses. Generally, viruses tend to have only a small number of beneficial mutations in their evolutionary outcomes, leading to minimal changes in the variants and far fewer beneficial mutations compared to harmful ones. To address this challenge, we propose the evolution-inspired framework called ProtFound-V, leveraging customized protein language models pretrained from scratch. In this work, we conduct an in-depth exploration of SARS-CoV-2, a representative pathogen of emerging viral infections.

Specifically, our approach incorporates a local-global feature coupling mechanism to address the challenge posed by indistinguishable sequences of highly similar variants. This mechanism effectively fuses amino-acid-level interactions at different scales, enabling us to discern the subtle effects caused by a small number of mutations. As demonstrated in the dimension reduction visualization, we are able to clearly identify the subtle effects of few-site mutations. Additionally, we introduce a multi-task focal loss function that tackles the issue of extreme imbalance between beneficial and harmful mutations. This novel loss function alleviates the negative impact of rare samples, which are often difficult to learn in a multi-task setting. Through ablation experiments and case studies, we validate the accuracy of our high-risk mutation predictions, showcasing the sensitivity of our multi-task focal loss in identifying high-risk mutations. In summary, our evolution-inspired designs, including the local-global feature coupling and multi-task focal loss, empower ProtFound-V to effectively learn the evolutionary routes of viruses and explore their underlying mechanisms.

As an example, ProtFound-V proposes a new state-of-the-art performance on the single-site mutation benchmark dataset for ACE2 binding prediction. Notably, even though our protein language model has a significantly smaller parameter count compared to other models in the comparison, ProtFound-V still exhibits a competitive edge in key performance indicators. This outcome underscores the success of our pre-training strategy. To expand the predictive capabilities of ProtFound-V to multi-site mutations, we integrate data from multiple lineages, enabling the incorporation of updated mutational effects influenced by different genetic backgrounds. Results on the multi-site mutation benchmark dataset demonstrate the superior performance of ProtFound-V in discriminating mutations. Moreover, when we transfer ProtFound-V, trained solely on the single-site mutation benchmark dataset, to other lineages with multi-site mutations under distinct genetic backgrounds, the model remains robust. However, it is essential to note that the substantial genetic background differences observed in Omicron have a slight impact on prediction performance.

The first value of ProtFound-V lies in its ability to discriminate the evolutionary advantages among different viral lineages. In our ACE2 binding predictions for the five well-known VOCs (Alpha, Beta, Gamma, Delta, and Omicron), we have observed consistent results with wet-lab measurements from previous studies. Notably, our findings indicate that the troublesome Omicron variant exhibits a binding affinity similar to that of the wild type. This leads us to conclude that the evolutionary advantage of Omicron may not be in binding strength but rather in its potential for immune escape and increased transmissibility. Nevertheless, it is important to emphasize that maintaining qualified binding strength remains a fundamental requirement for SARS-CoV-2 variants, even if the binding affinity compensates for other abilities.

The second and most important value of ProtFound-V lies in its ability to predict evolutionary trends. By constructing an evolutionary trend prediction process with ProtFound-V as the core, we can conduct *in silico* mutational scanning on different lineages. This pipeline performs effectively in identifying evolutionary differences between various lineages and has revealed an important reason for the dominance of the current popular strain XBB.1.5. Most importantly, ProtFound-V demonstrates competitive performance with wet-lab experiments in predicting evolutionary trends,

highlighting the potential of using deep learning predictions to replace costly wet-lab experiments. By simulate the evolutionary trajectory of viruses using large language models and evolution-inspired design, ProtFound-V can facilitate efficient responses to emerging viral infections in a way that predicts mutation trends.

Through ProtFound-V, we have presented a paradigm that can be utilized to explore evolutionary mechanisms by leveraging large language models. This approach is also transferable and applicable for processing other biological sequences. Notably, the multi-task focal loss function we proposed for regression and classification addresses the challenges of multi-task hard sample mining, effectively dealing with issues such as limited data and class imbalance commonly encountered in the life sciences field. As a result, the potential applications of ProtFound-V extend beyond its current use, and it can be applied to optimize antibodies, enzymes, and more.

ProtFound-V, despite its notable achievements, still exhibits limitations that stem from three aspects: training data, research scale, and *in vitro* selection pressure. Regarding the training data, the dataset used in this study is derived from deep mutational scanning experiments, which may be limited in terms of variety and quantity for effectively training deep learning models. As data availability increases in the future, the prediction performance of ProtFound-V is expected to further improve. In terms of research scale, this study focuses on the key region RBD in the Spike protein of SARS-CoV-2 as an illustrative example to showcase ProtFound-V's effectiveness in deciphering the laws of virus evolution. Expanding its scope to encompass larger regions could enhance its understanding capabilities to a higher level. Furthermore, while this study involves predicting key properties of virus variants at the molecular level, the spread of viruses in the real world is also influenced by various *in vitro* selection pressures. Integrating an *in vitro* selection pressure module in the future would enable ProtFound-V to extend its prediction ability to the spread scale of emerging viruses. Addressing these limitations in future research would undoubtedly enhance ProtFound-V's effectiveness and applicability, further contributing to our understanding of virus evolution and aiding in the prediction and response to emerging viral infections.

4 Methods

4.1 Dataset

The UniRef90 dataset [43]is adopted for the pretraining of language model. It contains approximately 144 million raw protein sequences.

For ACE2 binding prediction, the deep mutational scanning (DMS) datasets of previous work [29, 30] reported the difference in $\log 10 K_d$, i.e. apparent dissociation constant, relative to wild type or a specific lineage. Specifically, the original data form is that one mutation corresponds to one $\Delta \log_{10} K_d$. First, based on the mutations, we perform residue substitutions 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* as

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

After the above two-step processing, we can obtain the mutated RBD sequences and corresponding labels representing the binding strength. Regarding the single-site mutation benchmark dataset, there are 3648 samples, including 3361 negative samples with label less than 1, and 287 positive samples with label greater than 1. Regarding the multi-site mutation benchmark dataset, we integrate seven groups of DMS dataset [30] including wild type, Alpha, Beta, Delta, Eta, Omicron BA.1 and Omicron BA.2. Finally, there are 26,369 samples, including 23,629 negative samples with label less than 1, and 2,740 positive samples with label greater than 1.

The *in silico* deep mutational scanning consists of three steps, namely ACE2 binding prediction, expression prediction and antibody escape prediction. For ACE2 binding prediction and expression prediction, we use the DMS dataset of previous work [30]. We only use the samples of Omicron BA.1 and Omicron BA.2 due to the significant mutation differences between the lineages before and after Omicron. We normalize with the $\log_{10} K_d$ and the expression value of Omicron BA.2 as a reference to get the final 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 ACE2 binding dataset, while there are 7,182 negative samples and 639 positive samples in the expression dataset.

For antibody escape prediction, we use the DMS dataset of wild type with different antibodies from previous work [41]. The dataset is the DMS results of 3051 antibodies with wild type, including antibody names, mutations on wild type, and the corresponding escape scores. Here we use the escape scores as labels directly, and set 0.4 as the classification threshold, i.e. samples that are greater than 0.4 are considered as positive samples, others are negative samples. According to previous work [44], 0.4 is a high escape score among all the scores, so we assume that 0.4 is a proper standard to distinguish a strong escape. Finally, there are 57,740 positive samples and 284,032 negative samples in the dataset.

4.2 Protein language model pretraining

As the universal basic unit of life, protein can be regarded as the language of life. There are three levels of similarity between natural language and protein sequences, i.e., amino acids corresponding to letters, words corresponding to conserved protein fragments, and sentences corresponding to structures carrying different functions [45]. UniRef90 [43], a database with approximately 144 million protein sequences clustered on the base of UniRef100 [46], is adopted as the training data of transformer-based language model [10], i.e., ProtFound-V here. The model architecture of ProtFound-V is similar to $BERT_{LARGE}$ [47], except that the next sentence prediction task is not adopted. The configuration is as follows: 24-layer, 1,024 hidden dimension, 16 attention heads, and 340 million parameters. Inspired by LAMB [48], we adopt the two-step training strategy which facilitates the convergence of the model and promotes the training efficiency when dealing with longer sequences. In the first training stage, protein sequences with a maximum length of 128 are used for model training with a learning rate of 0.0004, warm-up steps of 40,000, and epochs of 65. In the second training stage, protein sequences with a maximum length of 256 are used for model

training with a learning rate of 0.0001 and epochs of 35. It is pretrained with only the objective of masked language modeling with a masking probability of 15%. The LAMB optimizer [48] and weight decay of 0.01 are adopted. The batch size of single Neural Processing Units (NPU) is 64, and the global batch size is 16,384. With the assistant of high performance computing strategy, our protein language model is pretrained with 256 NPUs and the training time is significantly reduced to about 100 hours. After the masked language modeling with negative log-likehood (NLL) loss function, the predicted probabilities per position in the protein sequences are obtained, which carry the reasonable distribution of amino acids in the natural protein sequence.

4.3 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. The length of the tokenized sequence is l, and the shape of the embedding for each RBD sequence is $d \times l$.

It is worth noting that there are usually only a few mutations on the RBD sequences, so the sequences input into the pretrained protein language models are usually similar to each other, which makes the embeddings of different sequences very similar. So after extracting embeddings, we conduct standardization to enlarge the difference between different sequences. Specifically, let $X_{d\times l}$ be the extracted embedding feature, 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

$$\begin{cases} std_i = \mathrm{STD}_N(X^f_i), \\ mean_i = \mathrm{MEAN}_N(X^f_i), \\ \tilde{X}^f_i = \frac{X^f_i - mean_i}{std_i}, \end{cases} \tag{2}$$

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 antibody escape prediction, the input involves RBD sequences and antibody sequences, and each antibody includes heavy chain sequence and light chain sequence. We do the operations separately to the RBD sequences, antibody heavy chain sequences and antibody light chain sequences as described above. We pad the heavy chain sequences and light chain sequences to 136, which is the largest length of all the antibody chains. We also conduct standardization to the extracted embeddings of antibody heavy chain sequences and light chain sequences.

4.4 Local-global feature coupling

After feature embedding and standardization, we use ContextPool-attention and CNN to capture the long-range dependencies and the near-range dependencies of the sequence embeddings, respectively.

4.4.1 ContextPool-attention

Self-attention [10] is widely used in language modeling. However, it needs to assume that each individual token constructs the query and key vectors over a fixed granularity, which may be sub-optimal for modeling context in different scales. Context pooling [18] 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.

Specifically, we learn a mapping function $m(\cdot)$ 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, as in Huang et al. [18], 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

$$y_i = f_{ave}(\tilde{X} \odot \gamma(w) \odot \gamma(g^i)) = \sum_{j=1}^{n} \tilde{x}_j \cdot w_j \cdot g_j^i, \tag{3}$$

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 $\odot$.

During the actual training of the above model, the value of the Gaussian mask is relatively large, and the training loss is large at the beginning of training. In order to improve the training stability, we normalize the Gaussian mask vector by dividing it by its maximum value. Specifically, for the Gaussian mask of the ith place, the normalization is calculated as

$$\tilde{g}^i = \frac{g^i}{\mathrm{MAX}_{j=1}^{l}(g_j^i)} \in [0, 1]. \tag{4}$$

Compared with the original module, the normalized one is more stable during training and the testing effect is slightly better.

4.4.2 CNN

We adopt a one-dimension convolution neural network to extract the local dependencies between neighboring residues. The CNN architecture has been widely used to integrate local dependencies and capture latent information of sequences in the area of protein-protein interactions [49, 50] and compound-protein interactions [51, 52]. 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.

4.4.3 Feature coupling

The output of ContextPool-attention and CNN both have the same shape as the input, so a shortcut can be added on the both. The outputs of the both are concatenated together to get the final extracted features of residues, and the feature of the whole RBD sequence is calculated through max pooling on the residue dimension. Specifically, for each RBD sequence, the shapes of the two outputs are both $1024 \times l$, and the shape of the concatenation is $2048 \times l$. The result of the max pooling is 2048×1. The extracted feature of the whole RBD sequence is used for the subsequent prediction tasks.

For antibody escape prediction, the embeddings of RBD, antibody heavy chain, and antibody light chain are input into independent local-global feature coupling channel respectively to capture long-range dependencies and near-range dependencies. The features of each kinds of sequence are calculated through max pooling on the residue dimension, and the results are concatenated as the feature to predict antibody escape. Specifically, for each kind of sequence, the shape of the max pooling is 2048×1, and the concatenation of features of three kinds of sequence is 6144×1, which is used for the subsequent prediction.

4.5 Loss function design

We design Multi-task focal loss function to alleviate the imbalance between positive and negative samples. It consists of two focal losses according to the characteristics of each task, i.e. the regression focal loss and the classification focal loss.

For classification task, the training intensity of a certain category can be adjusted by assigning different weights to samples with different labels in classification focal loss to alleviate the imbalance between positive and negative samples in training dataset. Specifically, the calculation of the classification focal loss is as follows

$$p_t = \begin{cases} pred, label = 1, \\ 1 - pred, label = 0. \end{cases} \tag{5}$$

$$\alpha_t = \begin{cases} \alpha, label = 1, \\ 1 - \alpha, label = 0. \end{cases} \tag{6}$$

$$ClsFocalLoss = -\alpha_t (1 - p_t)^{\gamma} \log p_t, \tag{7}$$

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 to alleviate the imbalance between positive and negative samples in training dataset.

For regression task, we propose the regression focal loss inspired by the above classification focal loss. The calculation of the regression focal loss is as follows.

$$\begin{cases} a_i = (score - target)^{\gamma}, \gamma = 1, 2, 3, ..., \\ \hat{a}_l = \frac{a_i}{\sum a_i}, \\ l_i = \hat{a}_l (score - target)^2, \\ RegFocalLoss = \sum_{i=1}^{B} l_i, \end{cases} \tag{8}$$

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 classification focal loss.

The final focal loss function is:

$$FocalLoss = \mu * RegFocalLoss + ClsFocalLoss, \tag{9}$$

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

4.6 Metrics

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

$$\begin{cases} MSE = \frac{1}{N} \sum_{i=1}^{N} (y_{pred_i} - y_{label_i})^2, \\ Corr = \frac{\sum_{i=1}^{N} (y_{pred_i} - \overline{y_{pred}})(y_{label_i} - \overline{y_{label}})}{\sqrt{\sum_{i=1}^{N} (y_{pred_i} - \overline{y_{pred}})^2} \sqrt{\sum_{i=1}^{N} (y_{label_i} - \overline{y_{label}})^2}}, \end{cases} \tag{10}$$

where N is the number of samples.

For classification task, we set the label of samples with K_a *ratio* > 1 as $True$, and the label of samples with K_a *ratio* ≤ 1 as $False$. The metrics for classification includes AUC, $Accuracy$, $F1 - Score$, $Precision$ and $Recall$ between label and prediction.

AUC 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

$$AUC = 1 - \frac{1}{m^+ m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} (\mathbb{I}(f(x^+) < f(x^-)) + \frac{1}{2} \mathbb{I}(f(x^+) = f(x^-))), \tag{11}$$

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 prediction 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

$$\begin{cases} Accuracy = \frac{TP+TN}{TP+TN+FP+FN}, \\ Precision = \frac{TP}{TP+FP}, \\ Recall = \frac{TP}{TP+FN}, \\ F1-Score = \frac{2*Precision*Recall}{Precision+Recall}. \end{cases} \tag{12}$$

In addition to the metrics mentioned above, we propose a new evaluation metrics, $Ordinal\ Pair\ Proportion(OPP)$ to evaluate prediction performance of relative rankings. We first combine all the prediction results into binary pairs. And for each pair, if the label order of the two predictions is consistent with the prediction result order, 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

$$\begin{cases} OPP = \frac{|M|}{C_n^2}, \\ M = \{(i,j)|\forall i,j \in N, i < j, (label_i - label_j)(pred_i - pred_j) > 0\}. \end{cases} \tag{13}$$

4.7 *In silico* deep mutational scanning

To explore the difference between XBB.1 and XBB.1.5 in growth advantage, we mutate the residue on site 486 of XBB.1 into other 19 residues. Subsequently, the ACE2 binding prediction model, expression prediction model, and antibody escape prediction model are adopted to predict properties of each variant. We use the regression output and do normalization within each kinds of property due to the difference of output range between properties.

To predict the evolutionary trends of Omicron BA.5 and XBB.1.5, we take them as reference sequences, and enumerate all the mutated variants with single mutation. The ACE2 binding, expression and antibody escape of each variants are predicted. As to the antibody sequence used during antibody escape prediction, we adopt BD57-0129 from previous work [41] as it achieves relatively strong neutralizing capability, and variants escaping it indicates strong escape capability. Based on the predicted properties of variants, we set different thresholds to do variant filtering to get possible high-risk variants. For the property that is more essential for variants to survive, like expression, we set a higher threshold. For the less vital property, we set the threshold relatively lower to avoid filtering out too many variants. Finally, we use classification output, and set the threshold of ACE2 binding as 0.25, expression as 0.7, antibody escape as 0.55. For the filtered variants, the number of possible mutations at each site is considered as indicator of evolutionary trends. In other words, the greater the number of possible mutations, the higher the risk of the site.

Data availability

For the pretraining of the protein language model, we downloaded UniRef 90 from https://www.uniprot.org/. For the training and testing of ACE2 binding prediction,

expression prediction, and antibody escape, we obtained and processed deep mutational scanning data from previous work. The processed datasets are available at https://github.com/ZhiweiNiepku/ProtFound-V.

Code availability

The relevant codes of model training and inference are available at https://github.com/ZhiweiNiepku/ProtFound-V, and we also released the pretrained protein language model at the link above.

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

[9] 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).

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

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

[12] 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).

[13] Wang, S., Guo, Y., Wang, Y., Sun, H. & Huang, J. *Smiles-bert: large scale unsupervised pre-training for molecular property prediction*, 429–436 (2019).

[14] Wu, Z. *et al.* Knowledge-based bert: a method to extract molecular features like computational chemists. *Briefings in Bioinformatics* **23**, bbac131 (2022).

[15] Bagal, V., Aggarwal, R., Vinod, P. & Priyakumar, U. D. Molgpt: molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling* **62**, 2064–2076 (2021).

[16] 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).

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

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

[19] Dong, Q., Gong, S. & Zhu, X. *Class rectification hard mining for imbalanced deep learning*, 1851–1860 (2017).

[20] Sheng, H. *et al.* Mining hard samples globally and efficiently for person reidentification. *IEEE Internet of Things Journal* **7**, 9611–9622 (2020).

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

[22] Zhang, Y. & Yang, Q. An overview of multi-task learning. *National Science Review* **5**, 30–43 (2018).

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

[24] Nelson, G. *et al.* Molecular dynamic simulation reveals e484k mutation enhances spike rbd-ace2 affinity and the combination of e484k, k417n and n501y mutations (501y. v2 variant) induces conformational change greater than n501y mutant alone, potentially resulting in an escape mutant. *BioRxiv* (2021).

[25] Rezaei, S., Sefidbakht, Y. & Uskoković, V. Comparative molecular dynamics study of the receptor-binding domains in sars-cov-2 and sars-cov and the effects of mutations on the binding affinity. *Journal of Biomolecular Structure and Dynamics* **40**, 4662–4681 (2022).

[26] Luan, B., Wang, H. & Huynh, T. Enhanced binding of the n501y-mutated sars-cov-2 spike protein to the human ace2 receptor: insights from molecular dynamics simulations. *FEBS letters* **595**, 1454–1461 (2021).

[27] Yang, Y. *et al.* Key residues of the receptor binding domain in the spike protein of sars-cov-2 mediating the interactions with ace2: a molecular dynamics study. *Nanoscale* **13**, 9364–9370 (2021).

[28] 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).

[29] 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).

[30] 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).

[31] Zhang, R., Ghosh, S. & Pal, R. Predicting binding affinities of emerging variants of sars-cov-2 using spike protein sequencing data: observations, caveats and recommendations. *Briefings in Bioinformatics* **23**, bbac128 (2022).

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

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

[34] Elnaggar, A. *et al.* Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225* (2020).

[35] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

[36] 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).

[37] Witte, L. *et al.* Epistasis lowers the genetic barrier to sars-cov-2 neutralizing antibody escape. *Nature Communications* **14**, 302 (2023).

[38] Moulana, A. *et al.* Compensatory epistasis maintains ace2 affinity in sars-cov-2 omicron ba. 1. *Nature Communications* **13**, 7011 (2022).

[39] Han, P. *et al.* Receptor binding and complex structures of human ace2 to spike rbd from omicron and delta sars-cov-2. *Cell* **185**, 630–640 (2022).

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

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

[42] Yisimayi, A. *et al.* Repeated omicron infection alleviates sars-cov-2 immune imprinting. *bioRxiv* 2023–05 (2023).

[43] Steinegger, M. & Söding, J. Clustering huge protein sequence sets in linear time. *Nature communications* **9**, 1–8 (2018).

[44] 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).

[45] Ferruz, N. & Höcker, B. Controllable protein design with language models. *Nature Machine Intelligence* **4**, 521–532 (2022).

[46] Suzek, B. E. *et al.* Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).

[47] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[48] You, Y. *et al.* Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962* (2019).

[49] Litfin, T., Yang, Y. & Zhou, Y. Spot-peptide: template-based prediction of peptide-binding proteins and peptide-binding sites. *Journal of chemical information and modeling* **59**, 924–930 (2019).

[50] Lei, Y. *et al.* A deep-learning framework for multi-level peptide–protein interaction prediction. *Nature communications* **12**, 1–10 (2021).

[51] Öztürk, H., Özgür, A. & Ozkirimli, E. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).

[52] Li, S. *et al.* Monn: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Systems* **10**, 308–322 (2020).

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).

Author information

Contributions

Z.N. designed the project. Z.N and X.L. trained the model, performed the analysis, and wrote the paper. Z.W. and F.X. contributed to the pre-training of protein language models. Y.L., H.S., T.D., G.S. and W.G. contributed to useful discussions. P.Z., Y.W., J.C., Y.T. contributed to scheme design. J.C., Y.T. supervised this project.

Corresponding authors

Correspondence to Jie Chen, Yonghong Tian.

Competing interests

The authors declare no competing interests.

Extended data tables

Table S1 Ablation experiments of our multi-task focal loss

Loss function	AUC	Accuracy	F1-Score	Precision	Recall	MSE	Correlation
BCE&MSE	84.06%	57.41%	26.33%	97.14%	15.56%	0.060	0.871
BCE&RegFocalLoss	88.59%	58.52%	28.93%	100.0%	17.04%	0.059	0.873
MSE&ClsFocalLoss	93.69%	85.56%	84.92%	88.28%	82.22%	0.055	0.879
Multi-task focal loss	92.98%	**91.11%**	**91.58%**	87.43%	**96.30%**	0.058	0.870

Table S2 Prediction performance on single-site mutation benchmark dataset

Method	Encoder	AUC	Accuracy	F1-Score	Precision	Recall	MSE	Corr
NN_MM-GBSA	Molecular Dynamics Simulation	-	80.41	-	-	-	0.20	0.79
BiLSTM	-	-	-	-	-	-	0.23	0.13
CNN	One-hot	-	-	-	-	-	0.11	0.76
ProtFound-V (without protein LM[1])	One-hot	61.22	58.89	51.07	58.07	49.63	0.304	0.255
	SeqVec	91.36	85.93	86.41	83.02	90.37	0.071	0.837
	ProtTrans T5	92.81	87.41	87.46	86.31	88.89	0.079	0.818
	ESM-1b	95.86	90.37	90.63	88.32	93.33	0.059	0.873
ProtFound-V (protein LM)	ESM-1v (mean)[2]	93.90	89.11	89.27	88.06	90.81	0.063	0.853
	ESM-2 (650M)[3]	88.26	85.19	85.44	83.71	88.89	0.080	0.802
	ESM-2 (15B)[4]	93.83	90.37	90.66	87.69	94.07	0.059	0.863
	Our protein LM (340M)	92.98	**91.11**	**91.58**	87.43	**96.30**	**0.058**	0.870

[1]LM represents language model.

[2]The results of five models with different seeds of ESM-1v are averaged here.

[3]The ESM-2 model with 650 million parameters is chosen here to obtain an embedding dimension (1280) close to that of ProtFound-V (1024).

[4]The largest ESM-2 model with 15 billion parameters.

Table S3 Prediction performance on multi-site mutation benchmark dataset

Method	Encoder	AUC	Accuracy	F1-Score	Precision	Recall	MSE	Corr
ProtFound-V (protein LM[1])	SeqVec	86.68	79.50	80.51	76.86	84.60	5.861	0.898
	ProtTrans T5	84.66	77.33	79.37	72.97	87.13	6.345	0.882
	ESM-1b	84.93	77.17	77.54	76.37	78.93	5.251	0.937
	ESM-1v (mean)[2]	85.27	78.23	79.17	76.14	82.77	5.064	0.922
	ESM-2 (650M)[3]	83.25	75.83	76.65	74.21	79.33	5.818	0.903
	ESM-2 (15B)[4]	84.74	76.57	76.97	75.95	78.80	6.323	0.898
	Our protein LM (340M)	**87.37**	**79.83**	79.97	**79.49**	80.53	6.181	0.912

[1]LM represents language model.

[2]The results of five models with different seeds of ESM-1v are averaged here.

[3]The ESM-2 model with 650 million parameters is chosen here to obtain an embedding dimension (1280) close to that of ProtFound-V (1024).

[4]The largest ESM-2 model with 15 billion parameters.

Table S4 ACE2 binding predictions of ProtFound-V on five VOCs.

VOCs	Alpha	Beta	Gamma	Delta	Omicron (B.1.1.529)
Predicted K_a *ratio*	1.71	1.28	1.23	1.05	**1.04**